TP 2 : Initiation au logiciel R[[1]](#footnote-1)

# Manipuler des données

Les données sont généralement dans un fichier texte (extension .txt ou .csv) avec les individus en lignes et les variables en colonnes.

1☺☺☺ Créer un dossier « TP2 » dans votre Dossier « Formation\_R »

## Importer des données

### Importer des données avec Rstudio

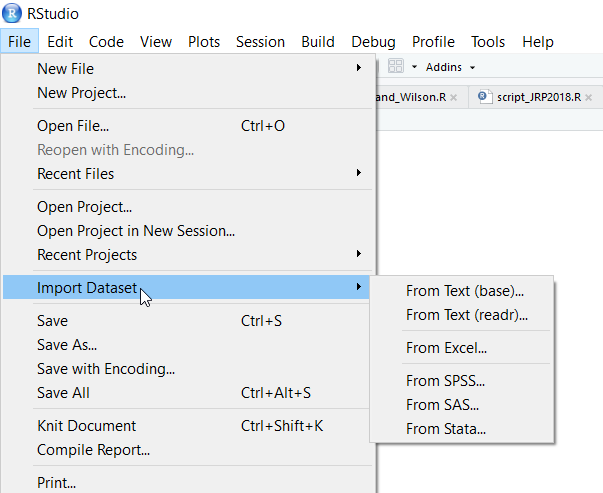
Rstudio permet d’importer des fichiers de plusieurs formats : fichiers texte, fichier Excel, fichier d’autres logiciels statistiques comme SAS, SPSS ou Stata).

**Importer un fichier Excel :**

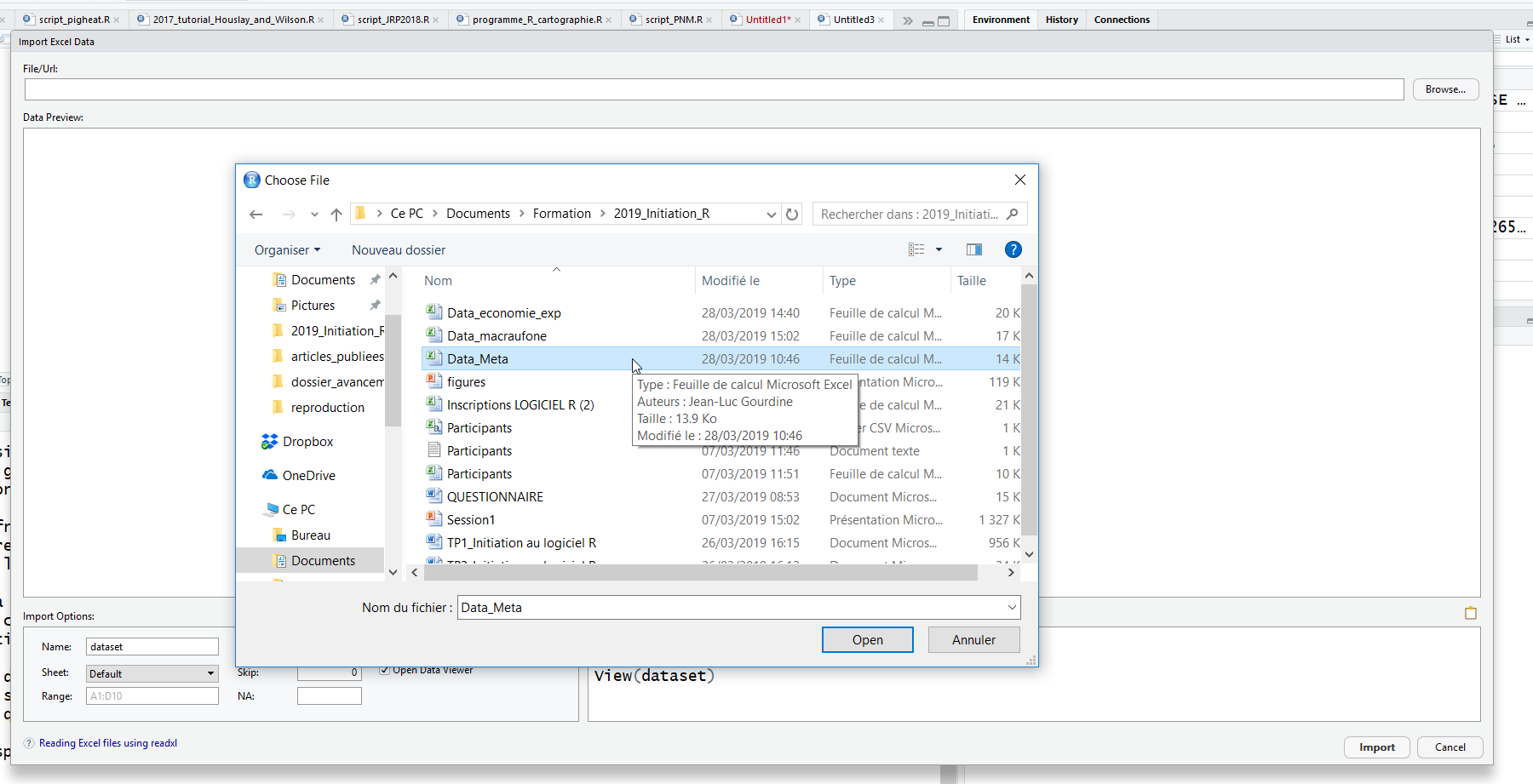
Il y a plusieurs façons d’importer des données. On va procéder à une importation avec l’aide de la souris qui est la manière la plus simple de procéder à une importation.

2☺☺☺ Cliquez sur File -> Import Dataset

3☺☺☺ Puis choisir le format approprié (dans l’exemple format Excel)



4☺☺☺ Sélectionner le fichier à importer en parcourant les dossiers avec « Browse… »



5☺☺☺ Une fois sélectionner, cliquez sur « Open »

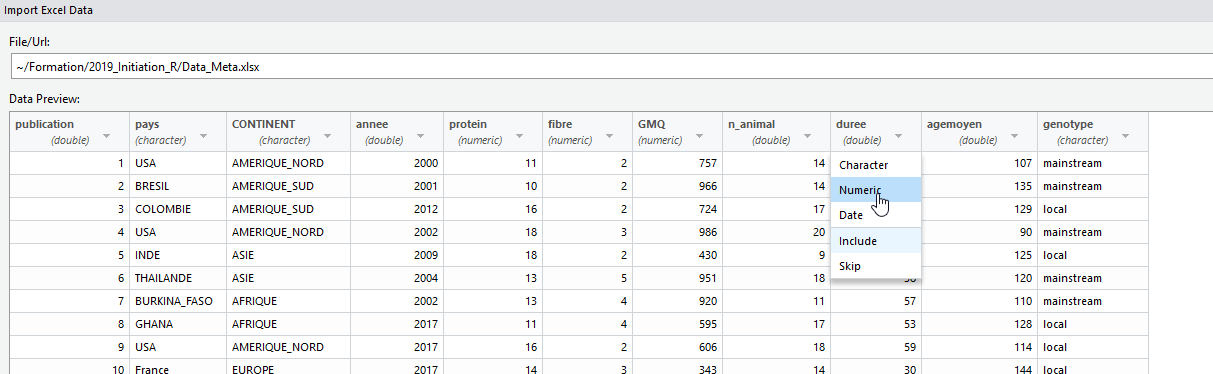
6☺☺☺ Contrôler dans « Data Preview : » les données et corriger si nécessaire:

* 7☺☺☺ Vérifier le type de chaque variable et corriger si nécessaire

Rstudio propose les types suivants :

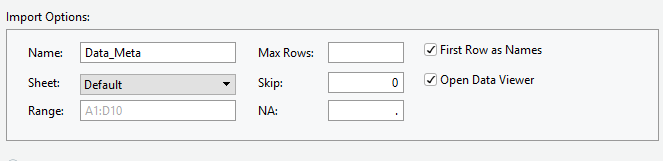
* Double (entier)
* Character (caractère)
* Numeric (numérique)
* Date (date)
* Include

Skip est utilisé quand on ne souhaite pas que la variable fasse partie du jeu de données.

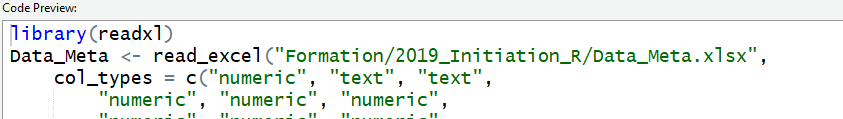


Ici on change le type de la variable « duree » en numérique.

8☺☺☺ Dans la fenêtre « Import Options », vérifier les options et modifier si nécessaire

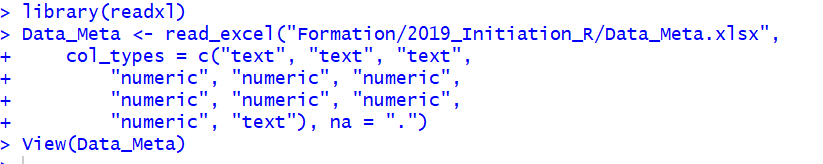


9☺☺☺ Dans la fenêtre «Code Previews », visualiser les codes correspondants à l’importation



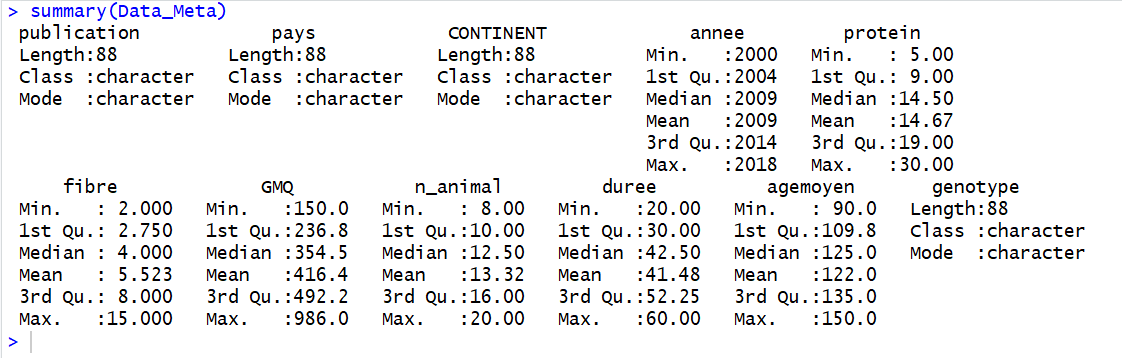
10☺☺☺ Après vérification, cliquez sur « Import »

11☺☺☺ Interpréter les codes résultant de l’importation :

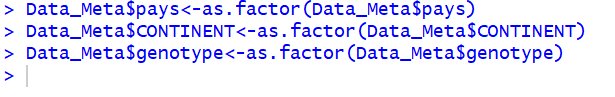


12☺☺☺ Déterminer le type de l’objet importé

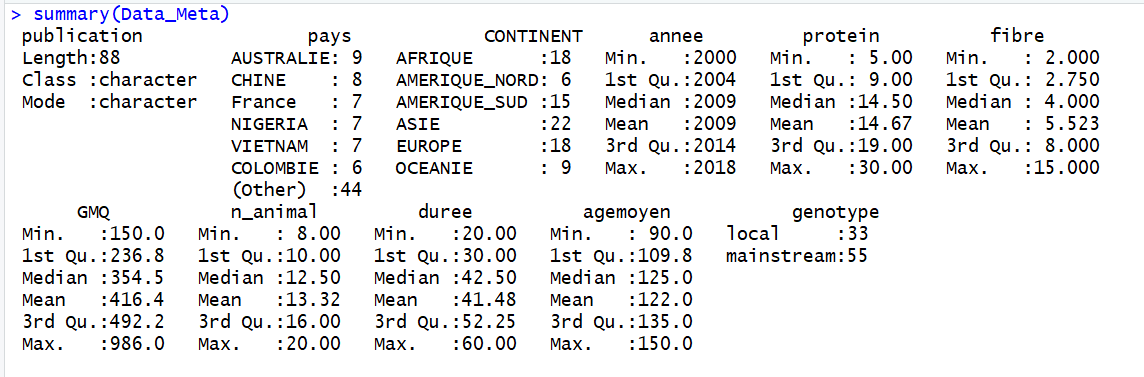
On peut vérifier les données à l’aide de la fonction **summary**.



Ici, les types de variables sont corrects pour « protein, fibre, GMQ, duree, agemoyen » car ce sont des variables quantitatives. En revanche, les variables « pays, continent, genotype » sont de type factor. On modifie avec les commandes suivantes :



Et on vérifie le changement en réexécutant la fonction **summary** :



**13☺☺☺ Quelles différences observez-vous entre la première et la deuxième exécution de « summary »**

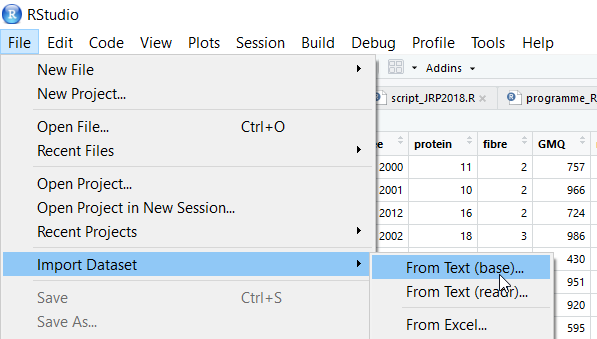
### 

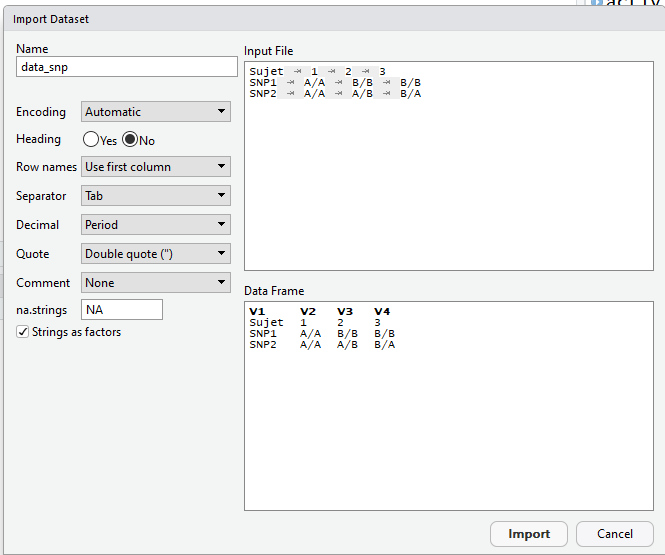
**Remarques importantes :**

Certains tableurs limitent le nombre de colonnes (donc le nombre de variables). C’est le cas des données de spectrométrie ou génomique. Dans ce cas, on a souvent des données de type variables x individus au lieu des données habituelles individus x variables. Dans ce cas, il suffit simplement de transposer le jeu de données variables x individus sous R, en vérifiantque la tranposition ne modifie pas le type des variables.

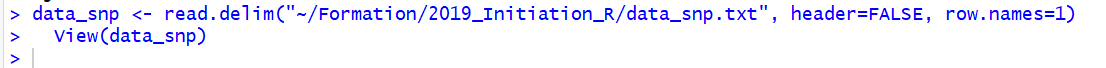
Voici un exemple, avec un jeu de données réduit avec le fichier data\_snp.txt

14☺☺☺ Importer le jeu de données



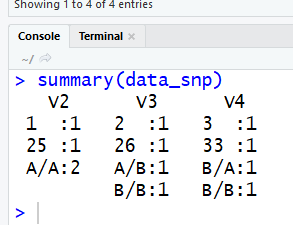


15☺☺☺ Interpréter les codes obtenus :



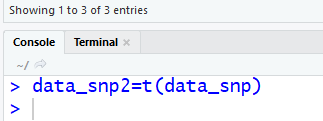
16☺☺☺ Vérifier que c’est un data-frame

17☺☺☺ Utiliser la fonction **summary** pour contrôler l’importation

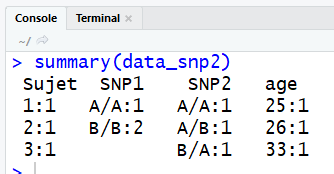


18☺☺☺ Que constatez-vous ?

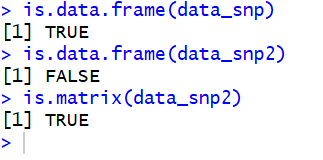
19☺☺☺ Réaliser la transposition du tableau avec la fonction **t**



20☺☺☺ Utiliser la fonction **summary** pour contrôler la transposition

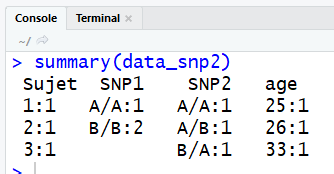


Vous remarquez que les variables sujets, SNP1, SNP2 et age sont résumées comme des facteurs. En fait, le résultat n’est plus un data-frame mais une matrice de caractères.

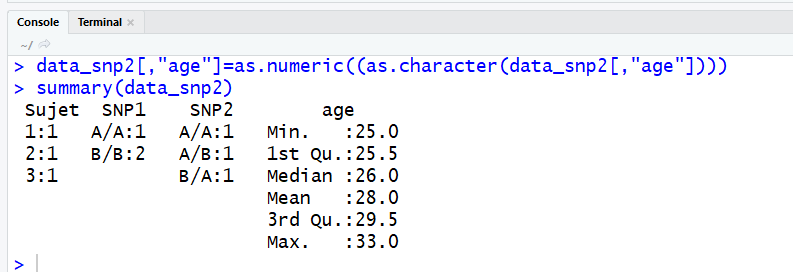


Pour avoir un tableau transposé convenable, il faut :

* Transformer la matrice en data-frame
* Puis transformer en numérique les variables qui doivent être numériques (ici la variable age)

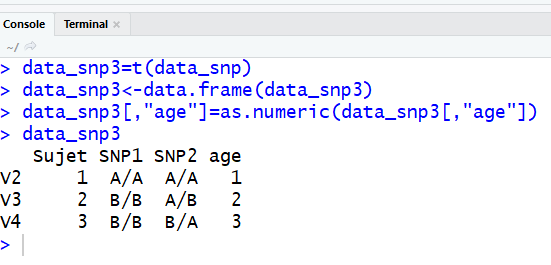


On constate que la variable **age** est considérer comme une variable qualitative.



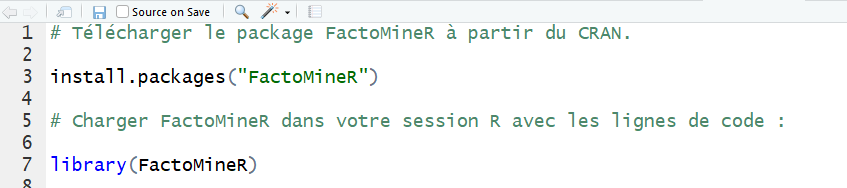
Le passage en numérique se fait en deux étapes car il faut i) transformer le type facteur en type caractère puis ii) transformer le type caractère en numérique.

21☺☺☺ Réaliser une conversion directe de facteur en numérique pour mieux comprendre

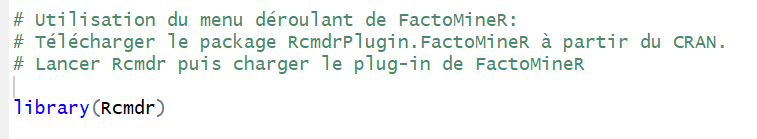


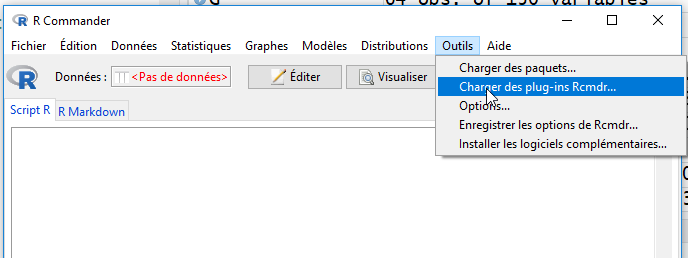
### Importer des données avec FactoMineR

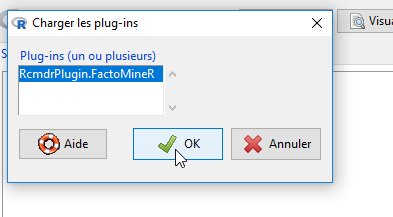
22☺☺☺ Télécharger, charger le package FactoMineR à partir du CRAN si ce n’est pas encore fait.

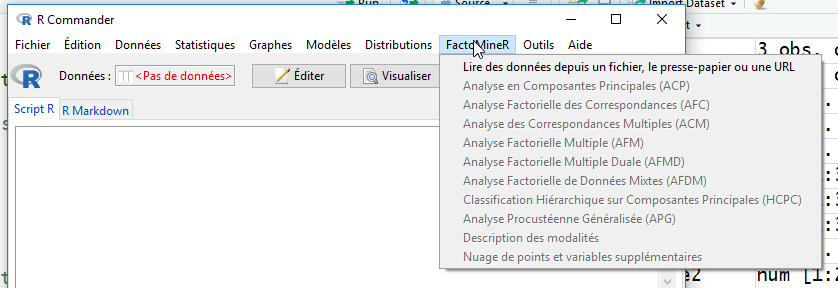


23☺☺☺ Utilisation du menu déroulant de FactoMineR: pour ce faire, télécharger le package RcmdrPlugin.FactoMineR à partir du CRAN et lancer Rcmdr puis charger le plug-in de FactoMineR

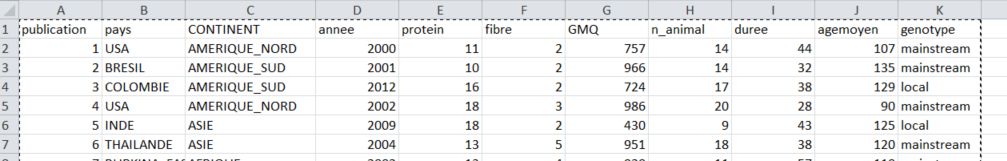


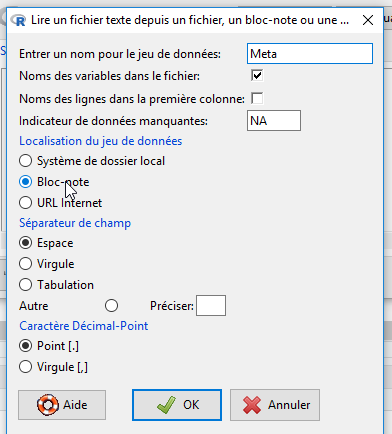




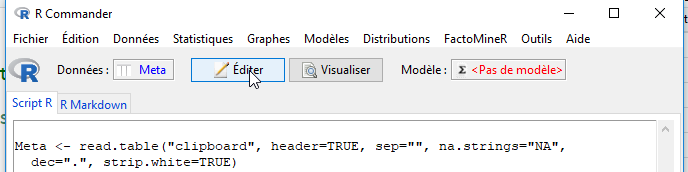


24☺☺☺ Importer le fichier Data\_Meta en copiant le jeu de données dans le Bloc-Notes, puis en fermant le fichier.

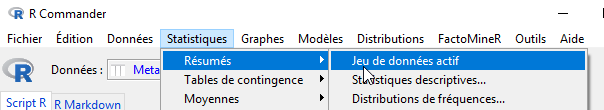




25☺☺☺ Cliquez sur « Visualiser » pour visualiser le jeu de données.

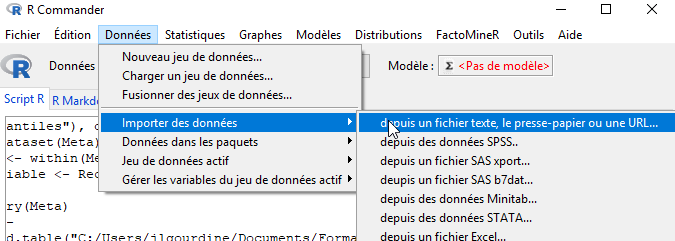


26☺☺☺ Obtenir un résumé du jeu de données en cliquant sur :



### Importer des données avec R Commander

27☺☺☺ Importer le jeu de données data\_snp.txt ;



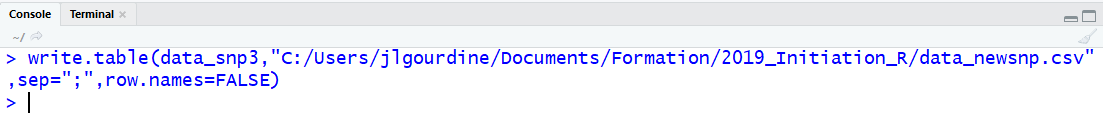
## Exporter des données

Lorsque les analyses ont été effectuées, il est souvent utile de communiquer les résultats à d’autres personnes ou pour d’autres logiciels. L’idéal est de réaliser une exportation sous le format texte.

### Exporter des données avec des commandes

On utilise la fonction write.table

28☺☺☺ Importer le jeu de données data\_snp3

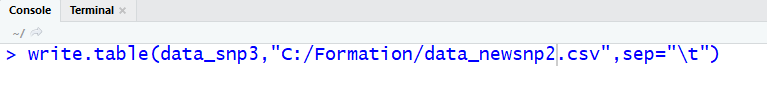




Noter aussi que par défaut la fonction write.table définit le fichier de résultats avec :

* le nom des lignes et des colonnes
* les chaines de caractères sont délimitées par des guillemets
* le séparateur de décimal (option dec) est « . »
* les valeurs manquantes sont définies par « NA »

Si l’on souhaite que le séparateur de colonne soit la tabulation on utilise l’option sep de la manière suivante :



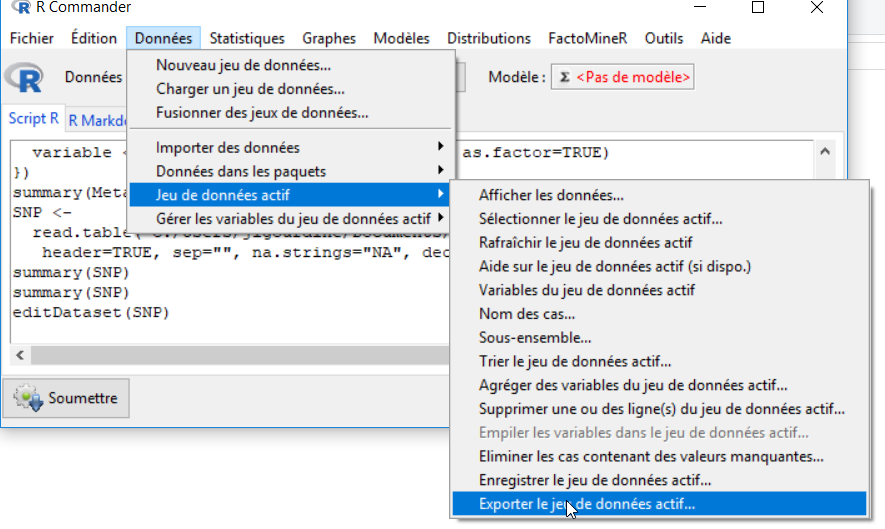
29☺☺☺ Exporter le jeu de données data\_snp3 dans un fichier nommé « data\_newsnp2.txt » avec la tabulation comme séparateur de colonnes.

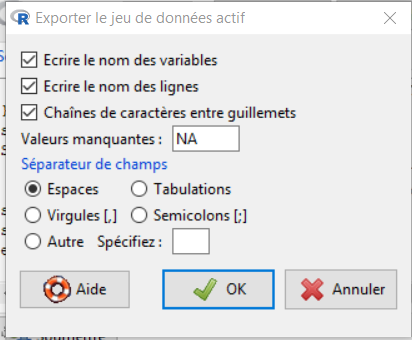
30☺☺☺ Ouvrer le fichier de résultats avec Excel.

Noter qu’il existe de nombreuses fonctions d’exportation autres que write.table. Par exemple, on peut utiliser write.csv.

Par ailleurs, la fonction write.infile du package FactoMineR permet d’écrire tous les objets d’une liste dans un même fichier, sans avoir à préciser les objets contenus dans la liste (voir 1.2.3).

### Exporter avec R Commander



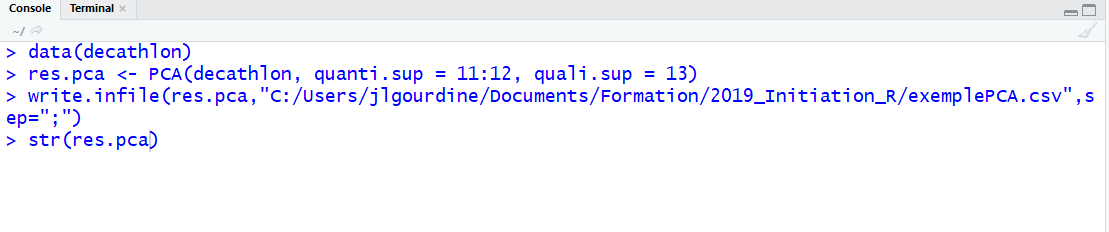


31☺☺☺ Exporter le jeu de données data\_meta avec R Commander.

### Exporter avec FactoMineR

32☺☺☺ Afficher l’aide de la fonction write.infile du package FactoMineR

33☺☺☺ Exécuter l’exemple disponible dans l’aide



34☺☺☺ Interpréter le rôle de la fonction str en exécutant la commande str(res.pca)

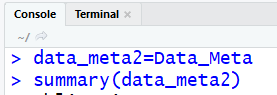
35☺☺☺ Vérifier le résultat de l’exportation en ouvrant le fichier résultat

## Manipuler les variables

### Changer de type avec Rstudio

Parfois il s’avère nécessaire de changer de type de variable. Par exemple, lors de l’importation, une variable qualitative dont les modalités sont codés en chiffres a été interprétée par R comme une variable quantitative.

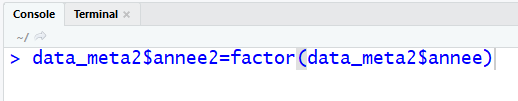
36☺☺☺ Sous Rstudio , copier le jeu de données data\_meta dans un nouvel objet data\_meta2 et réaliser un résumé des variables à l’aide de la fonction summary.



Vous constatez que la variable annee qui code pour l’année de publication est considérée comme une variable quantitative. On souhaite la transformer en variable qualitative.

37☺☺☺ Nous avons utilisé la fonction summary pour vérifier que la variable annee est quantitative. Trouver deux autres méthodes de vérification.

38☺☺☺ Créer une nouvelle colonne dans le jeu de données data\_meta2 contenant la variable annee2 qui correspond à la variable qualitative annee



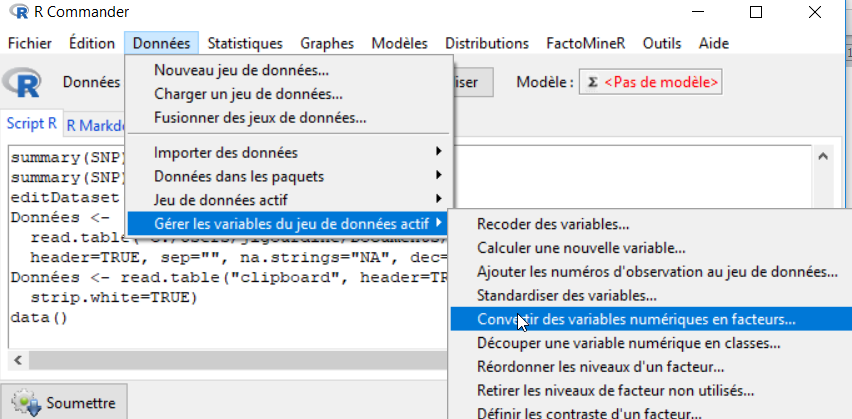
39☺☺☺ Vérifier avec une méthode de votre choix que la variable anne2 est qualitative

40☺☺☺ A l’aide de la fonction nlevels vérifier que la variable annee2 a 19 modalités.

### Changer de type avec R Commander

41☺☺☺ Sous R Commander, copier le jeu de données data\_meta dans un nouvel objet data\_meta3 et réaliser un résumé des variables.

42☺☺☺ Tranformer la variable annee en qualitative en stockant le résultat dans une nouvelle variable annee2



43☺☺☺ Vérifier avec une méthode de votre choix que la variable anne2 est qualitative

1. Formation INRA domaine de Duclos réalisée par Jean-Luc Gourdine : [jean-luc.gourdine@inra.fr](mailto:jean-luc.gourdine@inra.fr) [↑](#footnote-ref-1)