Chapitre 6 Structurer ses tables

6.1 Pourquoi se pencher sur la structuration des tables ?

Pour bien manipuler des données, leur structuration est fondamentale.

- Qu'est ce qu'une ligne de notre table ?
- Qu'est ce qu'une colonne de notre table ?

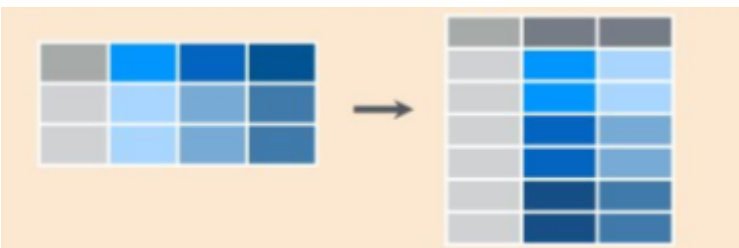
Sur une table non aggrégée (un répertoire, une table d'enquête…), la structuration naturelle est une ligne par observation (un individu, une entreprise…), une colonne par variable (âge, taille…) sur cette observation.

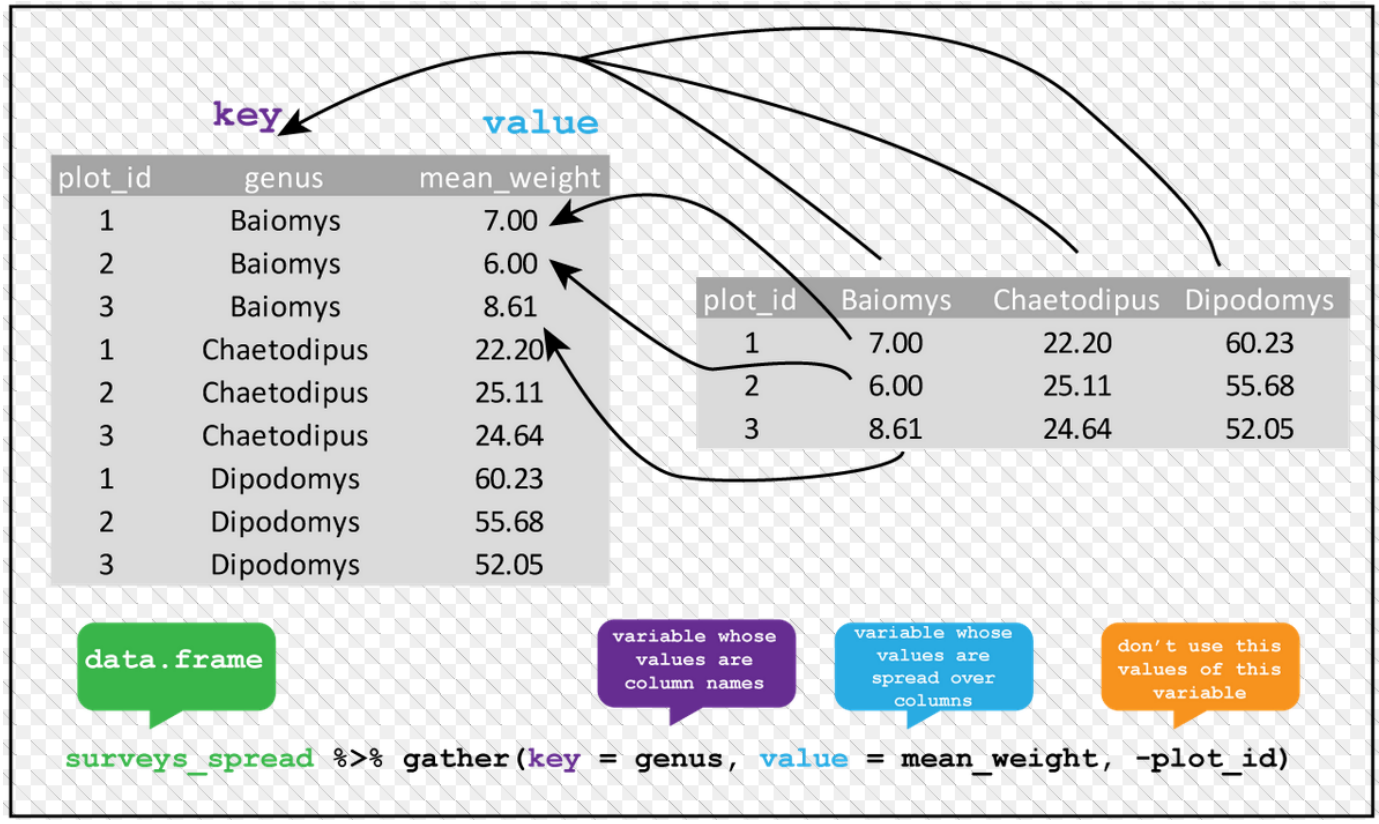
Mais dès qu'on aggrège une telle table pour construire des tables structurées par dimensions d'analyse et indicateurs, se pose toujours la question de savoir ce qu'on va considérer comme des dimensions et comme des indicateurs.

La bonne réponse, c'est que ça dépend de ce que l'on veut en faire. L'important est de pouvoir facilement passer de l'un à l'autre suivant ce que l'on doit faire. C'est l'intérêt du module *tidyr*.

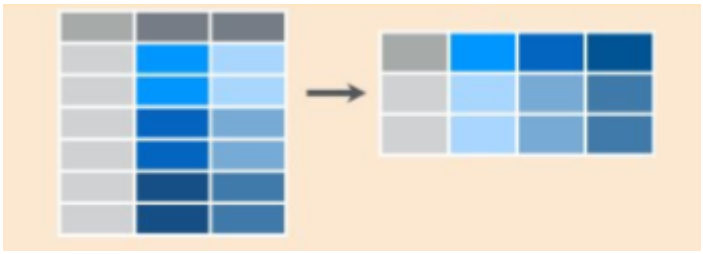
6.2 Les deux fonctions clefs de tidyr

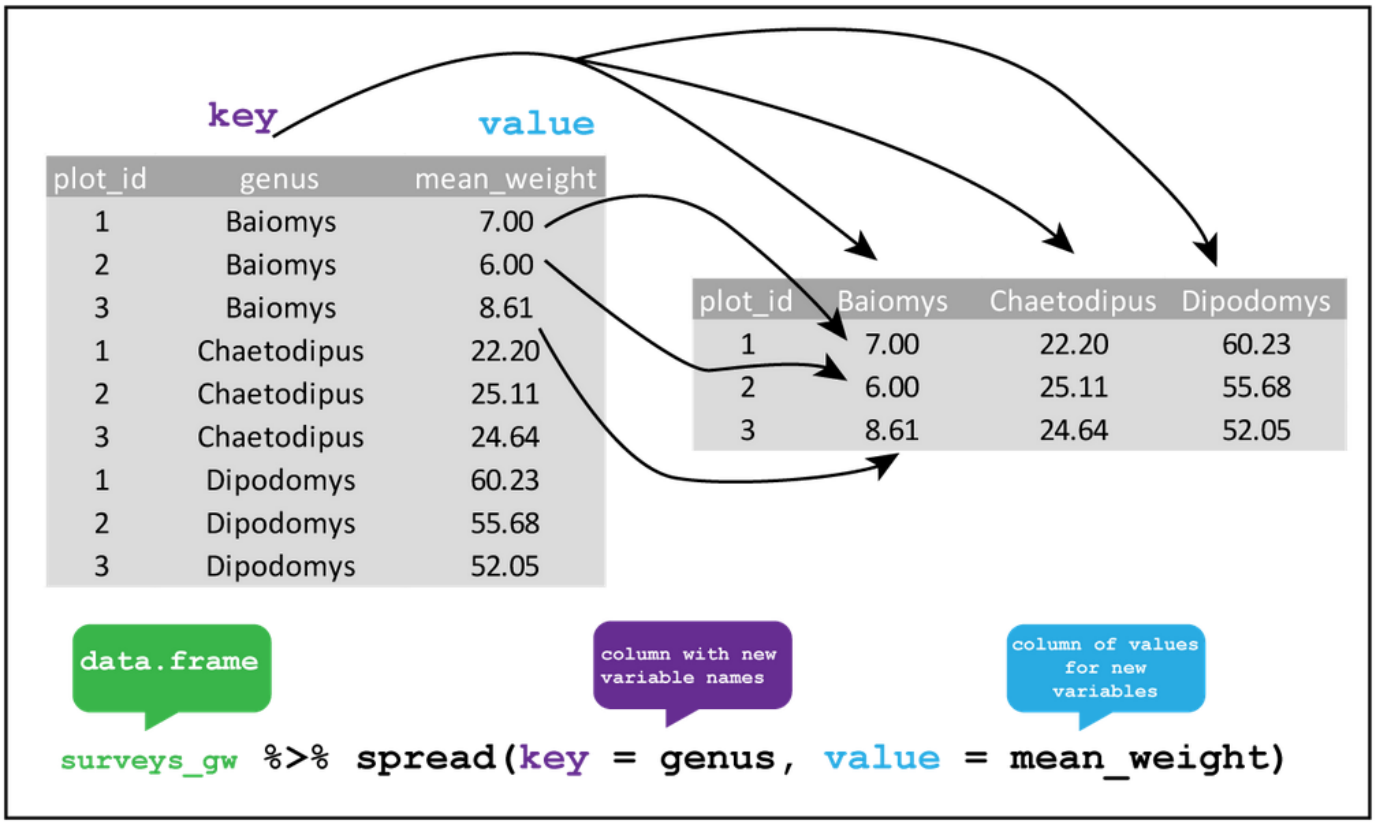
- **gather()** permet d'empiler plusieurs colonnes (correspondant à des variables quantitatives). Elles sont repérées par création d'une variable qualitative, à partir de leurs noms. Le résultat est une table au format *long*





- **spread()** fait l'inverse. Cette fonction crée autant de colonnes qu'il y a de modalités d'une variable qualitative, en remplissant chacune par le contenu d'une variable numérique. Le résultat est une table au format *large*





Un exemple : obtenir un fichier avec une ligne par région, et une colonne par année qui donne l'évolution en % de la construction neuve par rapport à l'année précédente

```
sitadel_long <- read_excel ("data/ROES_201702.xls", "AUT_REG") %>%
  mutate (ANNEE = str_sub (date, 1, 4)) %>%
  group_by (REG, ANNEE) %>%
    summarise_if (is.numeric, funs (sum (., na.rm = T))) %>%
    mutate_if (is.numeric, funs (EVO = 100 * . / lag (.) - 100)) %>%
    select (REG, ANNEE, log_AUT_EVO) %>%
  ungroup ()

sitadel_large <- sitadel_long %>%
  spread (key = ANNEE, value = log_AUT_EVO, sep = "_")

sitadel_long2 <- sitadel_large %>%
  gather (key = annee, value = log_aut_evo, -REG)
```