

Statistical modelling for biological data with R

Day 1 - Introduction to R

Marie-Pierre Etienne

<https://marieetienne.github.io>

Novembre 2019



Plan

① Préambule

② R, Rstudio and ecosystem

③ Manipulation données

④ Visualisation

⑤ Des ressources utiles

Plan

① Préambule

② R, Rstudio and ecosystem

③ Manipulation données

④ Visualisation

⑤ Des ressources utiles

Rendre à César . . .

Ce cours est essentiellement celui proposé par l'équipe de statistique d'Agrocampus Ouest dans le cadre de la formation continue *Manipulation et visualisation de données* <https://dsr-rennes.github.io>

Il s'inspire grandement de

- R for Data science (Wickham & Grolemund, 2016), <https://r4ds.had.co.nz/>



- R pour la statistique et la science des données (Cornillon et al., 2018),
<https://r-stat-sc-donnees.github.io/>



- Cours de Julien Chiquet <https://github.com/jchiquet/CourseAdvancedR>

Plan

① Préambule

② R, Rstudio and ecosystem

③ Manipulation données

④ Visualisation

⑤ Des ressources utiles

Plan

② R, Rstudio and ecosystem

L'écosystème R

Utiliser R

Utilisation de Rstudio

Bonnes pratiques

Tirer profit de la communauté

Qu'appelle-t-on R?

R est à la fois un **logiciel**, un **langage** et un **environnement** informatique dédié au calcul et à l'analyse statistique.

R est un projet open source (projet GNU)

R est un logiciel multi-plateforme (Linux, Mac, Windows)

La structure R

R est composé d'un socle (base) et de bibliothèques de fonctions thématiques regroupées sous le nom de **package**

Il est possible de connecter R avec d'autres langages : C, Fortran, Java, Javascript, Python...

Il est possible d'appeler des fonctions R depuis Matlab, Excel, SAS, SPSS...

Des connecteurs pour tous les types de bases de données : RODBC, RMySQL, ROracle, RJDBC, RMongo...

Les packages

R a été pensé comme un **langage ouvert et modulaire**.

L'ensemble des fonctionnalités de R est inclus dans des **packages**.

A l'installation de R, les packages de base sont déjà installés dans votre environnement

R permet à n'importe qui de proposer et déposer son package sur le serveur du **CRAN**. De nombreux chercheurs utilisent R donc les nouvelles méthodes sont implémentées.

Aujourd'hui, 13903 packages sont disponibles

<https://cran.r-project.org/web/packages/>

Si vous avez besoin d'une fonctionnalité spécifique, vous pouvez avoir besoin d'installer un package!

R histoire

L'ère pré-R

- 1970's développement de S au Bell labs
- 1980's développement de S-PLUS au AT&T. Lab

Les débuts

- 1993 développement de R sur le modèle de S par Robert Gentleman et Ross Ihaka au département de statistique de l'université d'Auckland.
- 1995 dépôts des codes sources sous licence GNU/GPL

Le succès

- 1997 élargissement du groupe
- 2002 la fondation R dépose ses statuts sous la présidence de Gentleman et Ihaka

Développement entièrement bénévole

- “R development core team” (12aine de personnes)
- Participation de nombreux chercheurs (\approx 13900 packages)

La communauté R

La page web de la **fondation R**

- les statuts, des liens, des références.
- <http://www.r-project.org/>

La page web du **CRAN** (Comprehensive R Arxiv Network)

- binaires d'installation, packages, documentations, . . .
- <http://cran.r-project.org/>

La **conférence** annuelle des utilisateurs de R :

- l'édition 2019 se déroule début juillet en France (Toulouse)
- l'édition 2009 s'est déroulée à Agrocampus Ouest

Depuis 2012, il existe une version française : "les rencontres R".

- l'édition 2018 s'est déroulée à Agrocampus Ouest

The **R journal** propose des articles sur :

- de nouvelles extensions, des applications, des actualités.
- <http://journal.r-project.org/>

Plan

② R, Rstudio and ecosystem

L'écosystème R

Utiliser R

Utilisation de Rstudio

Bonnes pratiques

Tirer profit de la communauté

Installer R

Allez sur <http://cran.r-project.org/> et choisissez votre système d'exploitation



[CRAN](#)
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

[About R](#)
[R Homepage](#)
[The R Journal](#)

[Software](#)
[R Sources](#)
[R Binaries](#)
[Packages](#)
[Other](#)

[Documentation](#)
[Manuals](#)
[FAQs](#)
[Contributed](#)

The Comprehensive R Archive Network

[Download and Install R](#)

Precompiled binary distributions of the base system and contributed packages. **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

[Source Code for all Platforms](#)

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2019-03-11, Great Truth) [R-3.5.3.tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#)

[Questions About R](#)

- If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

What are R and CRAN?

R is 'GNU S', a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc. Please consult the [R project homepage](#) for further information.

CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R. Please use the CRAN [mirror](#) nearest to you to minimize network load.

Submitting to CRAN

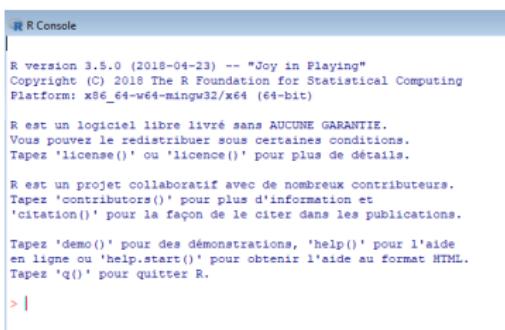
To "submit" a package to CRAN, check that your submission meets the [CRAN Repository Policy](#) and then use the [web form](#).

If this fails, upload to <http://CRAN.R-project.org/incoming/> and send an email to CRAN-submissions@R-project.org following the policy. Please do not attach submissions to emails, because this will clutter up the mailboxes of half a dozen people.

Note that we generally do not accept submissions of precompiled binaries due to security reasons. All binary distribution listed above are compiled by selected maintainers, who are in charge for all binaries of their platform, respectively.

Lancement de R

Cliquer sur l'icône  pour lancer R, une fenêtre, appelée **Console** s'ouvre :



R version 3.5.0 (2018-04-23) -- "Joy in Playing"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R est un logiciel libre livré sans AUCUNE GARANTIE.
Vous pouvez le redistribuer sous certaines conditions.
Tapez 'license()' ou 'licence()' pour plus de détails.

R est un projet collaboratif avec de nombreux contributeurs.
Tapez 'contributors()' pour plus d'information et
'citation()' pour la façon de le citer dans les publications.

Tapez 'demo()' pour des démonstrations, 'help()' pour l'aide
en ligne ou 'help.start()' pour obtenir l'aide au format HTML.
Tapez 'q()' pour quitter R.

> |

R attend une instruction : ceci est indiqué par > en début de ligne. Cette instruction doit être validée par **Entrée** pour être exécutée.

- instruction correcte, R exécute et redonne la main >
- instruction incomplète, R retourne +, il faut alors compléter l'instruction ou sortir avec **Echap**

R et RStudio

R

- Element de base : c'est le cœur de l'outil
- Utilisation "bas niveau"
- Convivialité réduite

RStudio

RStudio est un IDE (integrated development environment) pour R.

Principaux avantages de convivialité:

- Editeur de code intégré
- Débogage
- Outil de visualisation de l'environnement de travail

Plan

② R, Rstudio and ecosystem

L'écosystème R

Utiliser R

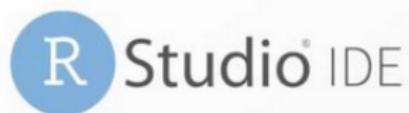
Utilisation de Rstudio

Bonnes pratiques

Tirer profit de la communauté

Installer RStudio

Allez sur <https://www.rstudio.com/products/RStudio/>

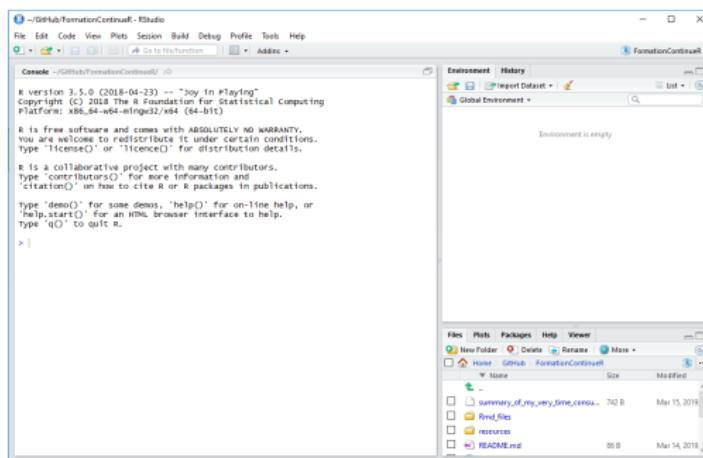


- Data analysis scripts
- Interactive web applications
- Documents
- Reports
- Graphs
- More



Lancement de RStudio

Cliquer sur l'icône  pour lancer RStudio



RStudio est divisé en 3 (4) quadrants :

- Console
- Espace de travail, historique, importation, ...
- Visualisation (graphiques), aide, ...
- Editeur de texte, de codes, ... (4ème quadrant à insérer 

Console et environnement de travail

L'environnement de travail permet de :

- faire des calculs

1+1

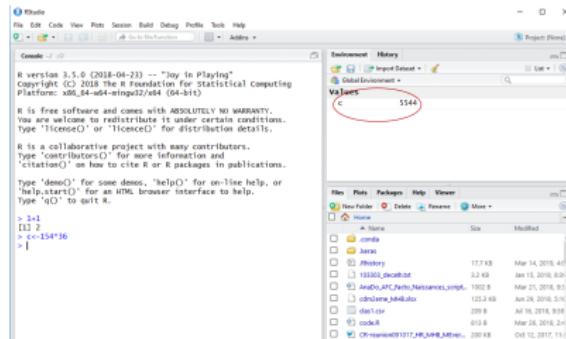
```
## [1] 2
```

- stocker les calculs dans des **variables** ou **objets**

c<-154*36

c

```
## [1] 5544
```



Plan

② R, Rstudio and ecosystem

L'écosystème R

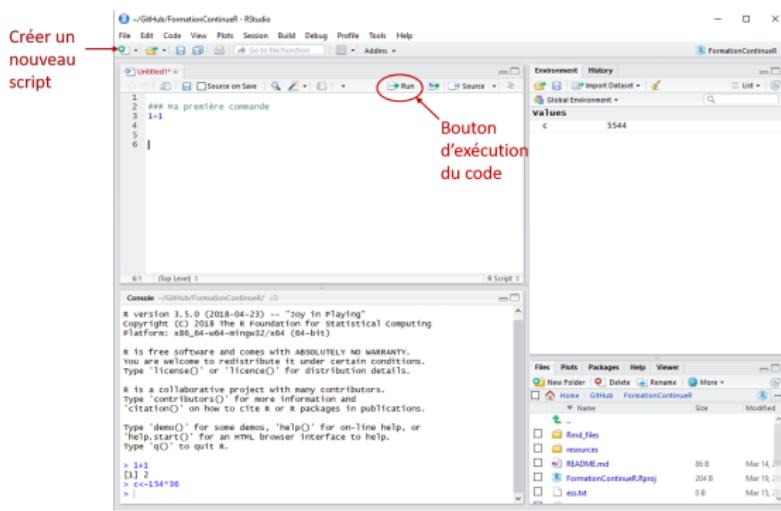
Utiliser R

Utilisation de Rstudio

Bonnes pratiques

Tirer profit de la communauté

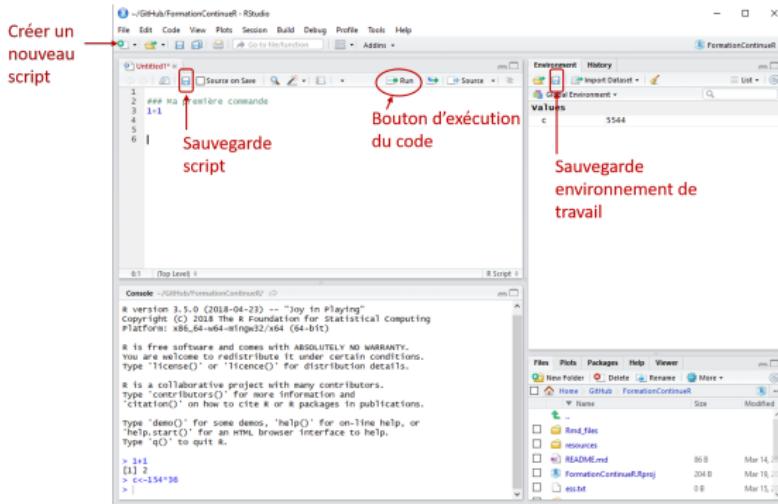
Faire un script



Le code se tape dans la fenêtre de script et s'exécute directement.

- `#` pour insérer des commentaires
- **Shift+Alt+k** pour les raccourcis clavier

Faire un script



- Possibilité de sauvegarder le script (.R),
- Possibilité de sauvegarder l'environnement de travail (.RData).

Les Projets dans RStudio

RStudio dispose d'une fonctionnalité très pratique pour organiser son travail en différents projets.

L'idée est de réunir tous les fichiers, documents (données, scripts, ...) relatifs à un même projet dans un répertoire dédié.

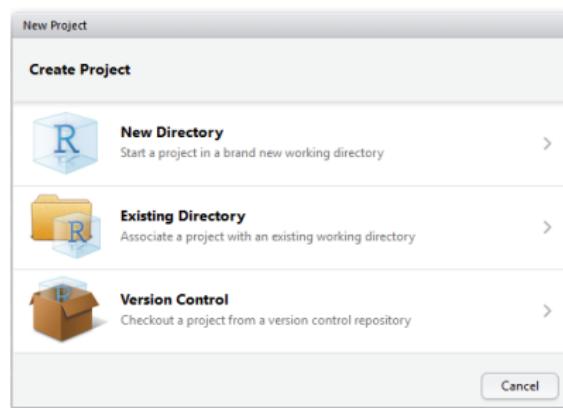
Les avantages

- Facilite l'accès aux fichiers de données à importer : le répertoire de travail de R est défini comme étant le répertoire du projet
- L'onglet Files de l'interface permet de naviguer dans les fichiers du projet
- Les objets créés (et sauvegardés dans le fichier .Rdata) lors d'une précédente séance de travail sont chargés en mémoire
- Les scripts ouverts lors d'une précédente séance de travail sont automatiquement ouverts

Lorsque l'on ouvre un projet RStudio, on revient à l'état de notre projet tel qu'il était la dernière fois que l'on a travaillé dessus.

Créer un nouveau projet

Pour créer un projet, File -> New Project



Choisir *Existing directory* ou *New directory* selon l'existence ou non du dossier du projet. Créer ou sélectionner le dossier, puis cliquer sur *Create project*

Une fois le projet créé, son nom est affiché dans un petit menu déroulant en haut à droite de l'interface de RStudio (FormationContinueR - (menu qui permet de passer facilement d'un projet à un autre)).

Plan

② R, Rstudio and ecosystem

L'écosystème R

Utiliser R

Utilisation de Rstudio

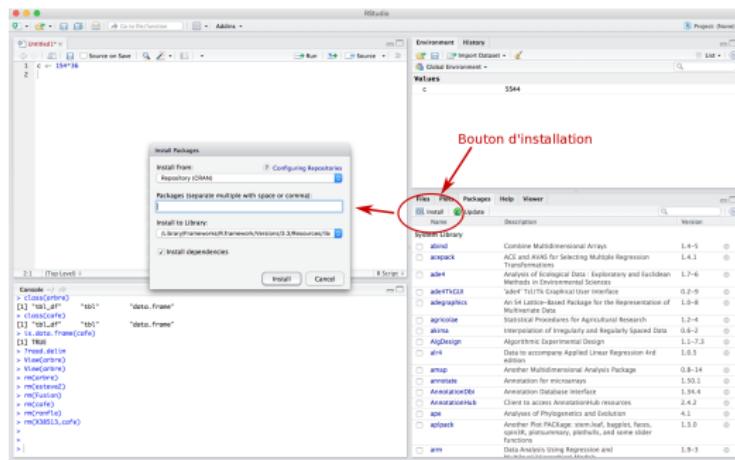
Bonnes pratiques

Tirer profit de la communauté

Installer un package

Le but de l'installation est de télécharger et placer au bon endroit les codes R contenus dans le package

- En mode interactif, directement à l'aide d'un site "miroir":



- En mode ligne de commande

```
install.packages('ggplot2')
```

- Vérifier si le package est disponible et l'installer uniquement si besoin

Charger le package

Il ensuite charger le package dans votre environnement lorsque l'on souhaite l'utiliser.

- En mode interactif

en cochant la case du package dans le menu Packages de RStudio

- dans la console : fonction library ou require

```
library('FactoMineR')
```

Le mode console est le mode compatible avec la production de documents.

Exercice

- Installer le package dplyr
- Charger le package dplyr

L'aide

Pour obtenir de l'aide :

- en ligne de code dans la console

```
help(dplyr) # lance l'aide associée à la commande dplyr
help.start() # lance l'aide HTML
```

- en mode interactif

Cliquer sur le nom du package dans l'onglet Packages

Name	Description	Version
<input type="checkbox"/> doBy	Groupwise Statistics, LSmeans, Linear Contrasts, Utilities	4.6-2
<input checked="" type="checkbox"/> dplyr	A Grammar of Data Manipulation	0.7.6
...		

A Grammar of Data Manipulation

Documentation for package 'dplyr' version 0.7.6

- [DESCRIPTION file](#)
- [User guides, package vignettes and other documentation.](#)

Help Pages

ABCDEFGHIJKLMNOPQRSTUVWXYZ

dplyr-package dplyr: a grammar of data manipulation

L'aide

La plupart des packages propose une documentation intitulée **vignettes**, décrivant l'utilisation du package

Pour accéder à la (aux) vignette(s) d'un package, taper dans la console :

```
browseVignettes("dplyr")
```

Il y a toujours des exemples que l'on peut exécuter directement.

Plan

① Préambule

② R, Rstudio and ecosystem

③ Manipulation données

④ Visualisation

⑤ Des ressources utiles

Choix de point de vue

Le cours présenté ici fait le choix de tirer largement partie des fonctionnalités offertes par la suite de package du tidyverse développé par Rstudio ainsi que par les possibilités graphiques de ggplot2. Il faut donc charger les packaes correspondants

```
library(tidyverse)

## -- Attaching packages ----- tidyverse 1.0.0
##   v ggplot2 3.2.1     v purrr    0.3.2
##   v tibble   2.1.3     v dplyr    0.8.3
##   v tidyr    0.8.3     v stringr  1.4.0
##   v readr    1.3.1     vforcats  0.4.0

## -- Conflicts ----- tidyverse_conflicts()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()   masks stats::lag()

library(ggplot2)
```

Bonnes pratiques pour les fichiers

- Préférer un format csv, éviter les formats propriétairesxlsx
- Eviter les accents, espaces et caractères spéciaux dans les noms de fichier et les noms de variables
- Garder une trace de toutes les opérations de pré traitement des données (typiquement dans un fichier RMarkdown)

Plan

③ Manipulation données

Importation d'un jeu de données

Manipulation de données - R base

Un tour dans le tidyverse

Opérations sur les individus (les lignes)

Opération sur les variables (les colonnes)

Des traitements par sous groupes

Sauvegarder des tables de données

Importer un fichier en mode interactif

File - Import Dataset - From Text (base) - choix du fichier

Exercice

- Importer le fichier SamaresEq.txt dans la variable SamaresEq_base et vérifier la classe de l'objet obtenu
- Importer le fichier SamaresEq.txt avec l'option From text (readr) SamaresEq_readr et vérifier la classe de l'objet obtenu

Importer un fichier en ligne de commande

```
mon_fichier <- "../Datasets/SamaresEq.txt"
SamaresEq_base <- read.table(file = mon_fichier, sep = " ", header = TRUE, n=3)

##   Site NomSite Distance Arbre Poids Surface Largeur Longueur Cir
## 1    4 Gornies      6.3     1 11.0 0.73816  0.427  2.31502
## 2    4 Gornies      6.3     1 15.0 0.69958  0.489  2.15873
## 3    4 Gornies      6.3     1 14.6 0.75847  0.444  2.27595
```

- file indique le nom complet du fichier (potentiellement avec le chemin d'accès)
- sep décrit le type de séparateur utilisé dans le fichier
- dec décrit le signe pour les décimales (utile pour les fichiers en Français)

La version ligne de commande préférable dans une optique de science reproductible.

Importer un fichier en ligne de commande

En utilisant le package `readr`

```
library(readr)
mon_fichier <- "../../Datasets/SamaresEq.txt"
SamaresEq_readr <- read_delim(file = mon_fichier, delim = " ")

## Parsed with column specification:
## cols(
##   Site = col_double(),
##   NomSite = col_character(),
##   Distance = col_double(),
##   Arbre = col_double(),
##   Poids = col_double(),
##   Surface = col_double(),
##   Largeur = col_double(),
##   Longueur = col_double(),
##   CircArbre = col_double()
## )
```

SamaresEq_readr

Importer depuis une url en ligne de commande

Même approche pour une importation depuis une url : Le fichier decathlon

```
monFichier <- 'https://husson.github.io/img/decathlon.csv'  
decathlon <- read.table(file = monFichier, header = TRUE, sep = ';')
```

Exercice

- Créer un fichier RMarkdown nommé Formation_R_exercices et insérer le code pour créer le tableau vins_base contenant le jeu de données (<http://factominer.free.fr/livre/vins.csv>) en utilisant la commande `read.table`

Solution

- Créer un fichier RMarkdown nommé Formation_R_exercices.Rmd et insérer le code pour créer le tableau vins_base contenant le jeu de données vin en utilisant la commande `read.table`
 - New File – R Markdown
 - Modifier auteur et titre
 - save as Formation_R_exercices.Rmd

Solution

- Insérer le code pour créer le tableau `vins_readr` contenant le jeu de données <http://factominer.free.fr/livre/vins.csv> en utilisant la commande `read_delim` du package `readr`.

```
library(readr)
vins_readr <- read_delim("http://factominer.free.fr/livre/vins.csv"
                           delim = ";",
                           escape_double = FALSE, locale = locale(),
                           trim_ws = TRUE)

## Warning: Missing column names filled in: 'X1' [1]

## Parsed with column specification:
## cols(
##   X1 = col_character(),
##   `1S-Michaud` = col_double(),
##   `2S-Renaudie` = col_double(),
##   `3S-Trotignon` = col_double(),
##   `4S-Buisse` = col_double(),
##   `5S-BuisseCristal` = col_double(),
##   `6S-AubSilou` = col_double()
```

Vérifier l'importation

- Afficher le fichier dans son ensemble (éviter pour des fichiers longs)

```
## Parsed with column specification:  
## cols(  
##   Site = col_double(),  
##   NomSite = col_character(),  
##   Distance = col_double(),  
##   Arbre = col_double(),  
##   Poids = col_double(),  
##   Surface = col_double(),  
##   Largeur = col_double(),  
##   Longueur = col_double(),  
##   CircArbre = col_double()  
## )
```

```
SamaresEq_readr <- read_delim(file = 'https://marieetienne.github.i  
SamaresEq_readr
```

```
SamaresEq_base
```

Importation d'un jeu de données à l'aide une requête SQL (base de données)

- Connexion à une base SQL.

```
library(RODBC)
# Liste les tables de la base de données connectée
sqlTables(connect_base, tableType = "TABLE")
# Liste les champs de la table DonneesTotales
sqlColumns(connect_base, sqtable = "DonneesTotales")
```

- Executer une requête sur la base

```
# execute une requete SQL
OtoYFT <- sqlQuery( channel = connect_base,
                      query =
"
SELECT * FROM DonneesTotales
WHERE (DonneesTotales.ProblemeSp = 'Ok' AND DonneesTotales.REC_Sp='
AND DonneesTotales.Otolithe = 'OT')
")
# Liste les champs de la table DonneesTotales
```

Plan

③ Manipulation données

Importation d'un jeu de données

Manipulation de données - R base

Un tour dans le tidyverse

Opérations sur les individus (les lignes)

Opération sur les variables (les colonnes)

Des traitements par sous groupes

Sauvegarder des tables de données

La structure des tableaux de données dans R

Un tableau de données est un objet `data.frame`.

- Connaître les dimensions d'un tableau

```
dim(SamaresEq_base)
```

```
## [1] 2380     9
```

- Connaître les noms de variables

```
colnames(SamaresEq_base)
```

```
## [1] "Site"        "NomSite"      "Distance"    "Arbre"       "Poids"  
## [7] "Largeur"     "Longueur"     "CircArbre"
```

- Accéder à une variable

```
head(SamaresEq_base$Poids, n=5)
```

```
## [1] 11.0 15.0 14.6 13.8 12.2
```

- Accéder à une ligne

```
SamaresEq_base[2, ]
```

Plan

③ Manipulation données

Importation d'un jeu de données

Manipulation de données - R base

Un tour dans le tidyverse

Opérations sur les individus (les lignes)

Opération sur les variables (les colonnes)

Des traitements par sous groupes

Sauvegarder des tables de données

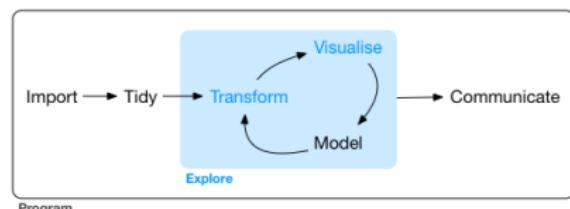
Présentation

Le tidyverse est ensemble de packages développés pour *{faciliter} la manipulation de données dans R.*

- *Installer le package tidyverse.*

```
install.packages("tidyverse")
```

D'après les créateurs dans Wickham & Grolemund (2016)



Charger le packahe tidyverse

```
library(tidyverse)
```

Objectif : Obtenir un code plus lisible

Plan

③ Manipulation données

Importation d'un jeu de données

Manipulation de données - R base

Un tour dans le tidyverse

Opérations sur les individus (les lignes)

Opération sur les variables (les colonnes)

Des traitements par sous groupes

Sauvegarder des tables de données

Selectionner les individus qui satisfont une condition

Selectionner les individus qui satisfont une condition

filter

```
SamaresEq_base %>% as_tibble %>% filter( Surface > 0.75) -> Grand_s
class(Grand_samares)

## [1] "tbl_df"     "tbl"        "data.frame"

Grand_samares

## # A tibble: 1,517 x 9
##   Site NomSite Distance Arbre Poids Surface Largeur Longueur C
##   <int> <fct>     <dbl> <int> <dbl>    <dbl>    <dbl>    <dbl>
## 1     4 Gornies     6.3     1  14.6    0.758    0.444    2.28
## 2     4 Gornies     6.3     1  13.8    0.779    0.474    2.26
## 3     4 Gornies     6.3     1  13.3    0.756    0.447    2.15
## 4     4 Gornies     6.3     1  13.7    0.782    0.474    2.24
## 5     4 Gornies     6.3     1  17.1    0.898    0.474    2.50
## 6     4 Gornies     6.3     1  15.3    0.790    0.483    2.13
## 7     4 Gornies     6.3     1  18.3    0.833    0.411    2.64
## 8     4 Gornies     6.3     2  12.2    0.827    0.446    2.71
## 9     4 Gornies     6.3     2  15.8    0.774    0.437    2.71
## # ... with 1,508 more rows, and 1 more variable:
## #   C: character
```

Exercice

- Sélectionner toutes les observations du site Gornies
- Sélectionner toutes les observations correspondantes à des largeurs supérieures à 0.45 mais le longueur inférieure à 2.32
- Sélectionner toutes les observations qui ne proviennent ni du site Gornies ni du site StEtienne.

Solution

filter

Plan

③ Manipulation données

Importation d'un jeu de données

Manipulation de données - R base

Un tour dans le tidyverse

Opérations sur les individus (les lignes)

Opération sur les variables (les colonnes)

Des traitements par sous groupes

Sauvegarder des tables de données

Sélectionner certaines variables

select

Pour modifier les variables présentes dans le jeu de données

- Ne garder que la variable NomSite

```
SamaresEq_base %>% select(NomSite)
```

- Supprimer la variable Site et Arbre

```
SamaresEq_base %>% select(-Site, -Arbre)
```

Exercice

select

-A partir de la table de données SamaresEq_base, créer une table SamaresEq_base_gornies qui contient les information concernant la largeur et la longueur de chaque samare uniquement pour les arbres du Site Gornies

Solution

`select`

-A partir de la table de données SamaresEq_base, créer une table SamaresEq_base_gornies qui contient les information concernant la largeur et la longueur de chaque samare uniquement pour les arbres du Site Gornies

```
SamarEsEq_base %>% filter( NomSite = 'Gornies') %>% select(-Site,  
head(SamarEsEq_base))
```

Créer des nouvelles variables

`mutate`

```
SamaresEq_readr %>%
  mutate(dispersion = Surface / Poids,
        log_disp = log( dispersion )) -> SamaresEq_disp
SamaresEq_disp %>% select(-Site, -Arbre, -Distance, -CircArbre)

## # A tibble: 2,380 x 7
##   NomSite Poids Surface Largeur Longueur dispersion log_disp
##   <chr>    <dbl>    <dbl>    <dbl>    <dbl>      <dbl>    <dbl>
## 1 Gornies  11     0.738   0.427    2.32     0.0671   -2.70
## 2 Gornies  15     0.700   0.489    2.16     0.0466   -3.07
## 3 Gornies  14.6   0.758   0.444    2.28     0.0520   -2.96
## 4 Gornies  13.8   0.779   0.474    2.26     0.0564   -2.87
## 5 Gornies  12.2   0.620   0.404    2.03     0.0509   -2.98
## 6 Gornies  11.1   0.537   0.339    2.04     0.0484   -3.03
## 7 Gornies  14.2   0.676   0.422    2.23     0.0476   -3.04
## 8 Gornies  15.6   0.708   0.452    2.15     0.0454   -3.09
## 9 Gornies  15.1   0.613   0.377    2.15     0.0406   -3.20
## 10 Gornies 13.3   0.756   0.447   2.15     0.0569   -2.87
```

Exercice

mutate

- A partir de la table de données SamaresEq_readr, ajouter une variable larg_x_long contenant le produit de la largeur et de la longueur et une colonne diff_surf qui calcule la différence entre la variable précédemment définie et la surface présente dans la table.

Solution

`select`

- A partir de la table de données SamaresEq_readr, ajouter une variable `larg_x_long` contenant le produit de la largeur et de la longueur et une variable `diff_surf` qui calcule la différence entre la variable précédemment définie et la surface présente dans la table.

```
SamaresEq_readr %>% mutate(larg_x_long = Largeur * Longueur,
                                diff_surf = larg_x_long - Surface) %>%
  select(-Site, -Arbre, -Distance, -CircArbre)
```

On peut limiter l'affichage

```
SamaresEq_readr %>% mutate(larg_x_long = Largeur * Longueur,
                                diff_surf = larg_x_long - Surface) %>%
```

```
  select(-Site, -Arbre, -Distance, -CircArbre) %>%
  print(n = 3)
```

```
## # A tibble: 2,380 x 7
##   NomSite Poids Surface Largeur Longueur larg_x_long diff_surf
##   <chr>    <dbl>    <dbl>    <dbl>    <dbl>      <dbl>      <dbl>
## 1 50/111
```

Résumer des variables

summarise

- Calculer des moyennes

```
SamaresEq_readr %>%
  summarise( longueur_m = mean(Longueur, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   longueur_m
##       <dbl>
## 1      2.49
```

- Calculer le nombre d'observations, les médianes pour plusieurs variables

```
SamaresEq_readr %>%
  summarise_at( vars(Largeur, Longueur), funs(n(), median))
```

```
## Warning: funs() is soft deprecated as of dplyr 0.8.0
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
```

Résumer des variables

summarise

- Calculer les moyennes de toutes les variables quantitatives

```
SamaresEq_readr %>%
  summarise_if(is.numeric, mean, na.rm=TRUE)

## # A tibble: 1 x 8
##   Site Distance Arbre Poids Surface Largeur Longueur CircArbre
##   <dbl>     <dbl> <dbl>  <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
## 1  3.98     8.99  10.9   26.6    0.823   0.451    2.49   26.2
```

Exercice

summarise

- Calculer la moyenne et l'écart-type pour les variables Surface et disp.

Exercice

summarise

- Calculer la moyenne et l'écart-type pour les variables Surface et dispersion.

```
SamarèsEq_disp %>%
  summarise_at( vars(Surface, dispersion),  funs( sd, mean), na.rm = TRUE)

## # A tibble: 1 x 4
##   Surface_sd dispersion_sd Surface_mean dispersion_mean
##       <dbl>        <dbl>      <dbl>        <dbl>
## 1     0.179      0.00948     0.823      0.0330
```

Plan

③ Manipulation données

Importation d'un jeu de données

Manipulation de données - R base

Un tour dans le tidyverse

Opérations sur les individus (les lignes)

Opération sur les variables (les colonnes)

Des traitements par sous groupes

Sauvegarder des tables de données

Calculer des moyennes pour chaque groupe

group_by

- Calculer la dispersion moyenne pour chaque site

```
SamaresEq_disp %>% group_by( NomSite) %>%  
  summarise( Surface_m = mean (Surface)) %>%  
  print(n = 3)
```

```
## # A tibble: 7 x 2  
##   NomSite   Surface_m  
##   <chr>       <dbl>  
## 1 Gornies     0.849  
## 2 Grenou_1    0.810  
## 3 Grenou_2    0.853  
## # ... with 4 more rows
```

Calculer des effectifs pour chaque groupe

group_by

- Calculer les effectifs par Site et par arbre

```
SamaresEq_disp %>% group_by( NomSite, Arbre ) %>%
  summarise( n_obs = n() ) %>% print(n = 3)
```

```
## # A tibble: 119 x 3
## # Groups:   NomSite [7]
##   NomSite Arbre n_obs
##   <chr>    <dbl> <int>
## 1 Gornies     1     20
## 2 Gornies     2     20
## 3 Gornies     3     20
## # ... with 116 more rows
```

Exercice

- Pour chaque site et chaque arbre, donner le nombre de samares échantillonés et leur poids moyen.
- Pour chaque site, donner le nombre d'arbres échantillonnés.

Solution

- Pour chaque site et chaque arbre, donner le nombre de samares échantillonés et leur poids moyen.

```
SamaresEq_disp %>% group_by( NomSite, Arbre ) %>%
  summarise( n_obs = n(), poids_m = mean(Poids)) %>% print(n = 3)
```

```
## # A tibble: 119 x 4
## # Groups:   NomSite [7]
##   NomSite Arbre n_obs  poids_m
##   <chr>    <dbl> <int>    <dbl>
## 1 Gornies     1     20     13.4
## 2 Gornies     2     20     11.5
## 3 Gornies     3     20     24.6
## # ... with 116 more rows
```

- Pour chaque site, donner le nombre d'arbres échantillonnés.

```
SamaresEq_disp %>% group_by( NomSite) %>% summarise( n_Arbre = n_
```

```
## # A tibble: 7 x 2
##   NomSite  n_Arbre
```

Plan

③ Manipulation données

Importation d'un jeu de données

Manipulation de données - R base

Un tour dans le tidyverse

Opérations sur les individus (les lignes)

Opération sur les variables (les colonnes)

Des traitements par sous groupes

Sauvegarder des tables de données

Sauvegarder dans un format texte

```
write_csv
```

```
write_csv(SamaresEq_disp, path = ".../.../Datasets/SamaresEq_disp.csv")
```

Plan

① Préambule

② R, Rstudio and ecosystem

③ Manipulation données

④ Visualisation

⑤ Des ressources utiles

Plan

4 Visualisation

Avec les fonctionnalités de base de R

Une présentation générale de ggplot2

Un graphique basique

Personnalisation

Pour aller plus loin

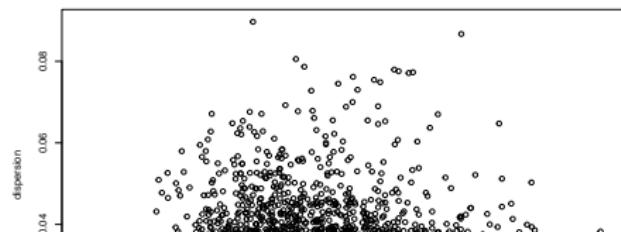
Exercice

Scatter plot

```
head(SamaresEq_disp)
```

```
## # A tibble: 6 x 11
##   Site NomSite Distance Arbre Poids Surface Largeur Longueur Ci
##   <dbl> <chr>     <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 4 Gornies     6.3    1  11  0.738  0.427  2.32
## 2 4 Gornies     6.3    1  15  0.700  0.489  2.16
## 3 4 Gornies     6.3    1 14.6  0.758  0.444  2.28
## 4 4 Gornies     6.3    1 13.8  0.779  0.474  2.26
## 5 4 Gornies     6.3    1 12.2  0.620  0.404  2.03
## 6 4 Gornies     6.3    1 11.1  0.537  0.339  2.04
## # ... with 2 more variables: dispersion <dbl>, log_disp <dbl>
```

```
plot(dispersion~Surface, data = SamaresEq_disp)
```



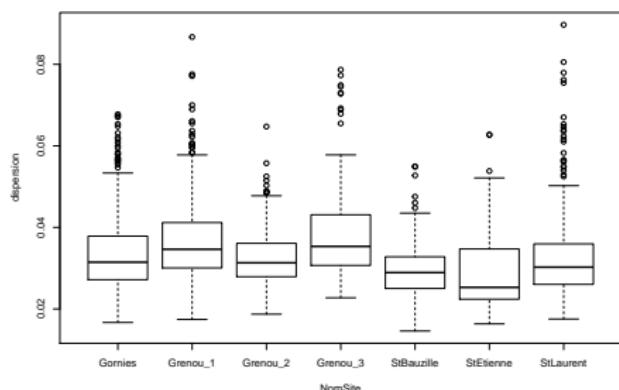
Boîtes à moustaches

```
plot(dispersion~NomSite, data=SamaresEq_disp)
```

la variable NomSite n'est pas un facteur.

```
SamaresEq_disp %>% mutate(NomSite = as.factor(NomSite)) -> SamaresEq_disp
```

```
plot(dispersion~NomSite, data=SamaresEq_disp)
```



Plan

④ Visualisation

Avec les fonctionnalités de base de R

Une présentation générale de ggplot2

Un graphique basique

Personnalisation

Pour aller plus loin

Exercice

Les idées derrière ggplot2

ggplot2 propose de construire des graphiques en suivant une grammaire graphique. Hadley Wickham créateur du package a écrit [ce papier](#) pour expliquer son approche.

Un graphique est composé

- d'un jeu de données
- dont on veut représenter certains aspects
- en utilisant une forme adaptée

Les idées derrière ggplot2

Un graphique est composé

- d'un jeu de données : `ggplot`
- dont on veut représenter certains aspects : `aes`
- en utilisant une forme adaptée : `geom` et `stat`

La côte est difficile mais la vue vaut le détour !

Les idées derrière ggplot2

Un graphique est composé

- d'un jeu de données : `ggplot`
- dont on veut représenter certains aspects : `aes`
- en utilisant une forme adaptée : `geom` et `stat`

La côte est difficile mais la vue vaut le détour !

Liens utiles

Le site de référence

Elegant graphics for data analysis.

Une galerie inspirante

les extensions de ggplot2

Cheatsheet

Plan

4 Visualisation

Avec les fonctionnalités de base de R

Une présentation générale de ggplot2

Un graphique basique

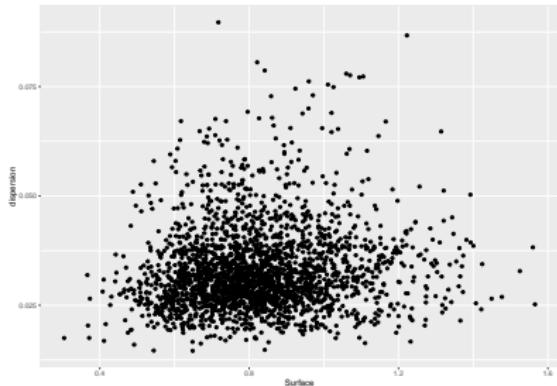
Personnalisation

Pour aller plus loin

Exercice

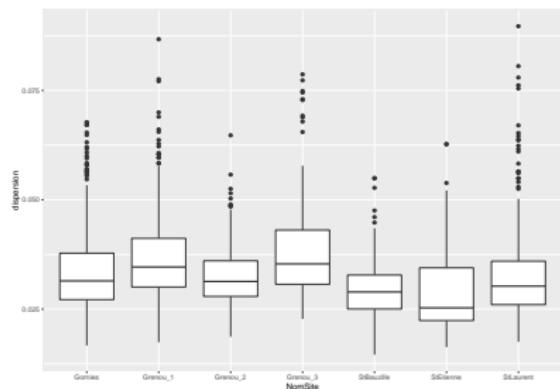
Scatter plot

```
library(ggplot2)
ggplot(data = SamaresEq_disp) + aes( x = Surface, y = dispersion) +
```



Boîtes à moustaches

```
ggplot(data = SamaresEq_disp) +  
  aes( x = NomSite, y = dispersion) +  
  geom_boxplot()
```



Plan

④ Visualisation

Avec les fonctionnalités de base de R

Une présentation générale de ggplot2

Un graphique basique

Personnalisation

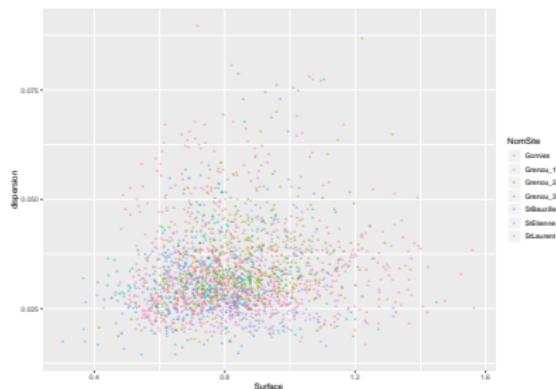
Pour aller plus loin

Exercice

Ajouter des couleurs

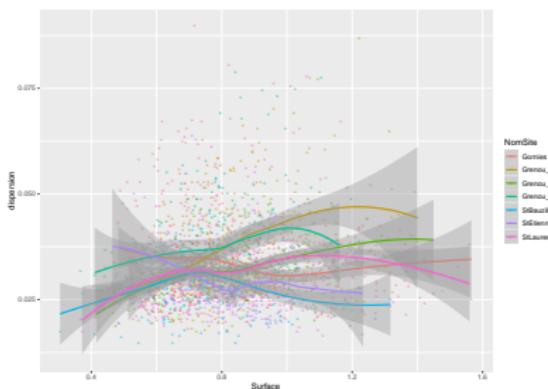
Ajouter de la couleur pour le Nom du Site et changer le symbole.

```
ggplot(data = SamaresEq_disp) +  
  aes( x = Surface, y = dispersion, col = NomSite) +  
  geom_point( shape = 'a')
```



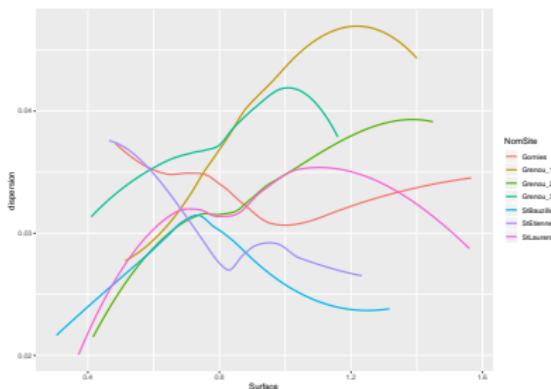
Ajouter une tendance

```
ggplot(data = SamaresEq_disp) +  
  aes( x = Surface, y = dispersion, col = NomSite) +  
  geom_point( shape = 'a') +  
  geom_smooth()  
  
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



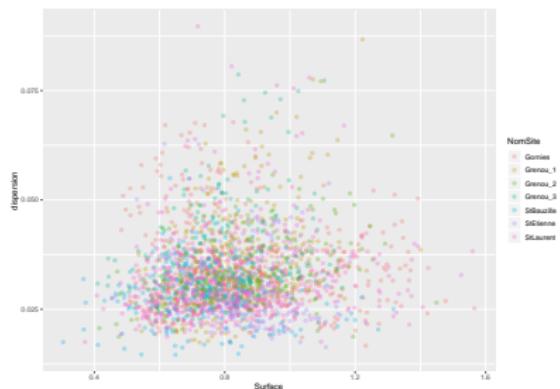
Ajouter une tendance seule

```
ggplot(data = SamaresEq_disp) +  
  aes( x = Surface, y = dispersion, col = NomSite) +  
  geom_smooth(se = FALSE)  
  
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



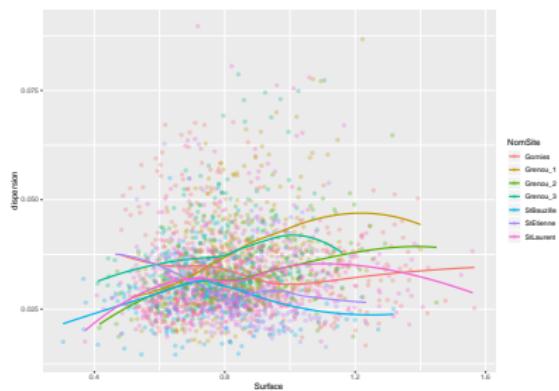
Jouer avec la transparence

```
ggplot(data = SamaresEq_disp) +  
  aes( x = Surface, y = dispersion, col = NomSite) +  
  geom_point( alpha= 0.3)
```



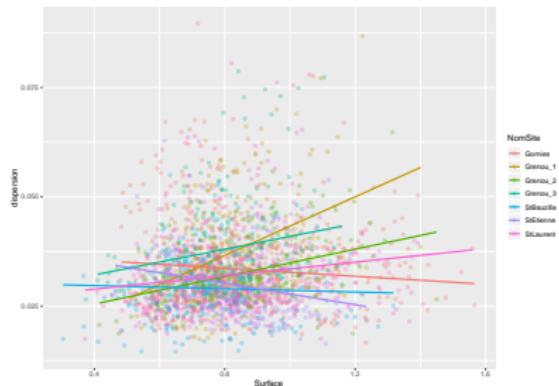
Tous ensemble

```
ggplot(data = SamaresEq_disp) +  
  aes( x = Surface, y = dispersion, col = NomSite) +  
  geom_point( alpha= 0.3) +  
  geom_smooth( se = FALSE)  
  
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Avec une tendance linéaire

```
ggplot(data = SamaresEq_disp) +  
  aes( x = Surface, y = dispersion, col = NomSite) +  
  geom_point( alpha= 0.3) +  
  geom_smooth(method = 'lm', se = FALSE)
```



Plan

④ Visualisation

Avec les fonctionnalités de base de R

Une présentation générale de ggplot2

Un graphique basique

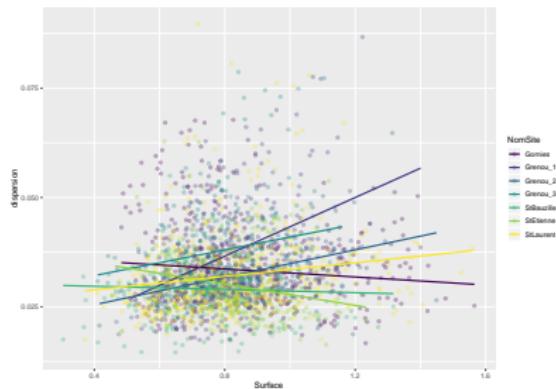
Personnalisation

Pour aller plus loin

Exercice

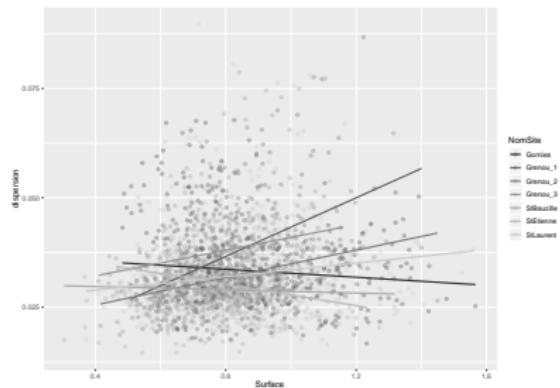
Une palette compatible daltonisme

```
ggplot(data = SamaresEq_disp) +  
  aes( x = Surface, y = dispersion, col = NomSite) +  
  geom_point( alpha= 0.3) +  
  geom_smooth(method = 'lm', se = FALSE) +  
  scale_color_viridis_d()
```



En noir et blanc

```
ggplot(data = SamaresEq_disp) +  
  aes( x = Surface, y = dispersion, col = NomSite) +  
  geom_point( alpha= 0.3) +  
  geom_smooth(method = 'lm', se = FALSE) +  
  scale_color_grey()
```

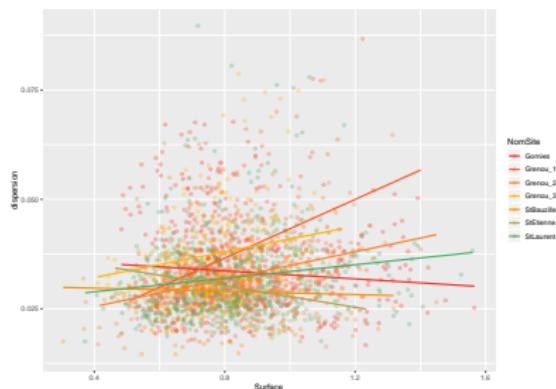


Utiliser ses propres couleurs

Un site utile pour la spécification des couleurs

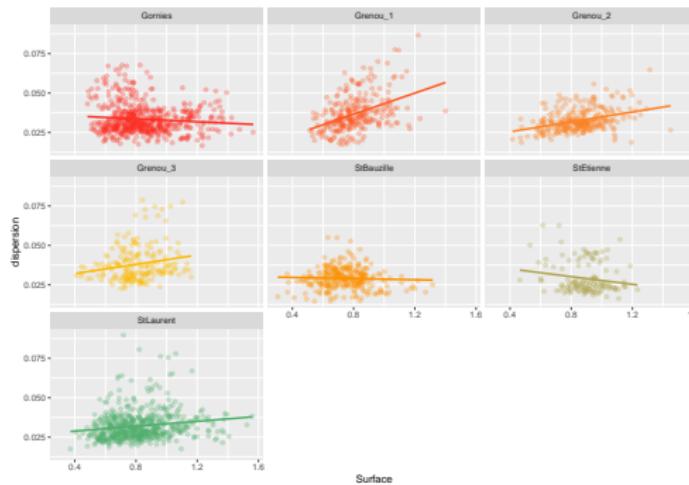
```
palette <- c('#FF2F24', '#FF5B24', '#FF8324', '#FFBb00', '#FF9100',
```

```
ggplot(data = SamaresEq_disp) +  
  aes( x = Surface, y = dispersion, col = NomSite) +  
  geom_point( alpha= 0.3) +  
  geom_smooth(method = 'lm', se = FALSE) +  
  scale_color_manual(values = palette)
```



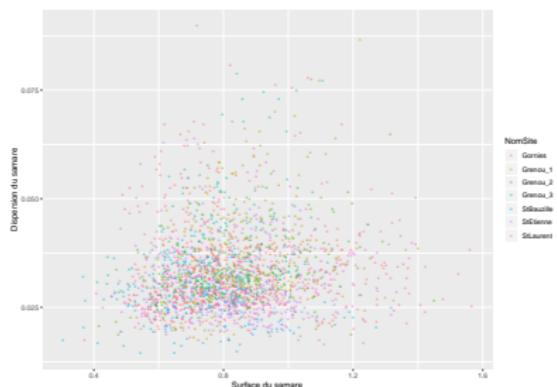
Un graphique par site

```
ggplot(data = SamaresEq_disp) +  
  aes( x = Surface, y = dispersion, col = NomSite) +  
  facet_wrap(~NomSite) + theme( legend.position = 'none' ) +  
  geom_point( alpha= 0.3) +  
  geom_smooth(method = 'lm', se = FALSE) +  
  scale_color_manual(values = palette)
```



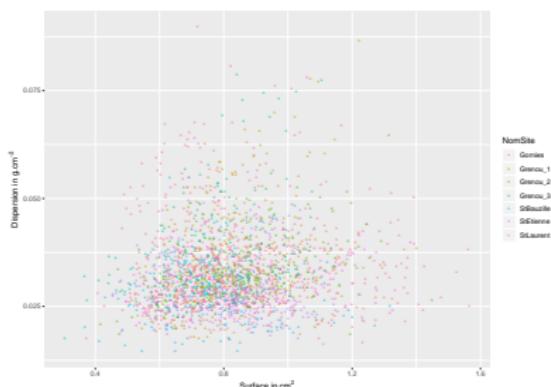
Nommer les axes

```
ggplot(data = SamaresEq_disp) +  
  aes( x = Surface, y = dispersion, col = NomSite) +  
  geom_point( shape = 'a') +  
  labs(y = 'Dispersion du samare', x = 'Surface du samare')
```



Intégrer des exposants et des indices dans les noms des axes

```
ggplot(data = SamaresEq_disp) +  
  aes( x = Surface, y = dispersion, col = NomSite) +  
  geom_point( shape = 'a') +  
  labs(y = expression ("Dispersion in"~g.cm^{-2}), x = expression
```



Plan

④ Visualisation

Avec les fonctionnalités de base de R

Une présentation générale de ggplot2

Un graphique basique

Personnalisation

Pour aller plus loin

Exercice

Exercice

Etude du rendement en fonction de l'indice de diversité,

Données issues de Kirwan et al. (2014)

- Charger le fichier `Biomass_diversity.csv` dans un tableau nommé `biomass`.

La variable `H` est l'indice de diversité de Shannon et `HARV_YIELD` le rendement de la parcelle.

- Quelle est le type de la variable `YEAR` ?
- Créer une variable qualitative (un facteur) `Year_fact`.
- Représenter la variabilité des rendements en fonction des pays, et des couples pays-années.
- Faire un graphique du rendement en fonction de l'indice de diversité
- Colorier les données selon l'année et indiquer par des symboles différents les différents pays.
- Ajuster une droite de régression par pays
- Modifier le nom des axes pour une publication en Français.

Solution

Etude du rendement en fonction de l'indice de diversité,

Données issues de Kirwan et al. (2014)

- Charger le fichier `Biomass_diversity.csv` dans un tableau nommé `biomass`.

```
biomass <- readr::read_csv(file = '../..../Datasets/Biomass_diversity.csv')

## Parsed with column specification:
## cols(
##   COUNTRY = col_character(),
##   YEAR = col_double(),
##   PLOT = col_double(),
##   G = col_double(),
##   L = col_double(),
##   HARV_YIELD = col_double(),
##   H = col_double()
## )
```

- Quelle est le type de la variable `YEAR` ?

`biomass`

Solution

- Représenter la variabilité des rendements en fonction des pays, et des couples pays-années.

```
p1 <- biomass %>% ggplot() + aes(y = HARV_YIELD, x = COUNTRY) + geom_boxplot(aes(fill = Year_fact)) + theme(text = element_text(size=8))
p2 <- biomass %>% ggplot() + aes(y = HARV_YIELD, x = COUNTRY) + geom_boxplot(aes(fill = Year_fact)) + theme(text = element_text(size=8))
library(ggpubr)

## Loading required package: magrittr

##
## Attaching package: 'magrittr'

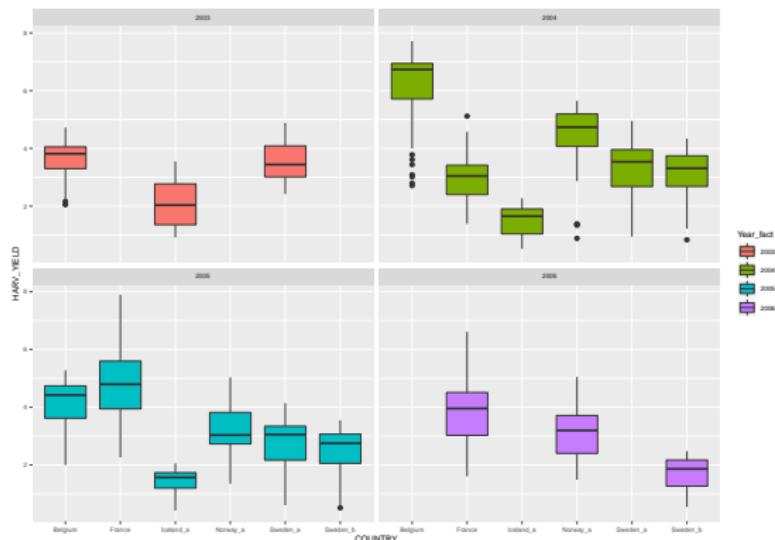
## The following object is masked from 'package:purrr':
##
##      set_names

## The following object is masked from 'package:tidyverse':
##
##      extract
```

Solution

- Représenter la variabilité des rendements en fonction des pays, et des couples pays-années.

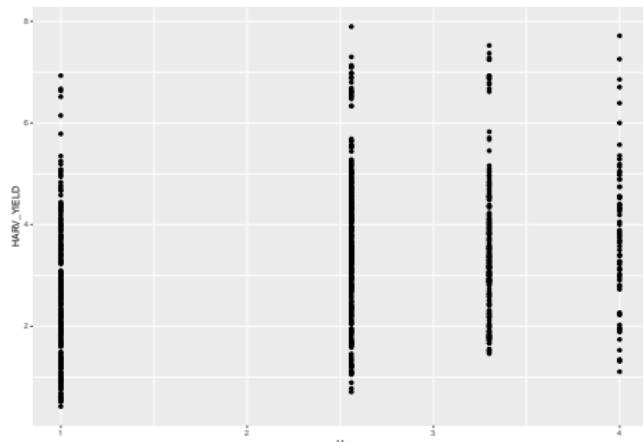
```
biomass %>% ggplot() + aes(y = HARV_YIELD, x = COUNTRY) +
  facet_wrap(~Year_fact) +
  geom_boxplot(aes(fill = Year_fact)) + theme( text = element_text
```



Solution

- Faire un graphique du rendement en fonction de l'indice de diversité

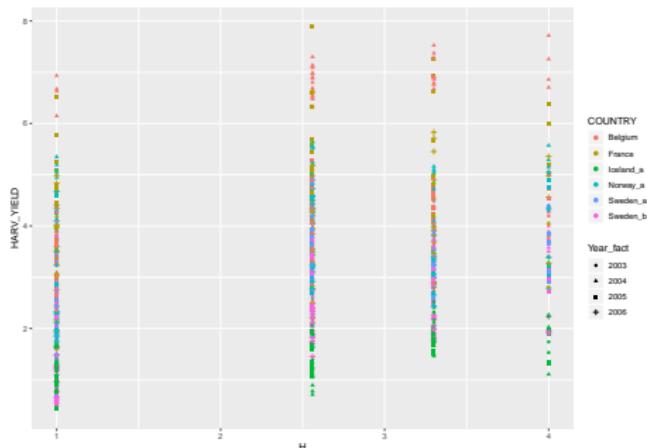
```
biomass %>% ggplot() + aes(y = HARV_YIELD, x = H) + geom_point()
```



Solution

- Colorer les données selon le pays et indiquer par des symboles différents les différentes années.

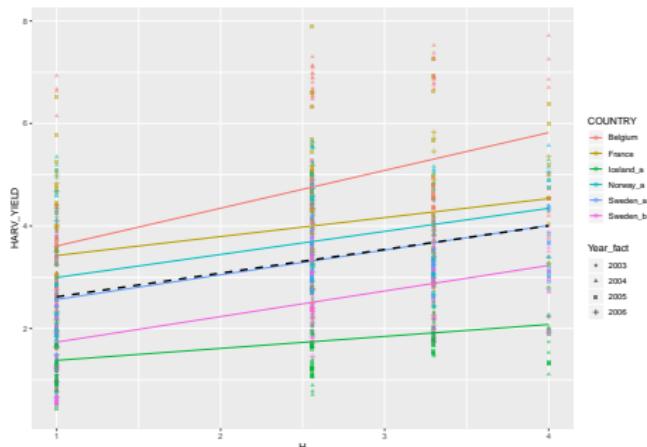
```
biomass %>% ggplot() +  
  aes(y = HARV_YIELD, x = H ) +  
  geom_point(aes(col = COUNTRY, shape = Year_fact))
```



Solution

- Ajuster une droite de régression par pays

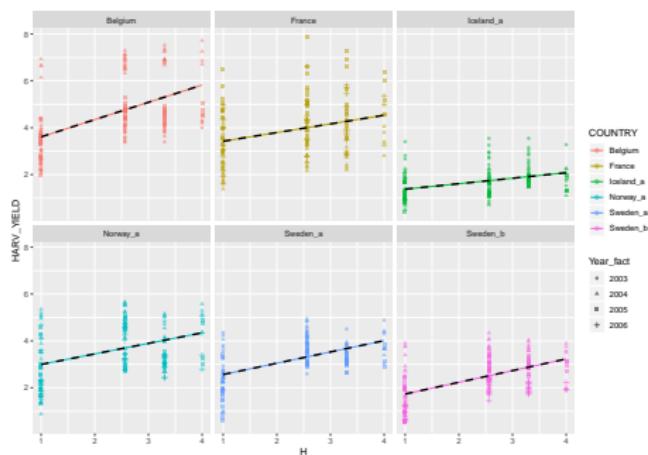
```
biomass %>% ggplot() + aes(y = HARV_YIELD, x = H) +  
  geom_point( aes(col = COUNTRY, shape = Year_fact ), alpha = 0.5 )  
  geom_smooth(method="lm", se= F, size = 0.5, aes(col = COUNTRY, gr  
  geom_smooth(method = 'lm',size = 1, linetype = 'dashed', colour =
```



Solution

- Ajuster une droite de régression par pays

```
biomass %>% ggplot() + aes(y = HARV_YIELD, x = H) +
  geom_point( aes(col = COUNTRY, shape = Year_fact ), alpha = 0.5 )
  geom_smooth(method="lm", se= F, size = 0.5, aes(col = COUNTRY, gr
p +   geom_smooth(method = 'lm',size = 1, linetype = 'dashed', col
```

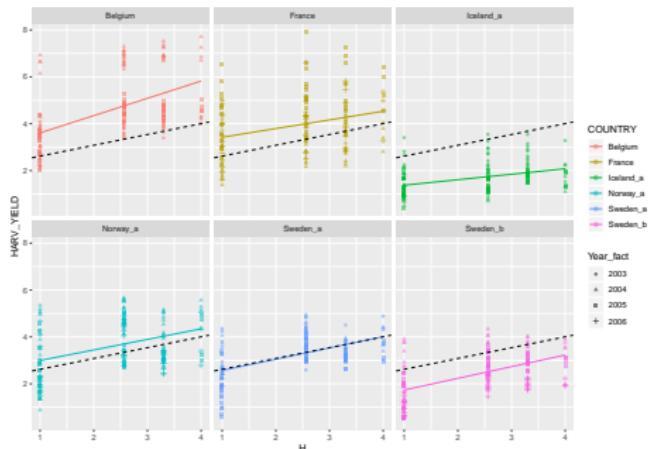


Solution

- Ajuster une droite de régression par pays

```
reg_coef <- coef(lm(HARV_YIELD ~ H, data = biomass))
```

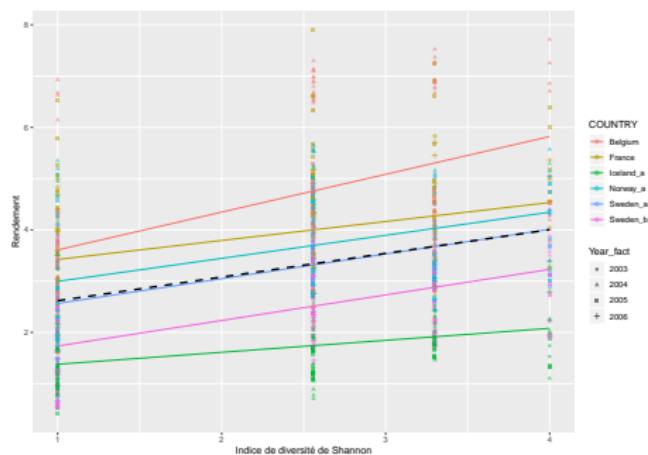
```
p + geom_abline( intercept = reg_coef[1], slope = reg_coef[2], line
```



Solution

- Modifier le nom des axes pour une publication en Français.

```
biomass %>% ggplot() + aes(y = HARV_YIELD, x = H) +  
  geom_point( aes(col = COUNTRY, shape = Year_fact ), alpha = 0.5 )  
  geom_smooth(method="lm", se= F, size = 0.5, aes(col = COUNTRY, group = COUNTRY))  
  geom_smooth(method = 'lm',size = 1, linetype = 'dashed', colour = "black")  
  labs( x = 'Indice de diversité de Shannon', y = 'Rendement' )
```



Plan

① Préambule

② R, Rstudio and ecosystem

③ Manipulation données

④ Visualisation

⑤ Des ressources utiles

Les anti sèches de RStudio

R base

RStudio

RMarkdown

Importation

Manipulation

Visualisation

Des livres

- A Language and Environment for Statistical Computing (R Core Team, 2017),
<https://www.R-project.org/>



- R for Data science (Wickham & Grolemund, 2016), <https://r4ds.had.co.nz/>



- R pour la statistique et la science des données (Cornillon et al., 2018),
<https://r-stat-sc-donnees.github.io/>



Références

- Cornillon, P.-A., Guyader, A., Husson, F., Jégou, N., Josse, J., Klutchnikoff, N., ... Thieurmel, B. (2018). *R pour la statistique et la science des données*. Presses universitaires de Rennes.
- Kirwan, L., Connolly, J., Brophy, C., Baadshaug, O. H., Belanger, G., Black, A., ... others. (2014). The agrodiversity experiment: Three years of data from a multisite study in intensively managed grasslands. *Ecology, 2014, Vol. 95, Num. 9, P. 2680-2680.*
- R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Wickham, H. (2014). *Advanced r*. CRC Press. Retrieved from <http://adv-r.had.co.nz/>
- Wickham, H., & Grolemund, G. (2016). *R for data science: Import, tidy, transform, visualize, and model data*. " O'Reilly Media, Inc.".