

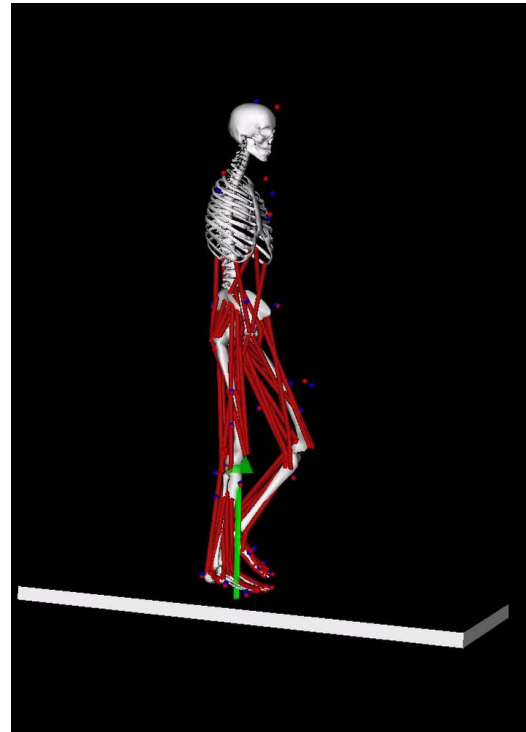
Calcul numérique : Exploration de données à l'aide d'outils informatiques

Fabien Leboeuf (pôle MPR, CHU de Nantes)

Ingénieur de recherche

Mon environnement

Le laboratoire d'analyse du mouvement (depuis 2009)



Staff du labo

- Médecin MPR
- Ingénieur
- kiné



Autour du labo

- Éducateur APA
- Unité d'investigation clinique

Beaucoup de mesures



Beaucoup de Données

Syndrome du « j'aimerais bien »...

- Vérifier la significativité statistique d'un facteur (p-value)
- Evaluer la corrélation/régression entre mes variables (R^2)
- Tracer un graphique de mes données/résultats
- Relancer des tests stats en intégrant de nouvelles données
-

Est-ce que, par hasard, tu saurais ... ?

➔ Un besoin d'autonomie



Conception d'une Etude

- Examen de la faisabilité
- Formulation de 1ere hypothèses

Redaction d'un article scientifique

- Intro
- Methode
- Resultats
- Discussion

Un bel Article = de belles figures

Ma proposition



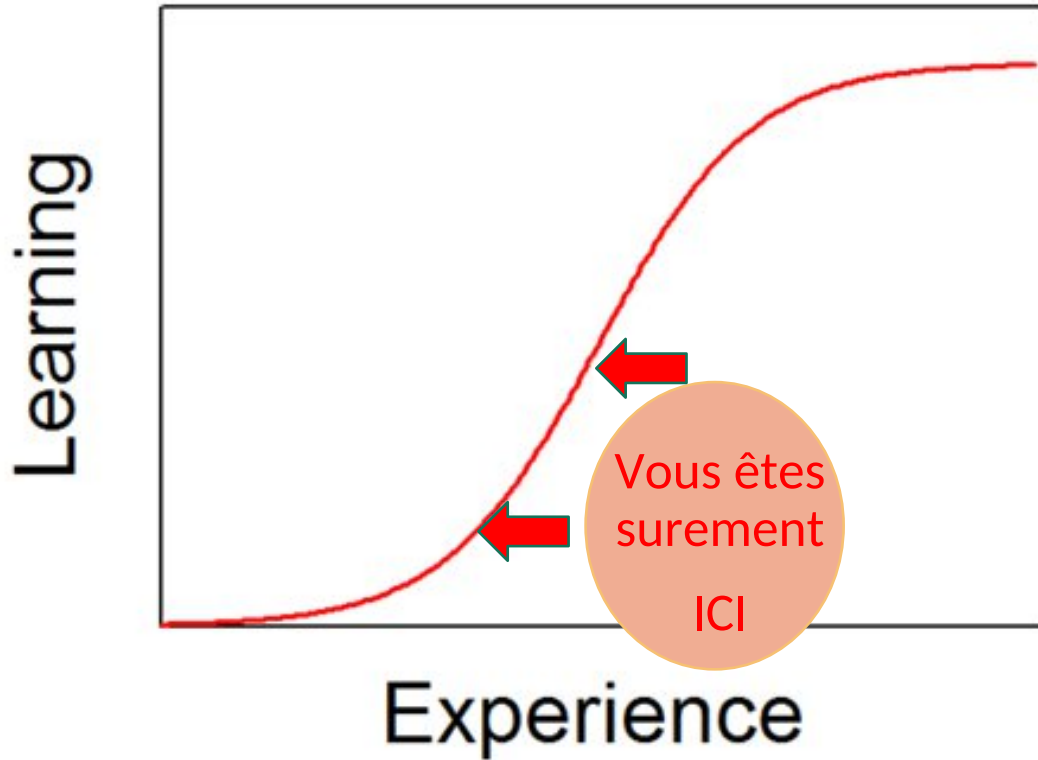
Ceci n'est ni un cours de statistiques ni un cours de programmation informatique.
C est une INITIATION

Comment **manipuler** les données

Comment **représenter** des résultats

Comment effectuer des tests stats
simples

Comment faire



Comment **manipuler** les données ★★ ★

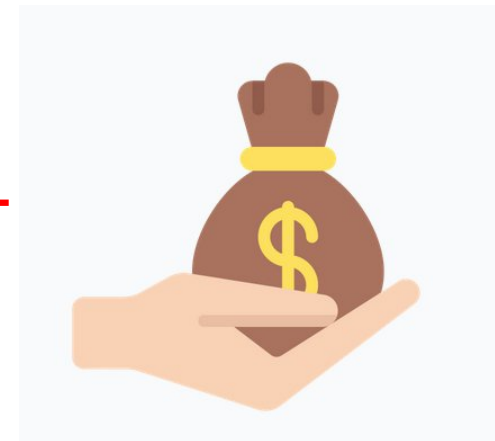
Comment **représenter** des résultats ★

Comment effectuer des tests stats ☆

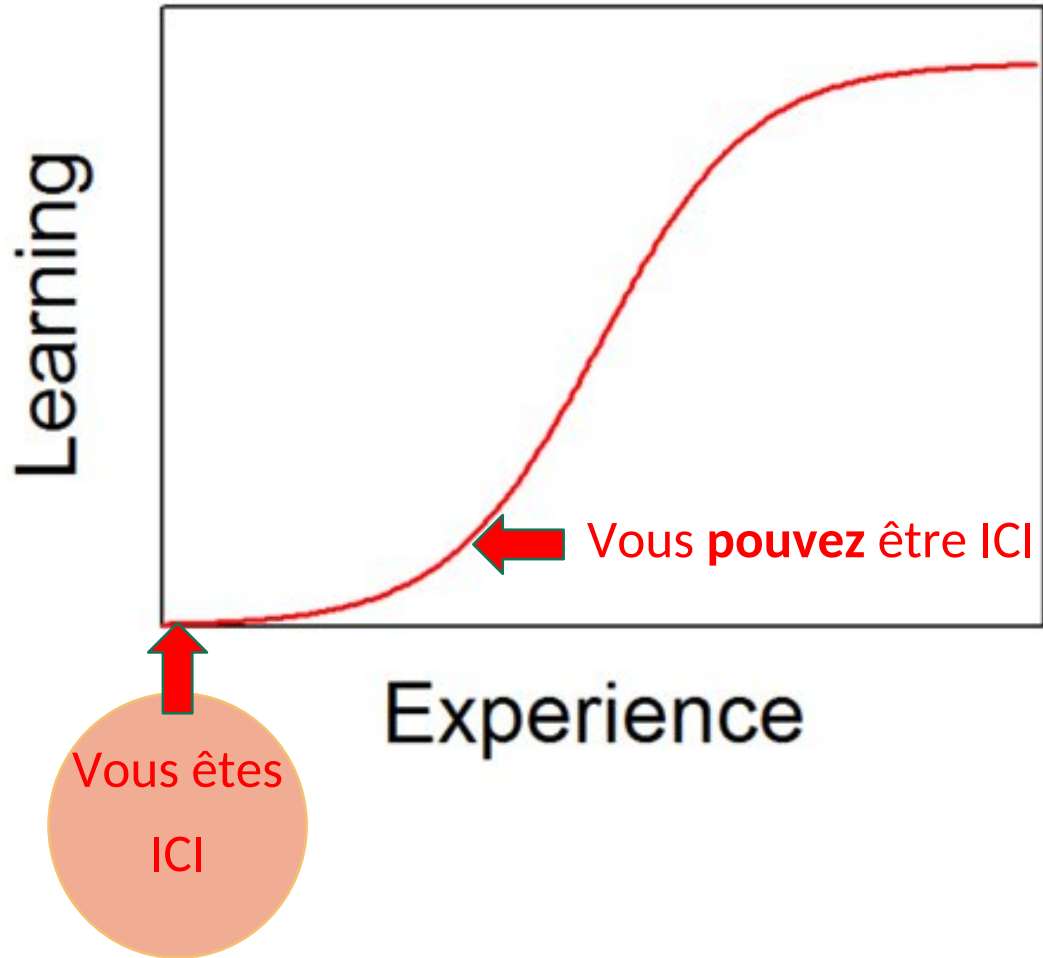
Quelques alternatives



Des outils **VALIDES** mais **PAYANT**



Un outil informatique **Verifié** et **Libre**



Comment **manipuler** les données ★★ ★

Comment **représenter** des résultats ★★ ★

Comment effectuer des tests stats ★★ ★

Progression de R

Classement en 2020

Jul 2020	Jul 2019	Change	Programming Language	Ratings	Change
1	2	▲	C	16.45%	+2.24%
2	1	▼	Java	15.10%	+0.04%
3	3		Python	9.09%	-0.17%
4	4		C++	6.21%	-0.49%
5	5		C#	5.25%	+0.88%
6	6		Visual Basic	5.23%	+1.03%
7	7		JavaScript	2.48%	+0.18%
8	20	▲▲	R	2.41%	+1.57%
9	8	▼	PHP	1.90%	-0.27%
10	13	▲	Swift	1.43%	+0.31%
11	9	▼	SQL	1.40%	-0.58%
12	16	▲▲	Go	1.21%	+0.19%
13	12	▼	Assembly language	0.94%	-0.45%
14	19	▲▲	Perl	0.87%	-0.04%
15	14	▼	MATLAB	0.84%	-0.24%

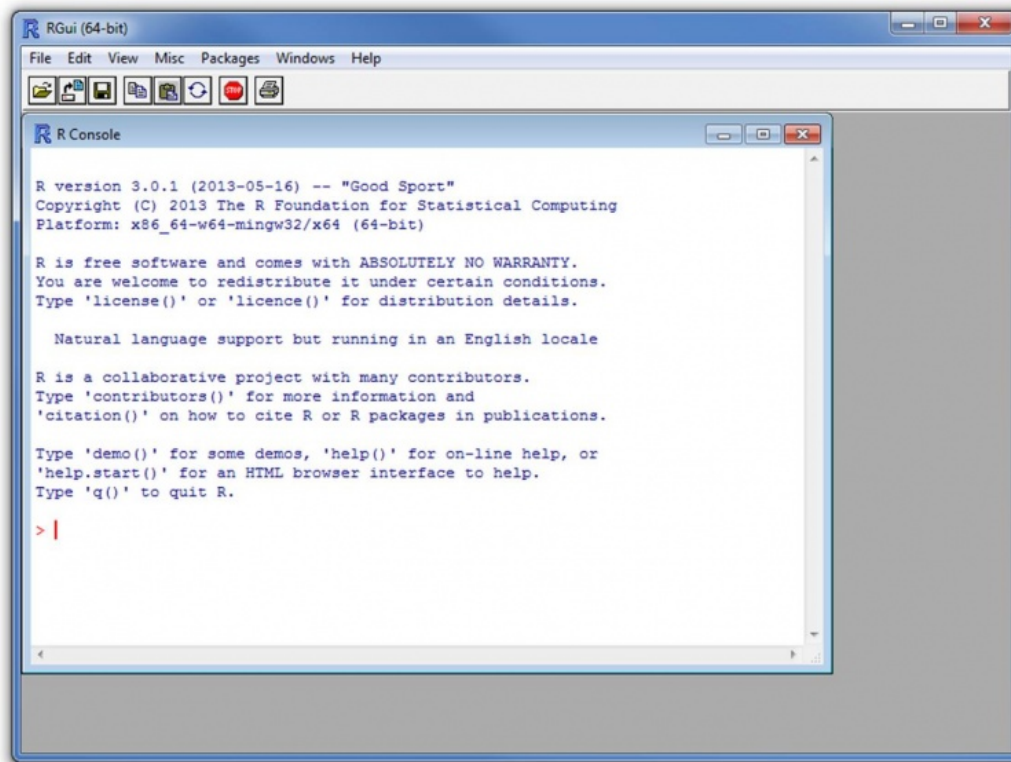
Progression due à



Un outils dédié à la manipulation de données

Progression de R

En 2009: une « console »



```
RGui (64-bit)
File Edit View Misc Packages Windows Help

R Console
R version 3.0.1 (2013-05-16) -- "Good Sport"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

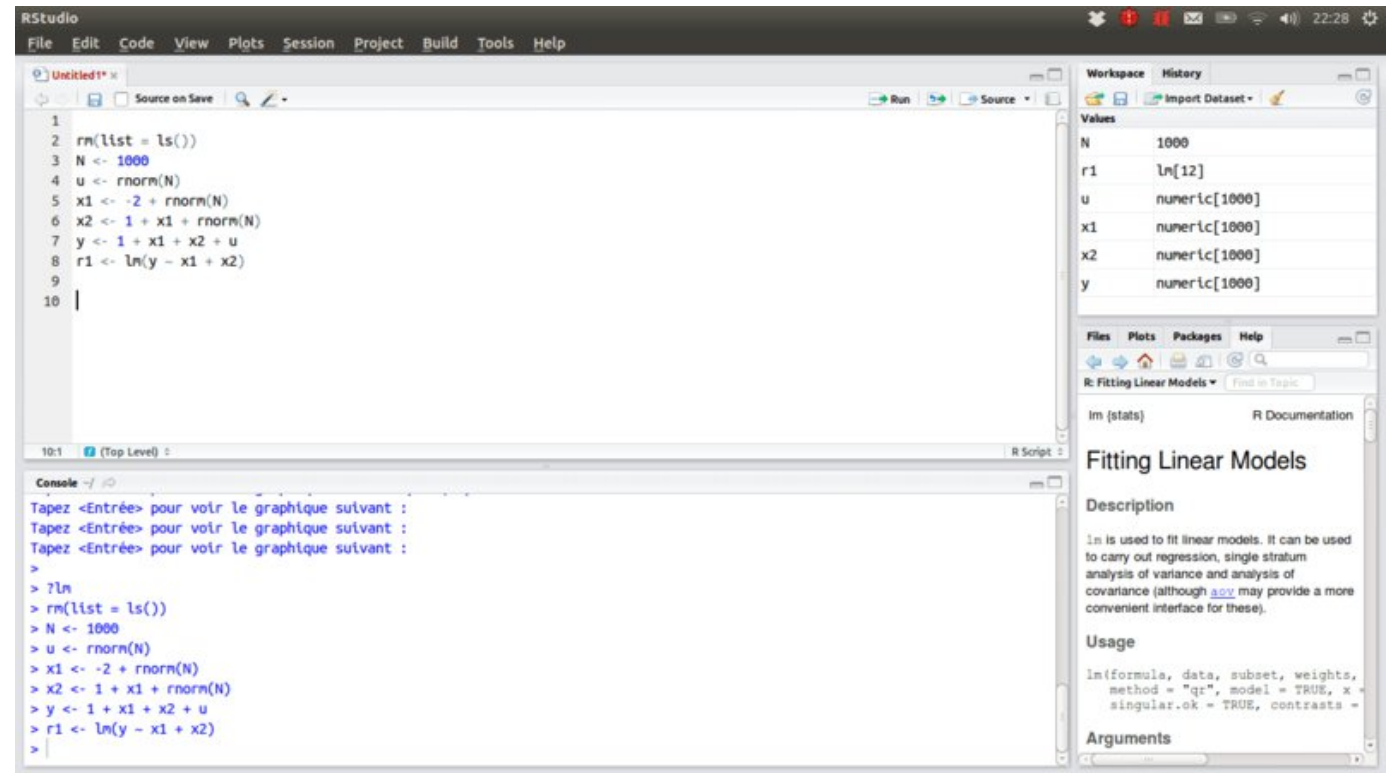
Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

En 2020: une **solution** informatique



```
RStudio
File Edit Code View Plots Session Project Build Tools Help

Untitled1*.R
1
2 rm(list = ls())
3 N <- 1000
4 u <- rnorm(N)
5 x1 <- -2 + rnorm(N)
6 x2 <- 1 + x1 + rnorm(N)
7 y <- 1 + x1 + x2 + u
8 r1 <- ln(y - x1 + x2)
9
10 |

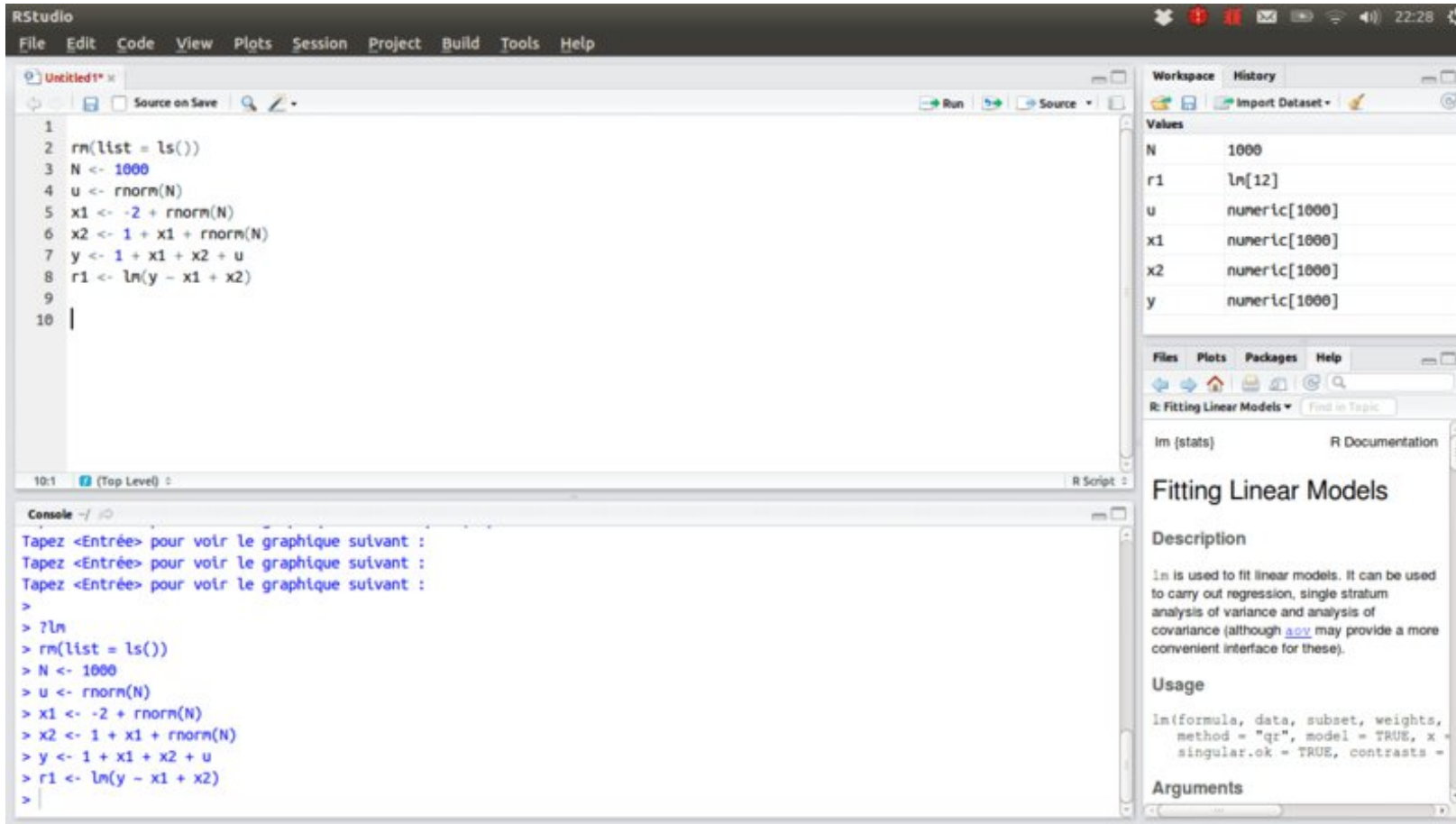
Workspace History
Values
N 1000
r1 ln[12]
u numeric[1000]
x1 numeric[1000]
x2 numeric[1000]
y numeric[1000]

Files Plots Packages Help
R: Fitting Linear Models
lm (stats) R Documentation
Fitting Linear Models
Description
lm is used to fit linear models. It can be used
to carry out regression, single stratum
analysis of variance and analysis of
covariance (although aov may provide a more
convenient interface for these).

Usage
lm(formula, data, subset, weights,
method = "qr", model = TRUE, x-
singular.ok = TRUE, contrasts =

Arguments
```

OUI... il y a du CODE



```
1 rm(list = ls())
2 N <- 1000
3 u <- rnorm(N)
4 x1 <- -2 + rnorm(N)
5 x2 <- 1 + x1 + rnorm(N)
6 y <- 1 + x1 + x2 + u
7 r1 <- ln(y - x1 + x2)
8
9
10 |
```

```
> ?ln
> rm(list = ls())
> N <- 1000
> u <- rnorm(N)
> x1 <- -2 + rnorm(N)
> x2 <- 1 + x1 + rnorm(N)
> y <- 1 + x1 + x2 + u
> r1 <- ln(y - x1 + x2)
> |
```

Vous ne serez pas
Développeur,
Vous resterez un
UTILISATEUR
de bibliothèques à
disposition



Hadley Wickham ✓
@hadleywickham



Installation de R studio

Se rendre sur <https://www.rstudio.com/products/rstudio/download/>

RStudio Desktop

Open Source License

Free

DOWNLOAD

[Learn more](#)

RStudio Desktop 2022.02.3+492 - [Release Notes](#)

1. Install R. RStudio requires R 3.3.0+.
2. Download RStudio Desktop. Recommended for your system:



DOWNLOAD RSTUDIO FOR WINDOWS

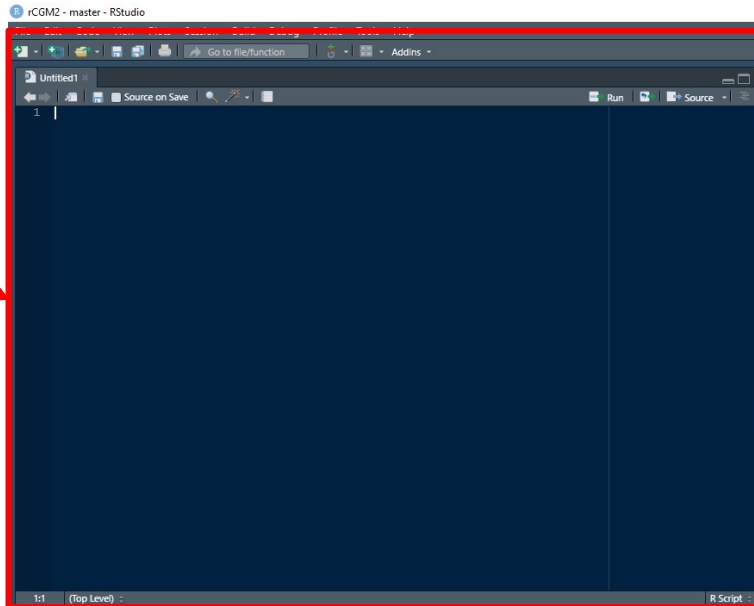
2022.02.3+492 | 177.26MB

Requires Windows 10/11 (64-bit)

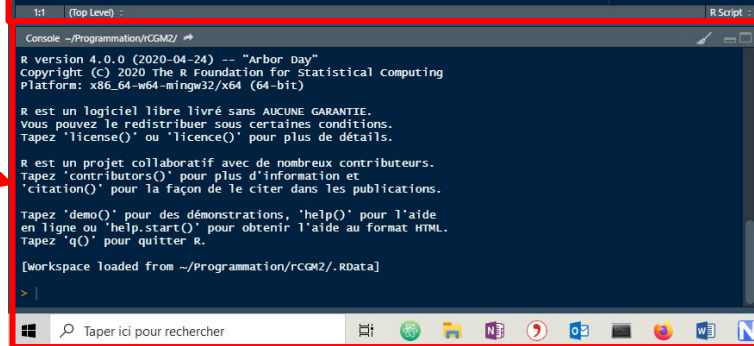


Présentation de Rstudio

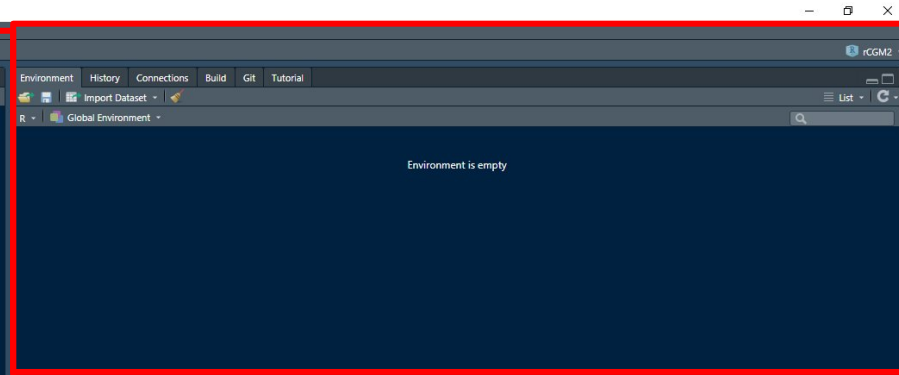
Editeur de scripts



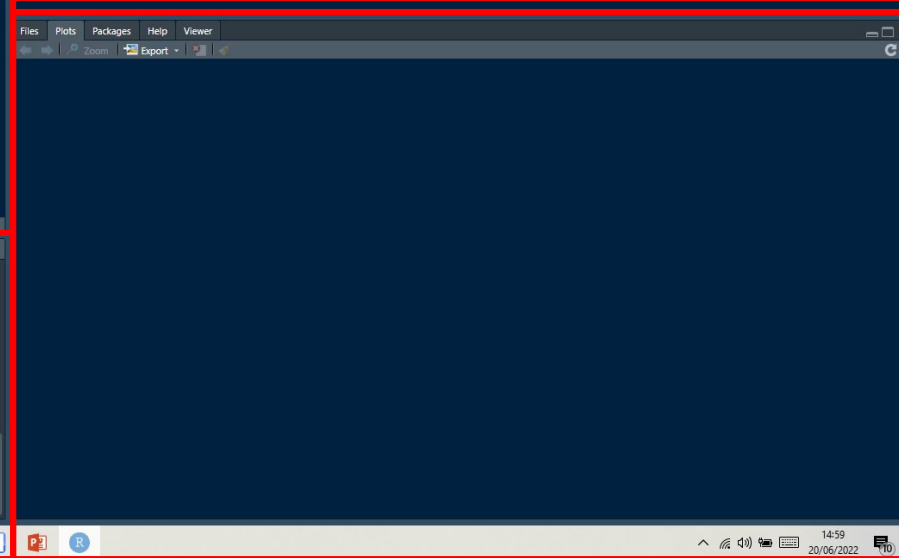
Console



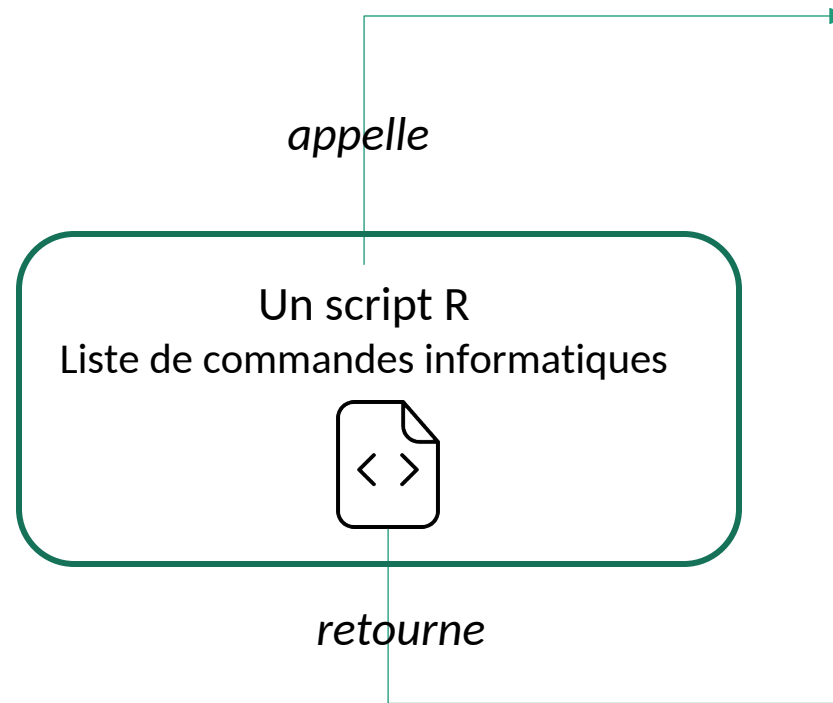
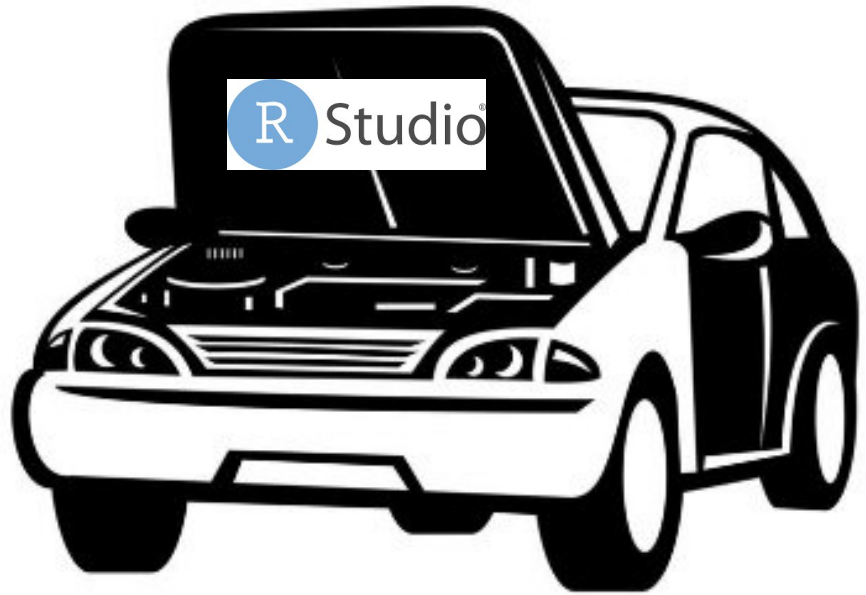
Environnement de travail



Fenêtre visualisation
(graphiques, fichiers...)

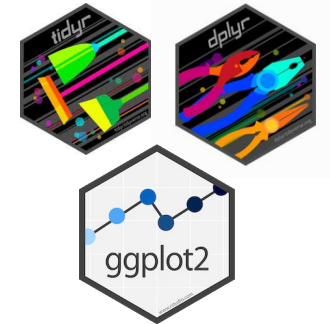


Exécution d'un script



Noyau de calcul R

Des bibliothèques de fonctions



<https://tidyr.tidyverse.org/>

Graphiques / résultats

Manipulons des données

UNE règle à respecter

Une LIGNE = un jeu d 'observations

Si le tableau est construit sous excel,
-> **oubliez la fusion de cellules**

Recommandations

Oubliez :

- Les accents
- Les espaces dans les noms de données

Observation	sujet	Session	TestDeMarche
01	AA	JO	250
02	AA	J30	350
03	AA	J60	400
04	BB	JO	250
05	BB	J30	500
06	BB	J60	800

Manipulons des données

Le jeu de données : Suivi de patients COVID

patientID	Session	FacteurRisque	Distance	Vitesse	LeverChaise	UpAndGo	ForcePinceDroit	ForcePinceGauche	FroceGraspDroit	ForceGraspGauche
AA011	M6	Surpoids	478	1,84	19	6,32	8	8,5	28	34
AH027	M6	Surpoids	503	1,48	33	5,85	6,5	4,25	12	8
AM009	M6	Surpoids	354	1,4	15	6,97	9,25	8	32	28
BK027	M6	Surpoids	542	2	31	5	8	8,5	30	36
BY025	M6	Surpoids	750	3,47	37	4,72	8,5	8,5	44	44
CN013	M6	Age	575	2,37	21	5	11	7,5	50	40
DA003	M6	Age	665	2,08	30	5,59	7,75	8,75	18	27
DB011	M6	Age	638	2,5	58	5	7,5	9,5	36	32
DB024	M6	Age	670	2,52	36	4,63	9,5	6	50	29
DN001	M6	Age	358	1,67	21	9,44	7	6	22	20
DP028	M6	Age	597	1,69	22	5,36	8,2	7,5	40	42
FJ003	M6	Hypertension	380	1,41	15	7,97	7,5	9	24	32
FJ005	M6	Hypertension	332	1,25	25	8	6,5	6	18	20

Session : M12 - M9-M18

Manipulons des données

Verbalisation



Bibliothèque « dplyr »
“dplyr is a grammar of data manipulation”

select()

filter()

arrange()

mutate()

summarise()

group_by()

Effectue des calculs par groupe

Manipulons des données

Verbalisation



select()

filter()

arrange()

mutate()

summarise()

```
1 #--- appel des bibliotheques---  
2 library(tidyverse)  
3  
4  
5 #---chargement des données---  
6  
7 table = read_excel("C:\\Users\\fleboeuf\\Desktop\\PRO\\COVID - stats (Hanae)\\CovidData_demo.xlsx")  
8  
9 # --- Manipulation de données avec la bibliotheque dplyr ---  
10  
11 # le verbe select()  
12 table_select = select(table,patientID,session,Distance)|
```

Manipulons des données

Verbalisation



select()

filter()

arrange()

mutate()

summarise()

```
# le verbe filter()
table_filter_onlyM6 = filter(table, session == "M6")
table_filter_allExceptM6 = filter(table, session != "M6")
```

Manipulons des données

Verbalisation



select()

filter()

arrange()

mutate()

summarise()

```
# le verbe arrange()
table_arrangeByDistance = arrange(table,Distance)
|
```

Manipulons des données

Verbalisation



select()

filter()

arrange()

mutate()

summarise()

Ajoute une colonne à la table de données
Ex: vitesse en kmH

```
# le verbe mutate()  
table_mutate = mutate(table, vitessekmh = vitesse*3.6)
```

	patientID	Session	Distance	Vitesse	LeverChaise	UpAndGo	ForcePinceDroit	ForcePinceGauche	ForceGraspDroit	ForceGraspGauche	VitesseKmh
1	AA011	M6	478	1.84	19	6.32	8.00	8.50	28.00	34.0	6.624
2	AH027	M6	503	1.48	33	5.85	6.50	4.25	12.00	8.0	5.328
3	AM000	M6	351	1.40	15	6.07	8.25	8.00	22.00	28.0	5.040

Manipulons des données

Verbalisation



select()

filter()

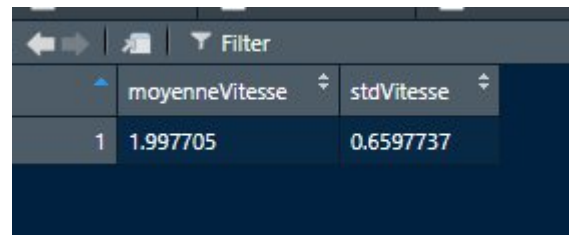
arrange()

mutate()

summarise()

Réduit la table aux nouvelles colonnes

```
# le verbe summarise()
table_summarise = summarise(table,
  moyenneVitesse = mean(vitesse, na.rm=TRUE),
  stdVitesse = sd(vitesse, na.rm=TRUE))
```



	moyenneVitesse	stdVitesse
1	1.997705	0.6597737

Manipulons des données

Verbalisation



select()

filter()

arrange()

mutate()

summarise()

group_by()

Effectue des calculs par groupe

```
# le verbe summarize associé a group_by|
tbl = group_by(table,patientID)
table_summarise_groupBy = summarise(tbl,
  moyennevitesse = mean(Vitesse,na.rm=TRUE),
  stdvitesse = sd(Vitesse,na.rm=TRUE))
```

```
# le verbe summarize associé a group_by ( avec operateur %>% (pipe))
table_summarise_groupBy2 = table %>%
  group_by(patientID)%>%
  summarise(moyennevitesse = mean(Vitesse,na.rm=TRUE),
  stdvitesse = sd(Vitesse,na.rm=TRUE))
```

Formulation à éviter !!

Manipulons des données

Verbalisation



select()

filter()

arrange()

mutate()

summarise()

group_by()

Effectue des calculs par groupe

```
# le verbe summarize associé a group_by|
tbl = group_by(table,patientID)
table_summarise_groupBy = summarise(tbl,
                                     moyenneVitesse = mean(Vitesse,na.rm=TRUE),
                                     stdVitesse = sd(Vitesse,na.rm=TRUE))
```

Formulation à éviter !!

```
# le verbe summarize associé a group_by ( avec operateur %>% (pipe))
table_summarise_groupBy2 = table %>%
  group_by(patientID)%>%
  summarise(moyenneVitesse = mean(Vitesse,na.rm=TRUE),
            stdVitesse = sd(Vitesse,na.rm=TRUE))
```

Utiliser l'operateur %>% pour
effectuer une succession de
manipulation

	patientID	moyenneVitesse	stdVitesse
1	AA011	1.960000	0.08000000
2	AH027	1.957500	0.42256163
3	AM009	1.703333	0.30005555
4	BK027	2.000000	0.00000000

Manipulons des données

Verbalisation



select()

filter()

arrange()

mutate()

summarise()

group_by()

Effectue des calculs par groupe

```
# succession d'operations
table_multipleoperations = table %>%
  filter(Session != "M6") %>%
  mutate(vitessekmh = vitesse*3.6) %>%
  group_by(patientID) %>%
  summarise(moyennevitesse = mean(vitessekmh, na.rm=TRUE),
            stdvitesse = sd(vitessekmh, na.rm=TRUE))
```


Manipulons des données

Transformation



Pivot_longer()

Pivot_wider()

patientID	Session	ForcePinceDroit	ForcePinceGauche
AA011	M6	8	8,5
AH027	M6	6,5	4,25
AM009	M6	9,25	8
BK027	M6	8	8,5
BY025	M6	8,5	8,5
CN013	M6	11	7,5
DA003	M6	7,75	8,75
DB011	M6	7,5	9,5
DB024	M6	9,5	6
DN001	M6	7	6
DP028	M6	8,2	7,5

longer

wider

patientID	Session	TypeForcePince	ForcePinceValue
AA011	M6	ForcePinceDroit	8
AA011	M6	ForcePinceGauche	8,5
AH027	M6	ForcePinceDroit	6,5
AH027	M6	ForcePinceGauche	4,5

```
# --- Manipulation de données avec la bibliothèque tidyR ---  
  
table_pivot_longer = table %>%  
  select(patientID,Session,ForcePinceDroit,ForcePinceGauche)%>%  
  pivot_longer(cols=ForcePinceDroit:ForcePinceGauche,  
              names_to = "TypeForcePince",  
              values_to = "ForceValue" )  
  
table_pivot_wider = table_pivot_longer %>%  
  pivot_wider(values_from = ForceValue,  
             names_from = TypeForcePince)
```

Représentons les données



<https://ggplot2.tidyverse.org/>

Grammaire de construction graphique

`ggplot(data, aes(x=?, y=?, color = class)) + geom_point()`



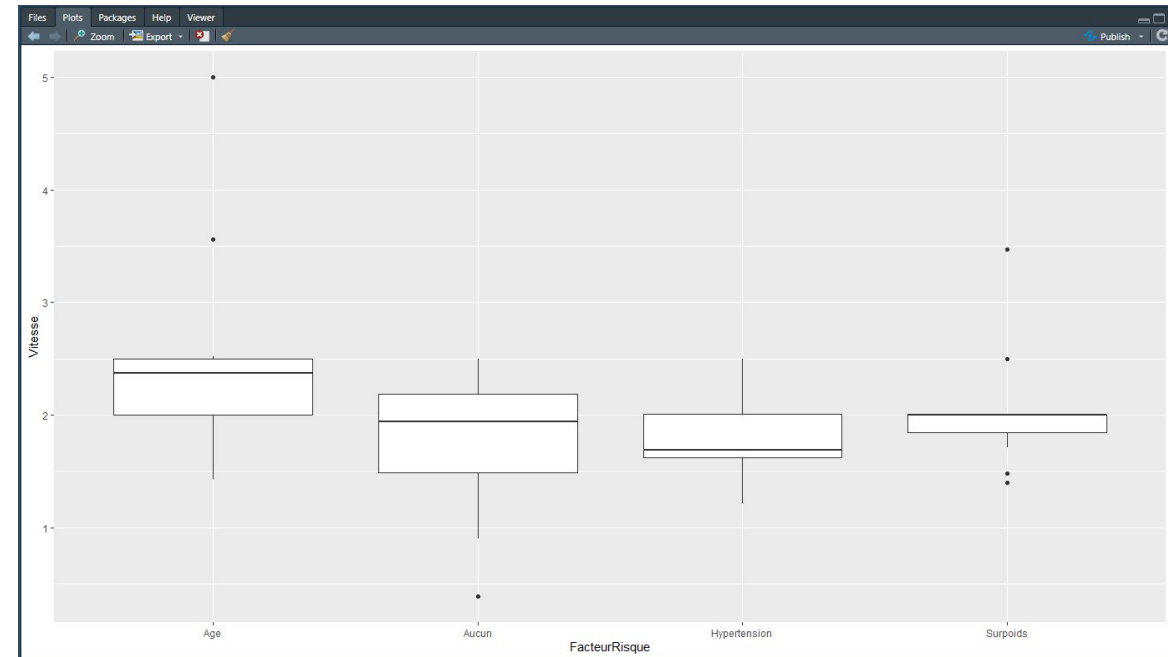
Données

Mapping



Geometrie

```
# exemple de graphique de base  
ggplot(table, aes(x=FacteurRisque, y=Vitesse)) + geom_boxplot()
```



Représentons les données



<https://ggplot2.tidyverse.org/>

De nombreuses géométries à disposition !!

<https://r-graph-gallery.com>

Distribution



Violin

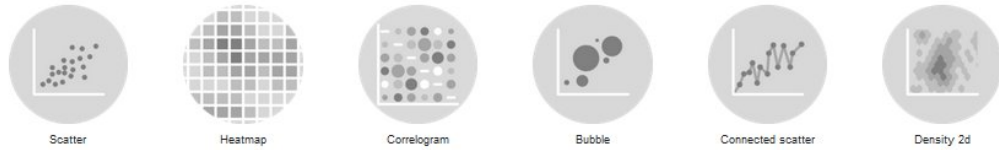
Density

Histogram

Boxplot

Ridgeline

Correlation



Scatter

Heatmap

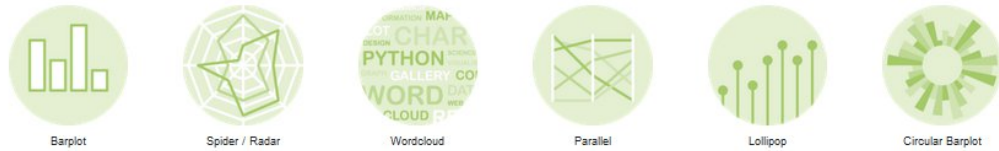
Correlogram

Bubble

Connected scatter

Density 2d

Ranking



Barplot

Spider / Radar

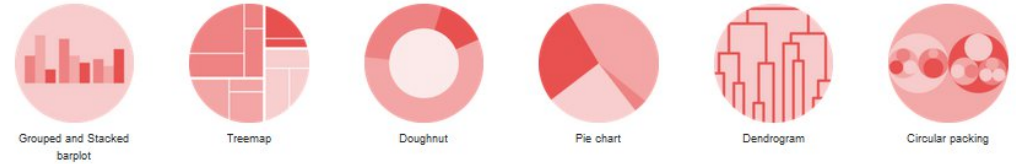
Wordcloud

Parallel

Lollipop

Circular Barplot

Part of a whole



Grouped and Stacked
barplot

Treemap

Doughnut

Pie chart

Dendrogram

Circular packing

Evolution



Line plot

Area

Stacked area

Streamchart

Time Series

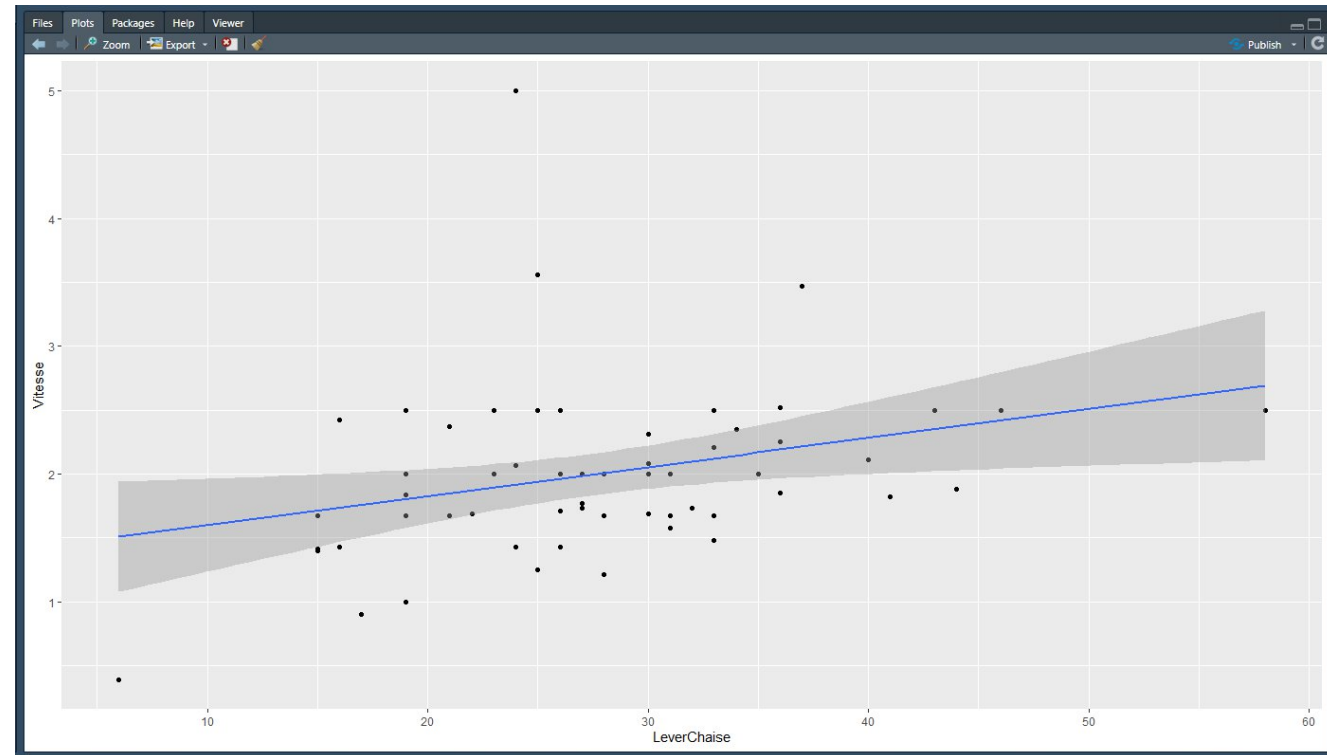
Représentons les données



<https://ggplot2.tidyverse.org/>

Explorer les relations linéaires.

```
ggplot(table, aes(x=LeverChaise, y=Vitesse)) +  
  geom_point()+  
  stat_smooth(method="lm")
```



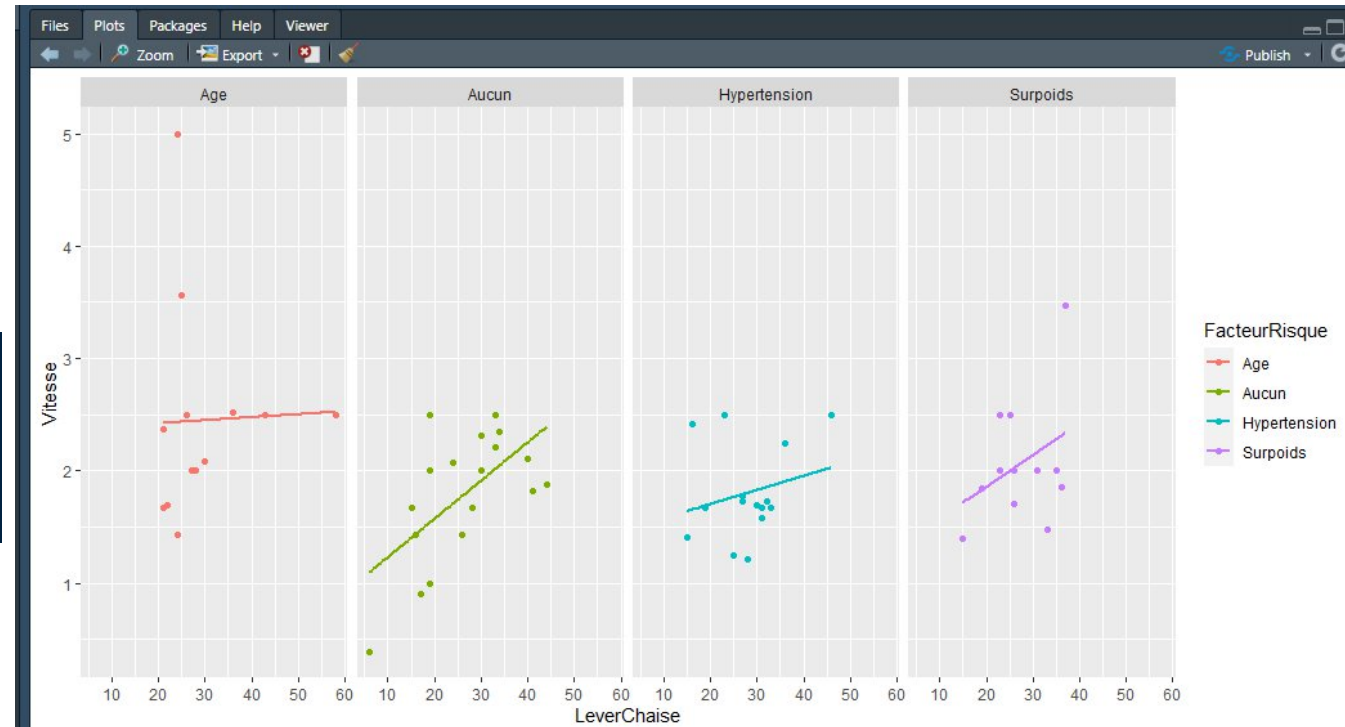
Représentons les données



<https://ggplot2.tidyverse.org/>

Explorer les relations linéaires.

```
ggplot(table, aes(x=LeverChaise, y=Vitesse, color=FacteurRisque)) +  
  geom_point()+  
  stat_smooth(method="lm, se=FALSE)+  
  facet_grid(.~FacteurRisque) # ajout d'une regression lineaire
```



Effectuons quelques tests statistiques

Régression lineaire – Significativité et R2

```
# regression lineaire - significativité et R2  
model = lm(Vitesse ~ LeverChaise,  
           filter(table,FacteurRisque == "Aucun"))  
summary(model)
```



console

```
> model = lm(Vitesse ~ LeverChaise,  
+           filter(table,FacteurRisque == "Aucun"))  
> summary(model)  
  
Call:  
lm(formula = vitesse ~ LeverChaise, data = filter(table, FacteurRisque ==  
"Aucun"))  
  
Residuals:  
    Min       1Q   Median       3Q      Max   
-0.70890 -0.44025  0.03737  0.34322  0.95854  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)      
(Intercept)  0.89464    0.31677   2.824  0.01221 *     
LeverChaise  0.03404    0.01125   3.027  0.00802 **    
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.4764 on 16 degrees of freedom  
(2 observations deleted due to missingness)  
Multiple R-squared:  0.3641,    Adjusted R-squared:  0.3243  
F-statistic: 9.16 on 1 and 16 DF,  p-value: 0.008022
```

Effectuons quelques tests statistiques

Anova (one-way)

```
# anova -  
model = lm(Vitesse ~ FacteurRisque, table)  
anova(model)  
summary(model)
```



console

```
> # anova -  
> model = lm(Vitesse ~ FacteurRisque, table)  
> anova(model)  
Analysis of Variance Table  
  
Response: Vitesse  
      Df Sum Sq Mean Sq F value Pr(>F)  
FacteurRisque  3  4.0083  1.33611  3.4446 0.02251 *  
Residuals    57 22.1097  0.38789  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
> summary(model)  
  
Call:  
lm(formula = Vitesse ~ FacteurRisque, data = table)  
  
Residuals:  
      Min       1Q   Median       3Q      Max  
-1.4011 -0.3611 -0.0500  0.2789  2.5523  
  
Coefficients:  
              Estimate Std. Error t value Pr(>|t|)  
(Intercept)      2.4477     0.1727  14.170 < 2e-16 ***  
FacteurRisqueAucun -0.6566     0.2267  -2.896  0.00534 **  
FacteurRisqueHypertension -0.6444     0.2360  -2.730  0.00840 **  
FacteurRisqueSurpoids -0.3977     0.2360  -1.685  0.09743 .  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.6228 on 57 degrees of freedom  
(9 observations deleted due to missingness)  
Multiple R-squared:  0.1535,    Adjusted R-squared:  0.1089  
F-statistic: 3.445 on 3 and 57 DF,  p-value: 0.02251
```

Effectuons quelques tests statistiques

Comparaisons multiples

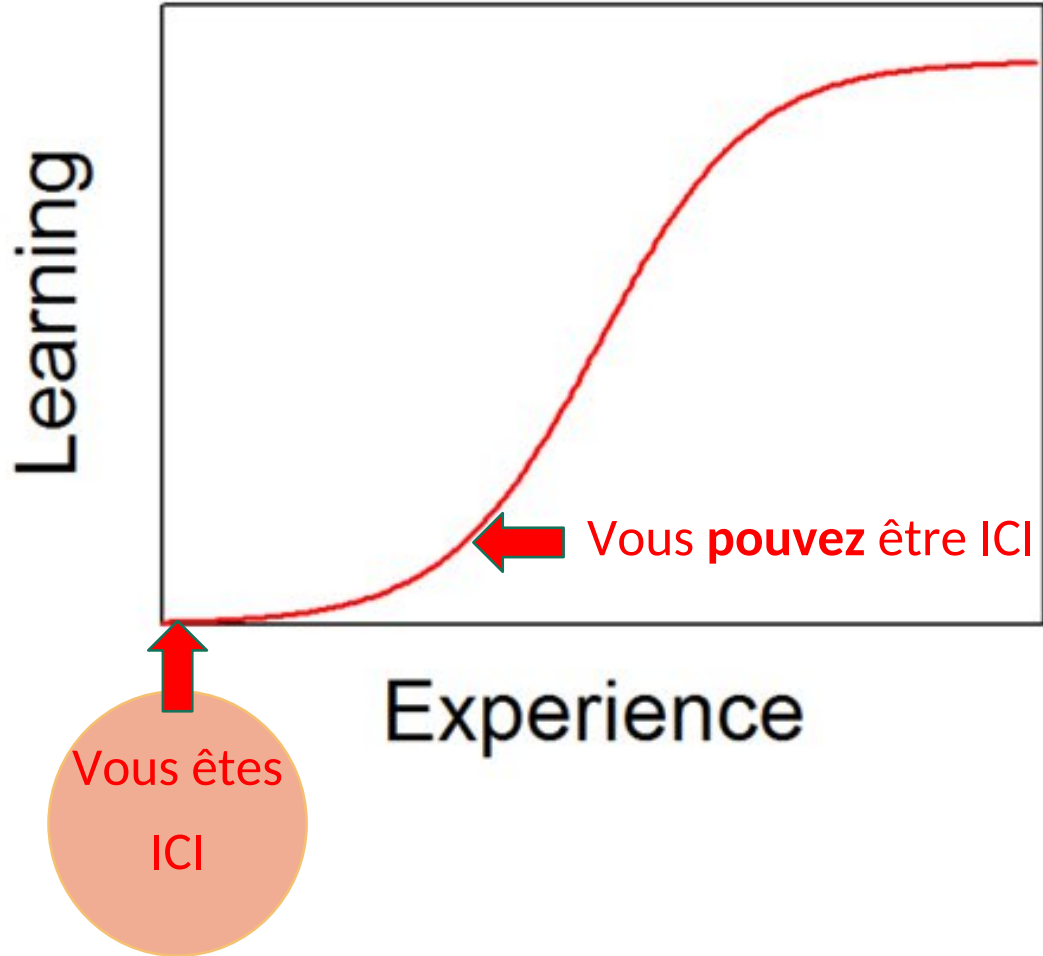
```
# comparaisons multiples  
pairwise.t.test(table$Vitesse,  
                table$FacteurRisque,  
                p.adjust.method = "bonferroni")
```



console

```
> pairwise.t.test(table$Vitesse,  
+                 table$FacteurRisque,  
+                 p.adjust.method = "bonferroni")  
  
Pairwise comparisons using t tests with pooled SD  
data: table$Vitesse and table$FacteurRisque  
  
          Age  Aucun Hypertension  
Aucun      0.032 -          -  
Hypertension 0.050 1.000 -  
Surpoids    0.585 1.000 1.000  
  
P value adjustment method: bonferroni
```


Comment aller plus loin



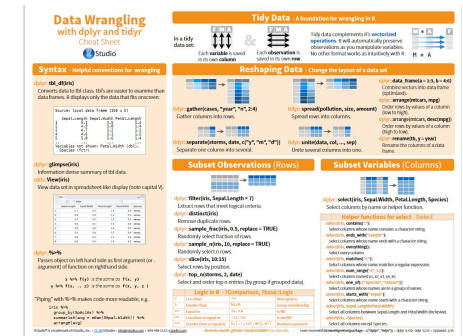
Pratique + Persévérance

Ressources

- Tutoriels web
- YouTube

NE PAS OUBLIER

Documentations
officielles



Comment aller plus loin

