

Préalables à une recherche par les données

It is often said that 80% of data analysis is spent on the process of cleaning and preparing the data.

Sébastien Déjean

Ingénieur de recherche
Institut de Mathématiques de Toulouse

`math.univ-toulouse.fr/~sdejean`



Plan

- Planifier si possible une future acquisition de données.
- Quand les données (planifiées ou pas) sont déjà là !
- Bien ranger ses données.
- Avec quels outils ?

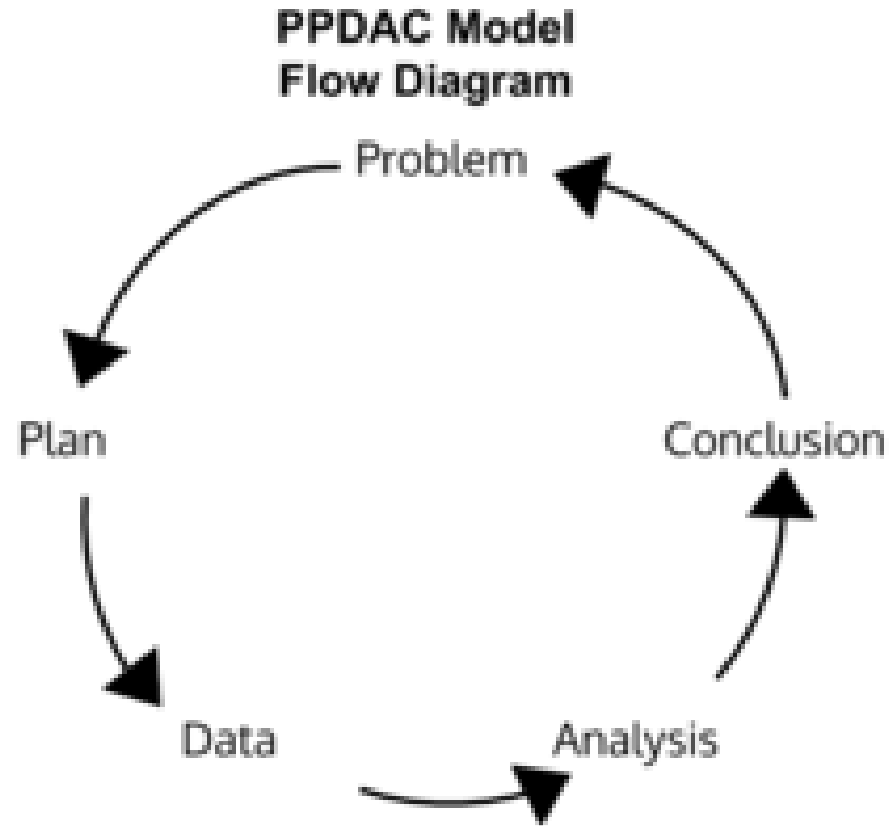
Planifier si possible une future acquisition de données

To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of. R.A. Fisher

While a good design does not guarantee a successful experiment, a suitably bad design guarantees a failed experiment—no results or incorrect results. K. Kerr (2003)

Une feuille de route possible

- 1) **P**roblème
- 2) **P**lanification
- 3) **D**onnées
- 4) **A**nalyse
- 5) **C**onclusion



Source : wiki.gis.com/wiki/index.php/PPDAC_Model

Éviter les confusion d'effet

2 conditions à l'étude : **Contrôle** / **Traitement**

Jour 1



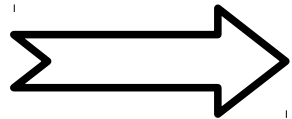
8 échantillons **Contrôle**

Jour 2



8 échantillons **Traitement**

Effet traitement ou effet jour ?



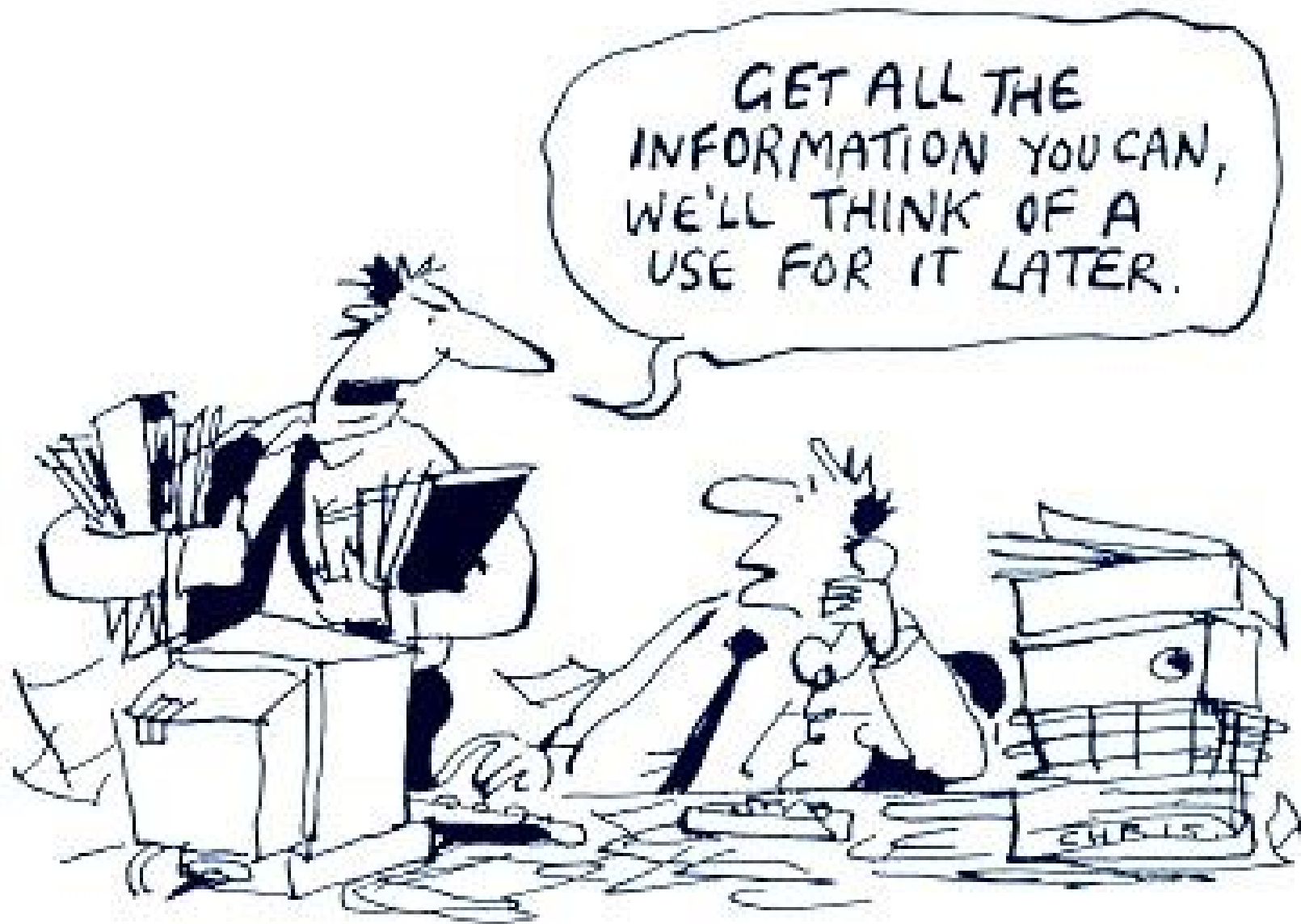
Jour 1



Jour 2



Randomisation



Source : ???

Quand les données (planifiées ou pas) sont déjà là !

- Données manquantes
- Données atypiques
- Distribution des données

Données manquantes

The best thing to do with missing values is not to have any.

Gertrude Mary Cox

- Situations
 - **Missing completely at random** (MCAR) : aucune relation entre la donnée non observée et une autre information contenue dans le jeu de données.
 - **Missing at random** (MAR) : le fait que la donnée ne soit pas observée ne dépend pas de la donnée elle-même mais d'une autre variable connue par ailleurs. Ex (fictif) : les femmes répondent moins souvent à une question concernant la consommation de cigarettes. L'absence de réponse à la question « Combien de cigarettes fumées par jour ? » est en partie liée au sexe de l'individu.
 - **Missing not at random** : le fait que la donnée ne soit pas observée dépend de la donnée elle-même. Ex : quelqu'un qui fume beaucoup ne répondra pas à une question concernant le nombre de cigarettes fumées par jour.
- Méthodes : imputation, interpolation, délétion partielle, faire sans !

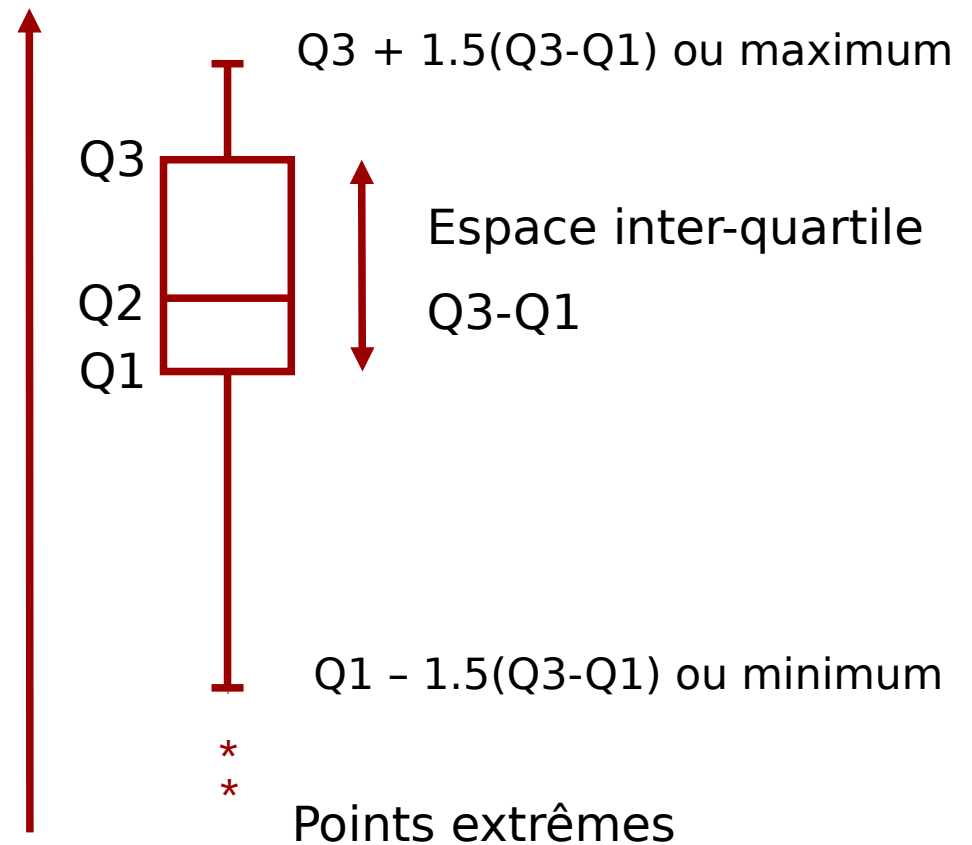
Données atypiques

- Pour une série de données :
boxplot

Q1 : 1^{er} quartile (25%)

Q2 : 2^{eme} quartile (50%)

Q3 : 3^{eme} quartile (75%)



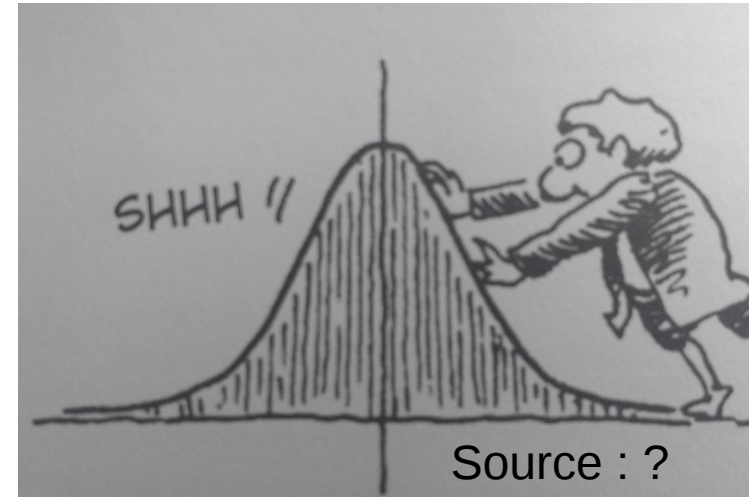
- Approche à plusieurs variables :
Exemple : package `mvoutlier` pour le logiciel R

Multivariate Outlier Detection Based on Robust Methods

Various Methods for Multivariate Outlier Detection

Distribution des données

- Centrage – réduction
 - Retrancher la moyenne, diviser par l'écart-type
- Conversion en logarithme
 - Ordre de grandeur
- Calcul de ratio, différence, taux...
 - Pour évaluer une évolution entre 2 temps



Conversion en log

Populations légales des communes au 1er janvier 2013
 Mise à jour : décembre 2012 en habitant
 Champ : Département de la Haute-Garonne, limites
 territoriales en vigueur au 1er janvier 2012
 Date de référence statistique : 1er janvier 2010
 Source : Insee, Recensement de la population 2010

Commune	Population	Log10
Toulouse	441 802	5.65
Colomiers	35 186	4.55
Tournefeuille	25 340	4.40
Muret	23 864	4.38
...		
Castanet-Tolosan	11 033	4.04
Saint-Orens...	10 918	4.04
Saint-Jean	10 259	4.01
Revel	9 361	3.97
Portet-sur-Garonne	9 435	3.97
Auterive	9 107	3.96
...		
La Magdelaine-sur-T/	1 006	3.00
Grépiac	990	2.99
Landorthe	946	2.98
Vigoulet-Auzil	944	2.97
...		
Belbèze-de-Lauragais	104	2.02
Saint-Germier	103	2.01
Seyre	102	2.01
Gouzens	95	1.98
Lourde	98	1.99
Pouze	97	1.99
...		
Saccourvielle	13	1.11
Cirès	13	1.11
Bourg-d'Oueil	8	0.90
Trébons-de-Luchon	8	0.90
Caubous	6	0.78
Baren	5	0.70

X	log2 (X)
16 = 2 ⁴	4
8 = 2 ³	3
4 = 2 ²	2
2 = 2 ¹	1
1 = 2 ⁰	0
0.5 = 2 ⁻¹	-1
0.25 = 2 ⁻²	-2
0.125 = 2 ⁻³	-3
...	

Bien ranger ses données

Des données bien rangées

H. Wickham, Tidy data, *Journal of Statistical Software*, 59(10), 2014.



Journal of Statistical Software

August 2014, Volume 59, Issue 10.

<http://www.jstatsoft.org/>

Tidy Data

Hadley Wickham
RStudio

Abstract

A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. This paper tackles a small, but important, component of data cleaning: data tidying. Tidy datasets are easy to manipulate, model and visualize, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. This framework makes it easy to tidy messy datasets because only a small set of tools are needed to deal with a wide range of un-tidy datasets. This structure also makes it easier to develop tidy tools for data analysis, tools that both input and output tidy datasets. The advantages of a consistent data structure and matching tools are demonstrated with a case study free from mundane data manipulation chores.

Keywords: data cleaning, data tidying, relational databases, R.

1. Introduction

It is often said that 80% of data analysis is spent on the process of cleaning and preparing

It is often said that 80% of data analysis is spent on the process of cleaning and preparing the data

... tidy datasets are all alike but every messy dataset is messy in its own way ().*

data tidying: structuring datasets to facilitate analysis.

This paper [...] provides a comprehensive "philosophy of data"

Since most real world datasets are not tidy...

Tidy datasets provide a standardized way to link the structure of a dataset (its physical layout) with its semantics (its meaning).

(*) Principe d'Anna Karenine



« *Les familles heureuses se ressemblent toutes. Les familles malheureuses sont malheureuses chacune à leur manière.* »

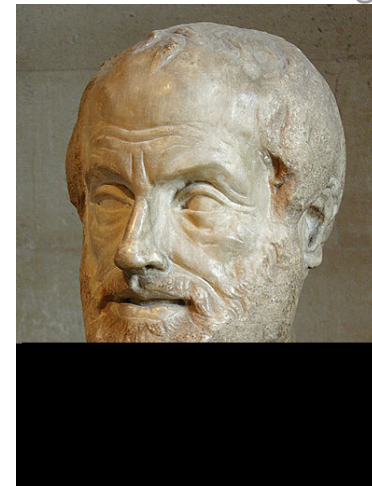
En d'autres termes, le succès demande que plusieurs conditions soient réunies. Une seule condition manquée est suffisante pour conduire à l'échec.

deselection.wordpress.com/2010/11/12/le-principe-danna-karenine/

Version Aristote en.wikipedia.org/wiki/Anna_Karenina_principle

Aristote, Éthique à Nicomaque :

Again, **it is possible to fail in many ways** [...] while to **succeed is possible only in one way** ([...] to miss the mark easy, to hit it difficult); [...] For men are good in but one way, but bad in many.



messy



tidy



LITTLE MISS TIDY

By Roger Hargreaves



Des données bien rangées

1. Chaque variable forme une colonne
2. Chaque observation forme une ligne
3. Chaque type « d'unité observationnelle » forme une table

Tout arrangement de données ne respectant pas ces 3 règles est considéré comme *messy*.

Messy data


Les jeux de données réels ne respectent quasiment jamais ces règles.

Exemples courants de violation de ces règles :

- Les noms de colonnes sont des valeurs pas des noms de variables.
- Plusieurs variables sont stockées dans une même colonne.
- Des variables sont stockées à la fois en ligne et en colonne.
- ...



Les noms de colonnes sont des valeurs, pas des noms de variable




religion	<\$10k	\$10–20k	\$20–30k	\$30–40k	\$40–50k	\$50–75k	...
Agnostic	27	34	60	81	76	137	
Atheist	12	27	37	52	35	70	
Buddhist	27	21	30	34	33	58	
Catholic	418	617	732	670	638	1116	
Don't know/refused	15	14	15	11	10	35	
Evangelical Prot	575	869	1064	982	881	1486	
Hindu	1	9	7	9	11	34	
Historically Black Prot	228	244	236	238	197	223	
Jehovah's Witness	20	27	24	24	21	30	
Jewish	19	19	25	25	30	95	

Extrait de H. Wickham, Tidy data, *Journal of Statistical Software*, 59(10), 2014.

3 variables :

- religion
- revenu
- effectif



religion	income	freq
Agnostic	<\$10k	27
Agnostic	\$10–20k	34
Agnostic	\$20–30k	60
Agnostic	\$30–40k	81
Agnostic	\$40–50k	76
Agnostic	\$50–75k	137
Agnostic	\$75–100k	122
Agnostic	\$100–150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

Chaque colonne représente une variable ; chaque ligne, une observation

Des variables sont stockées à la fois en ligne et en colonne



id	date	element	value
MX17004	2010-01-30	tmax	27.8
MX17004	2010-01-30	tmin	14.5
MX17004	2010-02-02	tmax	27.3
MX17004	2010-02-02	tmin	14.4
MX17004	2010-02-03	tmax	24.1
MX17004	2010-02-03	tmin	14.4
MX17004	2010-02-11	tmax	29.7
MX17004	2010-02-11	tmin	13.4
MX17004	2010-02-23	tmax	29.9
MX17004	2010-02-23	tmin	10.7

id	date	tmax	tmin
MX17004	2010-01-30	27.8	14.5
MX17004	2010-02-02	27.3	14.4
MX17004	2010-02-03	24.1	14.4
MX17004	2010-02-11	29.7	13.4
MX17004	2010-02-23	29.9	10.7
MX17004	2010-03-05	32.1	14.2
MX17004	2010-03-10	34.5	16.8
MX17004	2010-03-16	31.1	17.6
MX17004	2010-04-27	36.3	16.7
MX17004	2010-05-27	33.2	18.2



Cette colonne contient
un nom de variable

Une variable par
colonne, une
observation par ligne

Avec quels outils ?

Surprisingly, most messy datasets, including types of messiness not explicitly described above, can be tidied with a small set of tools: melting, string splitting, and casting.

Tidyverse

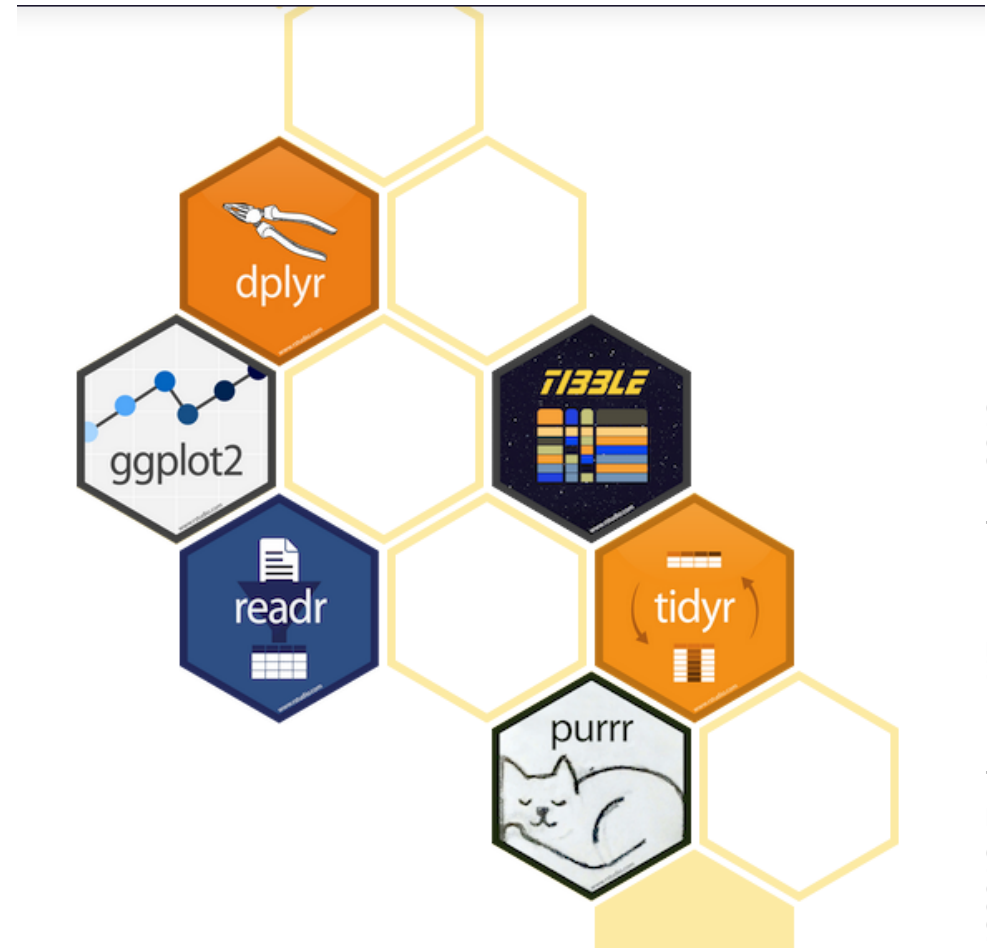


R packages for data science

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Objectif : faciliter

1/ la manipulation, 2/ la visualisation, 3/ la modélisation

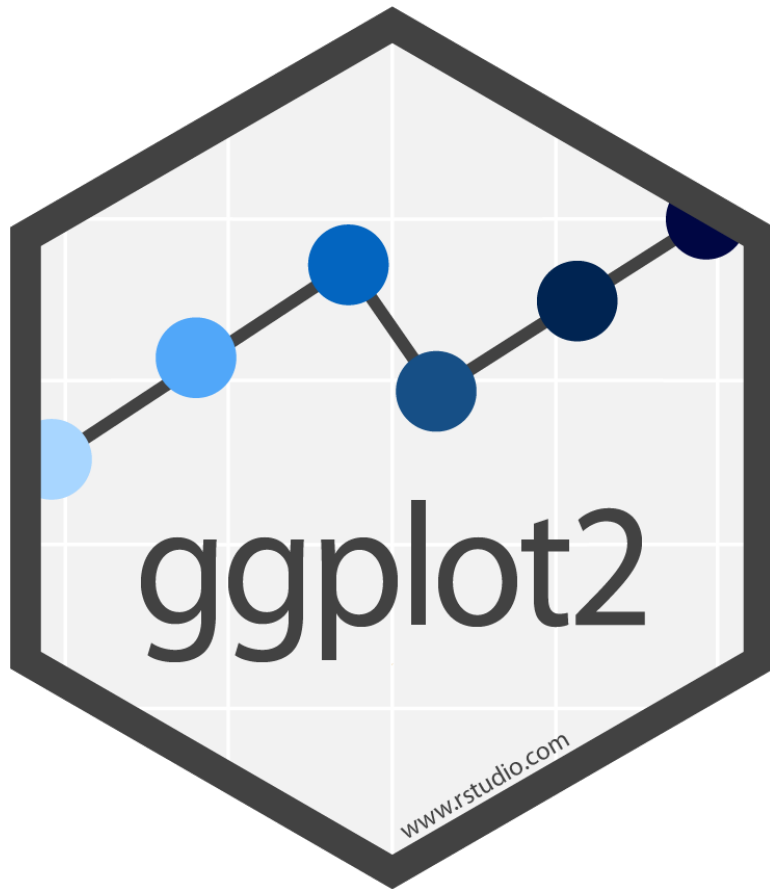


Manipulation

- **Filtrer** : extraire ou supprimer des sous-ensembles de données sur la base de conditions.
- **Transformer** : ajouter ou modifier des variables ; implication d'une ou de plusieurs variables.
- **Agréger** : résumer plusieurs valeurs en une seule.
- **Trier** : changer l'ordre des observations

Toutes ces opérations sont facilitées quand il existe un moyen cohérent d'accéder aux variables.
C'est le cas avec des données bien rangées.

Visualisation



- Package basé sur une grammaire des graphiques [*].
- On ne peut produire un graphique qu'à partir de données rangées dans une table.
- Principe :
 - Sujet : les données
 - Verbe : le type de graphique à produire
 - Complément : l'esthétique

[*] *The Grammar of Graphics*, L. Wilkinson, 2005
2nd Edition, Springer, Series Statistics and Computing

Modélisation

- Les données bien rangées facilitent la mise en œuvre de méthodes de modélisation.
- Tous les logiciels de statistique disposent d'un moyen de décrire un lien entre une variable réponse et des variables explicatives.
 - R (`lm()`): $y \sim a + b + c * d$.
 - SAS (`PROC GLM`): $y = a + b + c + d + c * d$.
 - SPSS (`glm`): `y BY a b c d / DESIGN a b c d c * d`.
 - Stata (`regress`): `y a b c#d`.

Des données dans un tableau

Groupe	Echantillon	Variable 1	Variable 2	Variable 3	Concentration produit X					Expression de gènes		
					T0	T1	T2	T3	T4	Gène 1	Gène 2	Gène 3
Groupe A	1	75,6	4,07	0,965	3,45	3,59	4,29	5,14	5,69	380	8	60
	3	70,8	7,22	0,143	4,63	4,64	5,21	5,73	6,24	208	452	97
	4	65,4	8,25	0,893	1,32	2,03	2,04	2,41	3,11	110	93	107
	5	61,4	8,65	0,505	2,99	3,64	4,08	4,36	4,91	124	458	205
	6	52,8	0,43	0,815	4,12	4,33	5,22	6,02	6,59	445	33	312
	7	82,8	6,20	0,244	1,81	2,06	2,47	2,68	3,17	215	461	154
	8	10,1	3,67	0,358	1,40	1,60	1,95	1,96	2,15	165	215	386
	Groupe B	1	84,8	0,08	0,867	1,26	1,30	1,58	1,82	1,91	432	389
2		46,8	7,31	0,020	3,67	3,97	4,13	5,00	5,77	174	135	188
3		64,4	2,16	0,169	4,78	5,43	5,79	6,21	6,34	378	202	418
4		65,1	5,76	0,334	2,97	3,11	3,89	4,04	4,58	451	272	349
5		29,9	2,64	0,873	3,82	4,38	4,66	5,44	5,57	103	34	131
6		79,4	7,93	0,985	4,88	5,53	6,17	6,60	7,25	51	1	359
7		51,8	4,82	0,507	0,46	1,03	2,03	2,45	2,61	11	359	312
8		86,4	2,48	0,231	0,23	0,52	0,83	1,04	1,48	96	147	282
Groupe C	1	72,6	1,76	0,703	2,76	3,60	4,31	4,36	4,99	341	78	120
	2	65,6	7,86	0,372	1,79	2,58	2,71	3,08	3,24	485	6	181
	3	90,1	0,08	0,735	3,45	3,86	4,27	4,42	5,40	461	32	394
	4	92,4	8,37	0,535	2,77	3,09	3,81	4,24	4,89	305	166	107
	5	48,5	7,75	0,109	1,27	1,78	2,00	2,38	3,27	365	410	390
	6	18,0	9,88	0,663	0,43	1,39	1,57	1,72	1,75	123	441	461
	7	57,2	5,04	0,671	0,40	1,26	1,98	2,16	2,72	390	352	15
	8	38,9	4,67	0,377	4,64	5,63	6,40	6,75	6,86	289	359	104

Des données prêtes à être importées dans un logiciel de statistique

```

Groupe;Echantillon;Var_1;Var_2;Var_3;Conc_T0;Conc_T1;Conc_T2;Conc_T3;Conc_T4;Gene_1;Gene_2;Gene_3;VC1;VC2
A;1;;75,6;4,07;0,965;3,45;3,59;4,29;5,14;5,69;380;8;60;0,35;0,73
A;3;;70,8;7,22;0,143;4,63;4,64;5,21;5,73;6,24;208;452;97;0,51;0,49
A;4;;65,4;8,25;0,893;1,32;2,03;2,04;2,41;3,11;110;93;107;0,50;0,73
A;5;;61,4;8,65;0,505;2,99;3,64;4,08;4,36;4,91;124;458;205;0,09;0,12
A;6;;52,8;0,43;0,815;4,12;4,33;5,22;6,02;6,59;445;33;312;0,79;0,81
A;7;;82,8;6,20;0,244;1,81;2,06;2,47;2,68;3,17;215;461;154;0,31;0,01
A;8;;10,1;3,67;0,358;1,40;1,60;1,95;1,96;2,15;165;215;386;0,13;0,10
B;1;;84,8;0,08;0,867;1,26;1,30;1,58;1,82;1,91;432;389;107;0,59;0,75
B;2;;46,8;7,31;0,020;3,67;3,97;4,13;5,00;5,77;174;135;188;0,87;0,41
B;3;;64,4;2,16;0,169;4,78;5,43;5,79;6,21;6,34;378;202;418;0,25;0,71
B;4;;65,1;5,76;0,334;2,97;3,11;3,89;4,04;4,58;451;272;349;0,79;0,57
B;5;;29,9;2,64;0,873;3,82;4,38;4,66;5,44;5,57;103;34;131;0,11;0,92
B;6;;79,4;7,93;0,985;4,88;5,53;6,17;6,60;7,25;51;1;359;0,42;0,71
B;7;;51,8;4,82;0,507;0,46;1,03;2,03;2,45;2,61;11;359;312;0,72;0,73
B;8;;86,4;2,48;0,231;0,23;0,52;0,83;1,04;1,48;96;147;282;0,14;0,47
C;1;;72,6;1,76;0,703;2,76;3,60;4,31;4,36;4,99;341;78;120;0,59;0,56
C;2;;65,6;7,86;0,372;1,79;2,58;2,71;3,08;3,24;485;6;181;0,55;0,36
C;3;;90,1;0,08;0,735;3,45;3,86;4,27;4,42;5,40;461;32;394;0,71;0,43
C;4;;92,4;8,37;0,535;2,77;3,09;3,81;4,24;4,89;305;166;107;0,84;0,82
C;5;;48,5;7,75;0,109;1,27;1,78;2,00;2,38;3,27;365;410;390;0,12;0,63
C;6;;18,0;9,88;0,663;0,43;1,39;1,57;1,72;1,75;123;441;461;0,70;0,83
C;7;;57,2;5,04;0,671;0,40;1,26;1,98;2,16;2,72;390;352;15;0,67;0,97
C;8;;38,9;4,67;0,377;4,64;5,63;6,40;6,75;6,86;289;359;104;0,15;0,37

```

Conclusion

- D'autres aspects de la préparation des données sont à prendre en compte :
 - Gestion des dates
 - Identification des données manquantes : NA, MV, 9999, " ", ...
 - Encodage des caractères (accents...)
 - Identification de données similaires mais enregistrées différemment...
- Donoho, 2017 [*] : *Greater Data Science*
 - 1/ Data Gathering, Preparation and Exploration
 - 2/ Data Representation and Transformation
 - 3/ Computing with Data
 - 4/ Data Modeling
 - 5/ Data Visualization and Presentation
 - 6/ Science about Data Science

[*] D. Donoho, 50 Years of Data Science
Journal of Computational and Graphical Statistics, 26:4, 745-766, 2017.