



# Introduction à R

**TSNAD Master TraDD - École des Ponts ParisTech**

Séance du 04 octobre 2024

# Sommaire

- 1) Qu'est-ce que **R**?
- 2) Principes et propriétés
- 3) Quels types de traitements?
- 4) Manipulations basiques
- 5) Importation de fichiers de données
- 6) Retraitement

# 1) Qu'est-ce que R?

R est un langage de **programmation objet** utilisé pour le **traitement** et l'**analyse** de jeux de **données**.

**Logiciel libre** disponible sur la plupart des supports (Windows, MAC OS, Linux...).

Créé par les **universitaires** américains Ross Ihaka & Robert Gentleman au **début des années 1990** à partir de programmes en **C, Fortran** et **Java**.

Logiciel de traitement et d'analyse de données **le plus utilisé au monde** par les entreprises, les universitaires, les organismes publics ou encore les ONG.

**2 types de méthodes** d'analyses statistiques:

- numériques
- graphiques

## 2) Principes et propriétés (1)

**R** = langage **objet**  $\Leftrightarrow$  on définit des objets (ex: valeurs numériques, vecteurs, fonctions, caractères texte, tableaux de données...) qui sont ensuite **stockés** dans la **mémoire** vive de la machine.

Pour créer un objet, il suffit de lui **attribuer un nom** et de **définir ce qu'il représente**.

Ex: **x <- 2.538** stocke le nombre réel 2,538 dans un espace mémoire que l'on nomme « x ». On peut donc créer à partir de x une infinité d'objets comme par exemple:

- **y <- 2\*x+3**
- **z <- y^2**
- **v <- exp(z)**

...etc.

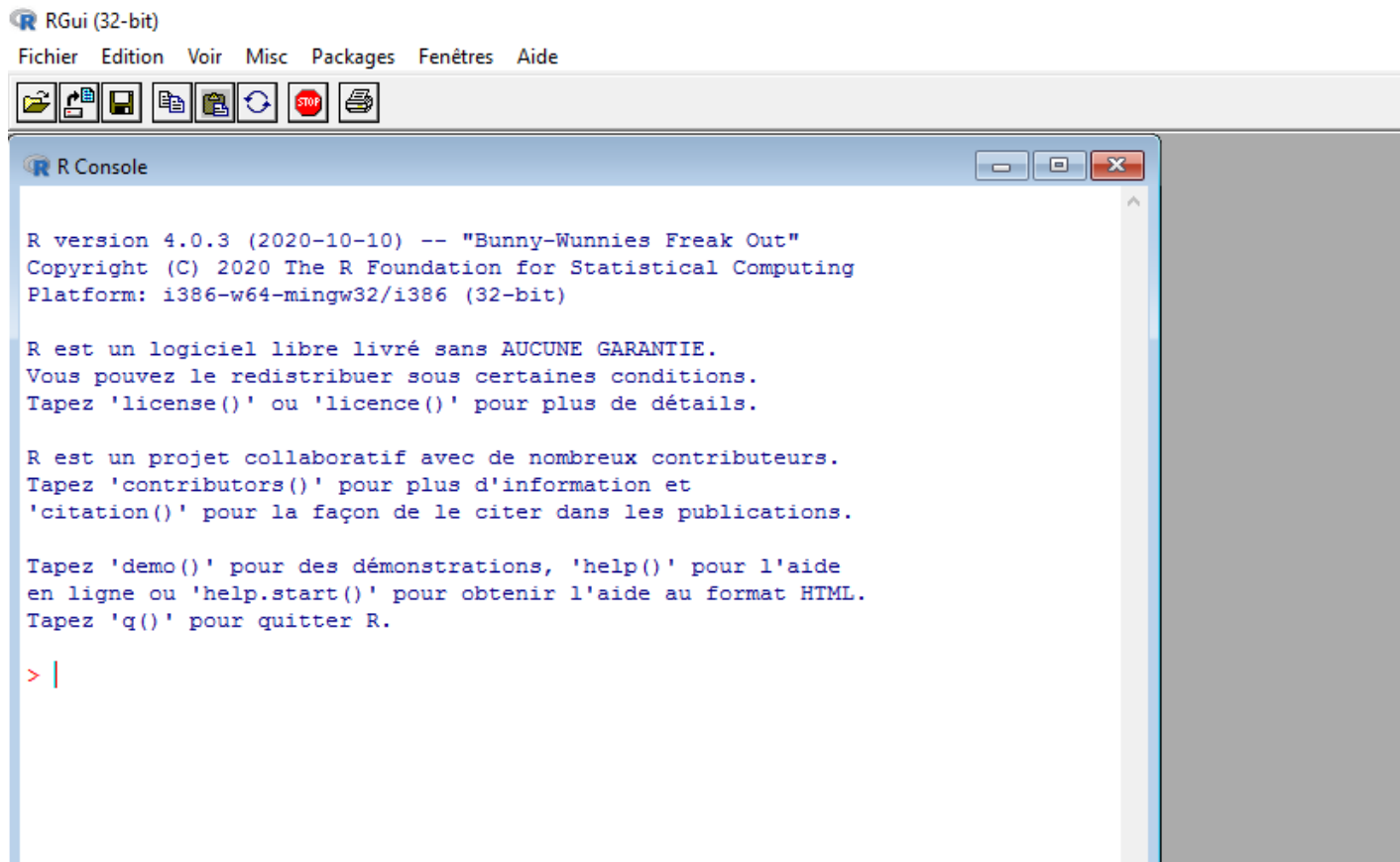
Pour le stockage, **R** alloue à chaque objet **un seul emplacement**  $\Leftrightarrow$  **unicité** de l'objet.

Par exemple, il est impossible d'avoir à la fois **x <- 2.538** et **x <- 0.654**.

Par ailleurs, **x  $\neq$  X** (distinction entre les minuscules et les majuscules).

# 2) Principes et propriétés (2)

La console de commande:



The screenshot shows the RGui (32-bit) application window. The title bar reads "RGui (32-bit)". The menu bar includes "Fichier", "Edition", "Voir", "Misc", "Packages", "Fenêtres", and "Aide". The toolbar contains icons for file operations (open, save, print, etc.) and a red stop button. The "R Console" window is open, displaying the following text:

```
R version 4.0.3 (2020-10-10) -- "Bunny-Wunnies Freak Out"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-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.

> |
```

## 2) Principes et propriétés (3)

Certaines commandes ou fonctionnalités du logiciel ne sont pas disponibles en les appelant directement dans la console.

Solution: télécharger et installer des modules complémentaires, également appelés «packages», contenant ces fonctionnalités => connexion aux serveurs universitaires sur lesquels se trouvent ces modules.

Secure CRAN mirrors



China (Shanghai 2) [https]  
Costa Rica [https]  
Denmark [https]  
East Asia [https]  
Ecuador (Cuenca) [https]  
Ecuador (Quito) [https]  
Estonia [https]  
France (Lyon 1) [https]  
France (Lyon 2) [https]  
France (Marseille) [https]  
France (Montpellier) [https]  
Germany (Erlangen) [https]  
Germany (Leipzig) [https]  
Germany (Göttingen) [https]  
Germany (Münster) [https]  
Germany (Regensburg) [https]  
Greece [https]  
Hungary [https]  
Iceland [https]  
India [https]  
Indonesia (Jakarta) [https]  
Iran [https]  
Italy (Milano) [https]  
Italy (Padua) [https]  
Japan (Tokyo) [https]  
Korea (Gyeongsan-si) [https]  
Korea (Seoul 1) [https]  
Korea (Ulsan) [https]  
Malaysia [https]  
Mexico (Mexico City) [https]  
Morocco [https]  
Netherlands [https]

Lorsqu'un package est téléchargé et installé, on va le chercher dans la bibliothèque à l'aide de la fonction **library**, comme suit: `library(Nom du package)`

# 3) Quels types de traitements? (1)

Traitements et analyses possibles sous R (liste non exhaustive):

- **Import/création** et **retraitement** de tables de données (ajout/suppression d'observations et de variables, concaténation...)
- Arithmétique (additions, multiplications, fractions...)
- Calcul vectoriel (sommes, produits scalaires...)
- Calcul matriciel (sommes, multiplications, transpositions, inversions, diagonalisations...)
- Représentations graphiques (nuages de points, courbes de fonctions...)

# 3) Quels types de traitements? (2)

- **Résumés statistiques** numériques (moyenne, quartiles, écart-type...) et graphiques (histogrammes, boîtes à moustaches...)
- Echantillonnage aléatoire, estimation de densité, test paramétrique/non paramétrique
- **Analyses multivariées** (Analyses en Composantes Principales, Analyses des Correspondances...)
- **Modèles de régression** (MCO, LOGIT, PROBIT, MLG...) et de séries temporelles
- Classification supervisée/non supervisée (classifieurs bayésiens, arbres de décision, clustering hiérarchique, k-means...)



# 4) Manipulations basiques (1)

Quelques exemples:

## 1) valeur numérique

```
a<- -3.27
```

```
a = -3.27 #autre écriture possible pour définir un objet#
```

## 2) fonction usuelles

```
b<- abs(a)
```

```
c<- 0.079-8.77*a+1.38*a^2-2*a^3
```

```
d<- sqrt(b)
```

```
e<- log(b) #log népérien#
```

```
f<-log10(b) #log en base 10#
```

```
g<- exp(c)
```

```
h<- cos(d)
```

```
i<- tan(-e)
```

# 4) Manipulations basiques (2)

## 3) dérivée d'une fonction en un point

```
t<- 3
```

```
numericDeriv(quote(t^2),"t")
```

```
#dérivée de la fonction  $f(t)=t^2$  au point  $t=3$ 
```

```
numericDeriv(quote(log(t)),"t")
```

```
numericDeriv(quote(sin(t)),"t")
```

## 4) suites et séries réelles

```
u<- 0
```

```
for (k in 1:100) {u[k]<- 1/k^2}
```

```
#on définit la suite  $u_k = \frac{1}{k^2}$  pour les 100 premiers entiers naturels non nuls
```

```
u[7]
```

```
#valeur de  $u_7$ 
```

```
sum(u)
```

```
#série  $\sum_{k=1}^{100} u_k$ 
```

# 4) Manipulations basiques (3)

## 5) calcul vectoriel

```
V1<- c(0,-1.32,4,-8.67)
```

```
V2<- 2*V1+3
```

```
V3<- V1+V2^2
```

*#somme de **V1** et du vecteur contenant chaque de terme de **V2** élevé au carré#*

```
V4<- V1*V2
```

*#multiplication terme par terme des vecteurs **V1** et **V2**#*

```
V5<- V2%*%V4
```

*#produit scalaire des vecteurs **V2** et **V4**#*

# 4) Manipulations basiques (4)

## 6) calcul matriciel

```
M1<- matrix(c(2,-1,0,4,-7,6),2,3)
```

*#matrice réelle 2x3 formée à partir d'un vecteur#*

```
M2 <- t(M1)
```

*#transposée de **M1**#*

```
M3<- 5.94*M1+1.31
```

*#combinaison linéaire de **M1**#*

```
M4<- exp(M1)
```

*#matrice 2x3 composée de l'exponentielle de chacune de coordonnées de **M1**#*

## 4) Manipulations basiques (5)

```
M5<- matrix(c(0,7,2,4,9,-2),2,3)
```

```
M6<- 0.3*M1-2.6*M5
```

*#combinaison linéaire des matrices 2x3 **M1** et **M5**#*

```
M7<- matrix(c(1,-3,6,-3,-5,8,6,8,11),3,3)
```

*#matrice carrée de dimension 3 formée à partir d'un vecteur de 9 valeurs numériques#*

```
q<- det(M7)
```

*#déterminant de **M7**#*

```
M8<- solve(M7)
```

*#inverse de **M7**#*

```
M9<- M7%*%M8
```

*#multiplication de **M7** par son inverse **M8** => on retrouve bien la matrice identité#*

```
eig<- eigen(M7)
```

*#valeurs propres et vecteurs propres de **M7**#*

# 4) Manipulations basiques (6)

## 7) statistiques descriptives

Id	Age	Sexe	Années_études	Salaire_net_mens	Dist_domicile_travail
I1	28	Femme	5	2340	entre 5 et 10 km
I2	46	Femme	6	3180	entre 5 et 10 km
I3	34	Homme	4	2760	entre 5 et 10 km
I4	24	Femme	2	1840	moins de 5 km
I5	41	Femme	6	2970	10 km et plus
I6	50	Homme	7	3670	entre 5 et 10 km
I7	45	Femme	3	2130	10 km et plus
I8	46	Homme	7	3430	entre 5 et 10 km
I9	34	Homme	4	2750	entre 5 et 10 km
I10	37	Femme	3	2090	entre 5 et 10 km
I11	40	Femme	5	2910	entre 5 et 10 km
I12	46	Homme	8	3850	entre 5 et 10 km
I13	25	Femme	2	1780	entre 5 et 10 km
I14	40	Femme	4	2650	10 km et plus
I15	32	Homme	4	2380	entre 5 et 10 km
I16	25	Homme	2	1760	10 km et plus
I17	32	Femme	2	1340	10 km et plus
I18	55	Homme	9	4210	entre 5 et 10 km
I19	51	Homme	6	3970	entre 5 et 10 km
I20	28	Homme	1	1780	10 km et plus
I21	31	Homme	3	2270	10 km et plus
I22	44	Femme	5	3350	entre 5 et 10 km
I23	38	Homme	6	2190	entre 5 et 10 km
I24	36	Femme	5	2900	entre 5 et 10 km
I25	39	Homme	4	2400	entre 5 et 10 km
I26	29	Homme	1	1480	entre 5 et 10 km
I27	48	Femme	8	3420	entre 5 et 10 km
I28	30	Femme	0	1670	entre 5 et 10 km
I29	27	Femme	0	1500	10 km et plus

## 4) Manipulations basiques (7)

### *Résumés numériques de la distribution d'une variable quantitative*

```
> summary(data$Salaire_net_mens)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1340	1840	2400	2585	3180	4210

```
> sd(data$Salaire_net_mens)
```

```
[1] 807.6139
```

*#écart-type#*

```
> sd(data$Salaire_net_mens)/mean(data$Salaire_net_mens)
```

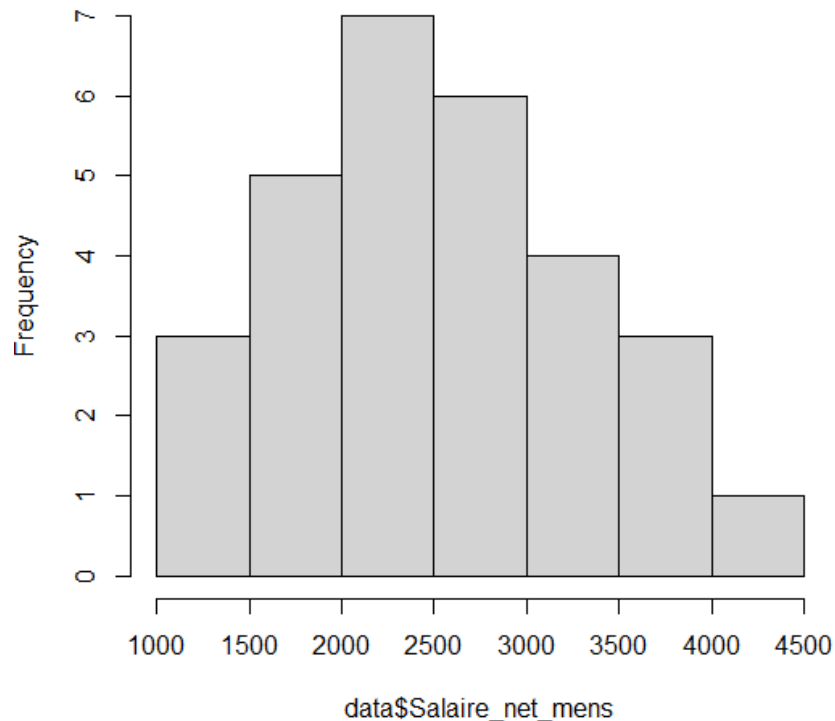
```
[1] 0.3124023
```

*#coefficient de variation#*

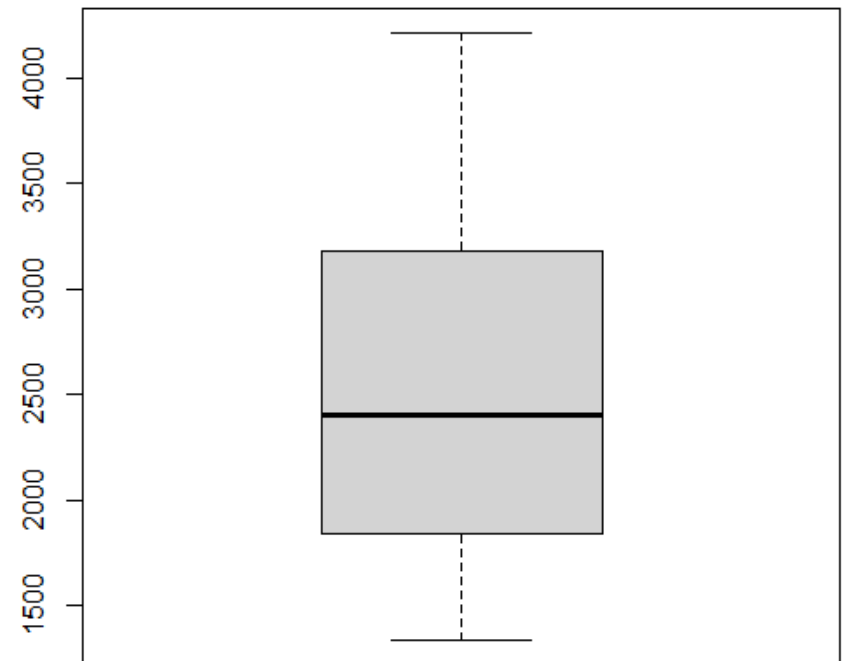
# 4) Manipulations basiques (8)

## *Résumés graphiques de la distribution d'une variable quantitative*

```
> hist(data$Salaire_net_mens)
```



```
> boxplot(data$Salaire_net_mens)
```



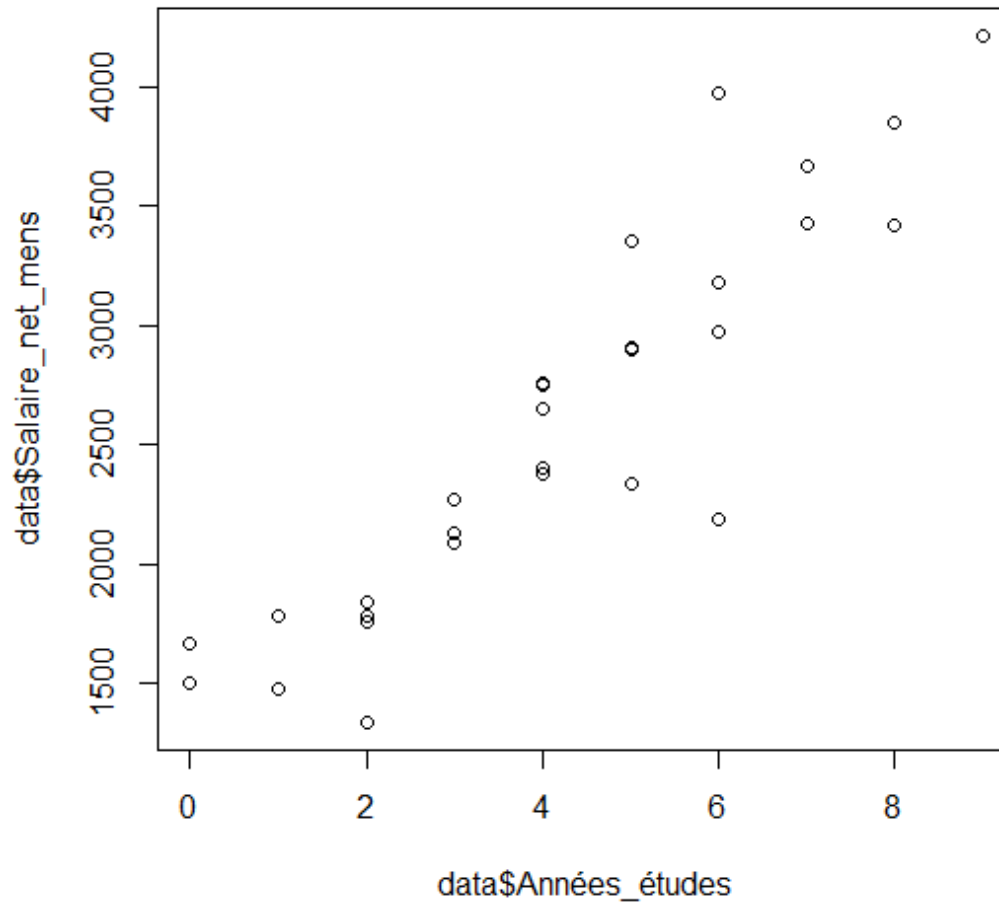


# 4) Manipulations basiques (9)

*Pour étudier la relation entre deux variables quantitatives*

- Le **coefficient de corrélation** mesure le degré de **dépendance linéaire** entre deux variables **quantitatives**.
- Il est compris entre **-1 et 1**.
- Plus il est **proche de 1 en valeur absolue**, plus la relation entre les deux variables est **significative**. A l'inverse, plus il est **proche de 0**, plus il est **difficile d'établir un lien** entre ces variables.

## 4) Manipulations basiques (10)



```
> cor(data$Salaire_net_mens,data$Années_études)  
[1] 0.9106582
```

## 4) Manipulations basiques (11)

*Pour des variables qualitatives*

- **Tri à plat:** on recense les **effectifs** de chaque **modalité** de la variable.

```
> table(data$Sexe)
```

Femme	Homme
15	14

## 4) Manipulations basiques (12)

- **Tri croisé:** on recense les effectifs relatifs au **croisement des modalités** de **deux** variables dans ce que l'on nomme un **tableau de contingence**.

```
> table(data$Dist_domicile_travail,data$Sexe)
```

	Femme	Homme
10 km et plus	5	3
entre 5 et 10 km	9	11
moins de 5 km	1	0

# 5) Importation de fichiers de données (1)

**R** permet d'importer des fichiers contenant des données que l'on souhaite (re)traiter et/ou analyser.

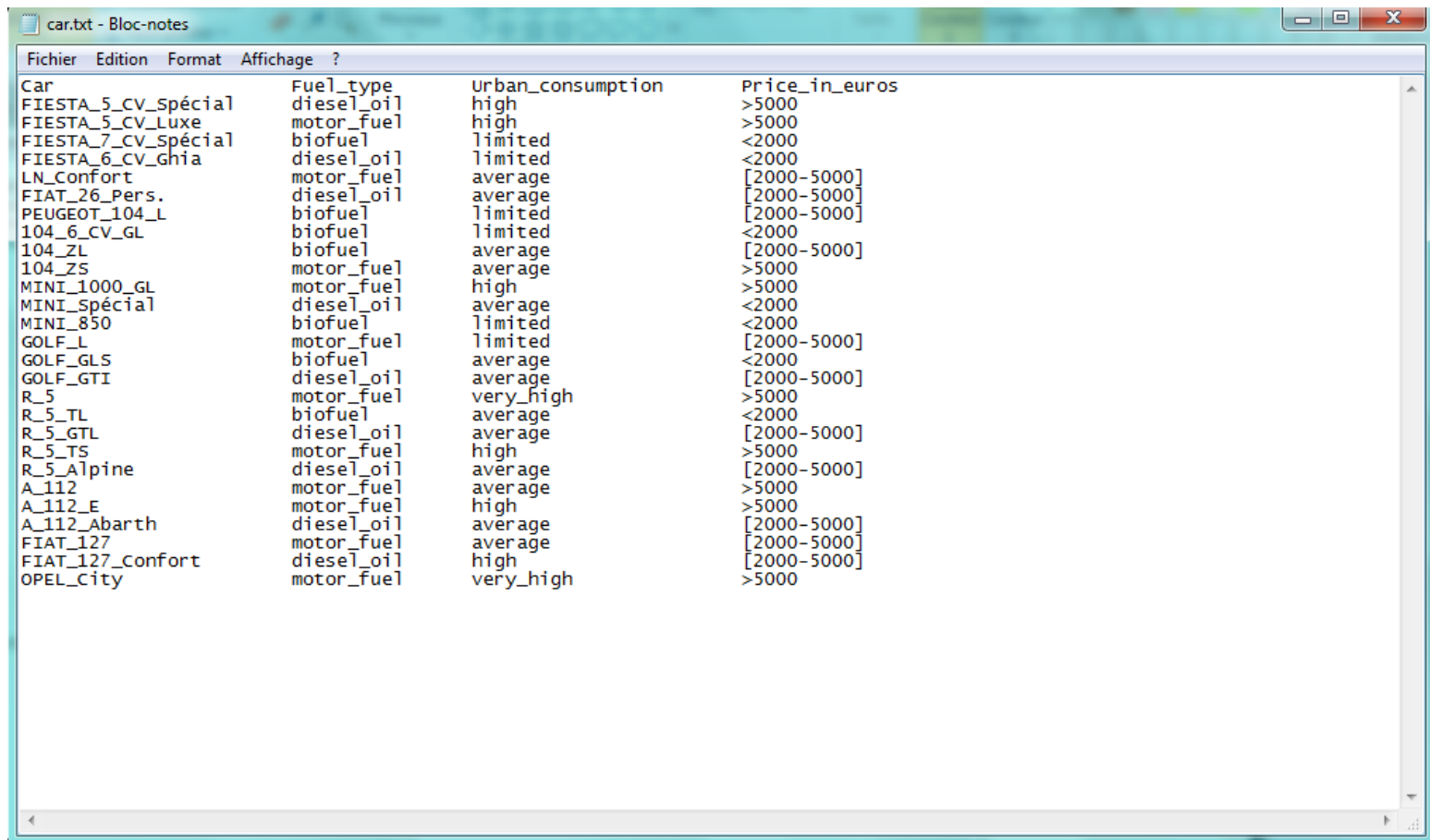
La plupart du temps, ces fichiers sont au format texte («.txt») ou Excel («.xls» ou «.xlsx»).

Problème: impossible d'importer un fichier au format Excel directement sous **R**.

Solution: le convertir en fichier texte («.txt») ou en fichier séparateur point-virgule («.csv»), qui sont importables sous **R**.

# 5) Importation de fichiers de données (2)

On souhaite importer le fichier texte suivant sous R:



Car	Fuel_type	Urban_consumption	Price_in_euros
FIESTA_5_CV_Spécial	diesel_oil	high	>5000
FIESTA_5_CV_Luxe	motor_fuel	high	>5000
FIESTA_7_CV_Spécial	biofuel	limited	<2000
FIESTA_6_CV_Ghia	diesel_oil	limited	<2000
LN_Confort	motor_fuel	average	[2000-5000]
FIAT_26_Pers.	diesel_oil	average	[2000-5000]
PEUGEOT_104_L	biofuel	limited	[2000-5000]
104_6_CV_GL	biofuel	limited	<2000
104_ZL	biofuel	average	[2000-5000]
104_ZS	motor_fuel	average	>5000
MINI_1000_GL	motor_fuel	high	>5000
MINI_Spécial	diesel_oil	average	<2000
MINI_850	biofuel	limited	<2000
GOLF_L	motor_fuel	limited	[2000-5000]
GOLF_GLS	biofuel	average	<2000
GOLF_GTI	diesel_oil	average	[2000-5000]
R_5	motor_fuel	very_high	>5000
R_5_TL	biofuel	average	<2000
R_5_GTL	diesel_oil	average	[2000-5000]
R_5_TS	motor_fuel	high	>5000
R_5_Alpine	diesel_oil	average	[2000-5000]
A_112	motor_fuel	average	>5000
A_112_E	motor_fuel	high	>5000
A_112_Abarth	diesel_oil	average	[2000-5000]
FIAT_127	motor_fuel	average	[2000-5000]
FIAT_127_Confort	diesel_oil	high	[2000-5000]
OPEL_City	motor_fuel	very_high	>5000

## 5) Importation de fichiers de données (3)

Pour cela, on utilise la fonction **read.table** avec comme argument l'emplacement du fichier sur l'ordinateur:

```
> car<-read.table("D:/cohene/Desktop/STATNUM TRADD 2019-2020/car.txt",header=TRUE)
> car
```

	Car	Fuel_type	Urban_consumption	Price_in_euros
1	FIESTA_5_CV_Spécial	diesel_oil	high	>5000
2	FIESTA_5_CV_Luxe	motor_fuel	high	>5000
3	FIESTA_7_CV_Spécial	biofuel	limited	<2000
4	FIESTA_6_CV_Ghia	diesel_oil	limited	<2000
5	LN_Confort	motor_fuel	average	[2000-5000]
6	FIAT_26_Pers.	diesel_oil	average	[2000-5000]
7	PEUGEOT_104_L	biofuel	limited	[2000-5000]
8	104_6_CV_GL	biofuel	limited	<2000
9	104_ZL	biofuel	average	[2000-5000]
10	104_ZS	motor_fuel	average	>5000
11	MINI_1000_GL	motor_fuel	high	>5000
12	MINI_Spécial	diesel_oil	average	<2000
13	MINI_850	biofuel	limited	<2000
14	GOLF_L	motor_fuel	limited	[2000-5000]
15	GOLF_GLS	biofuel	average	<2000
16	GOLF_GTI	diesel_oil	average	[2000-5000]
17	R_5	motor_fuel	very_high	>5000
18	R_5_TL	biofuel	average	<2000
19	R_5_GTL	diesel_oil	average	[2000-5000]
20	R_5_TS	motor_fuel	high	>5000
21	R_5_Alpine	diesel_oil	average	[2000-5000]
22	A_112	motor_fuel	average	>5000
23	A_112_E	motor_fuel	high	>5000
24	A_112_Abarth	diesel_oil	average	[2000-5000]
25	FIAT_127	motor_fuel	average	[2000-5000]
26	FIAT_127_Confort	diesel_oil	high	[2000-5000]
27	OPEL_City	motor_fuel	very_high	>5000

Note: *header=TRUE* signifie que la première ligne du fichier importé correspond aux noms des variables de la table.

## 5) Importation de fichiers de données (4)

On peut également créer un répertoire dans lequel on va lire les fichiers de données qu'il contient à l'aide de la fonction **setwd**.

```
> setwd("D:/cohene/Desktop/STATNUM TRADD 2019-2020")
> car<-read.table("car.txt",header=TRUE)
> car
```

	Car	Fuel_type	Urban_consumption	Price_in_euros
1	FIESTA_5_CV_Spécial	diesel_oil	high	>5000
2	FIESTA_5_CV_Luxe	motor_fuel	high	>5000
3	FIESTA_7_CV_Spécial	biofuel	limited	<2000
4	FIESTA_6_CV_Ghia	diesel_oil	limited	<2000
5	LN_Confort	motor_fuel	average	[2000-5000]
6	FIAT_26_Pers.	diesel_oil	average	[2000-5000]
7	PEUGEOT_104_L	biofuel	limited	[2000-5000]
8	104_6_CV_GL	biofuel	limited	<2000
9	104_ZL	biofuel	average	[2000-5000]
10	104_ZS	motor_fuel	average	>5000
11	MINI_1000_GL	motor_fuel	high	>5000
12	MINI_Spécial	diesel_oil	average	<2000
13	MINI_850	biofuel	limited	<2000
14	GOLF_L	motor_fuel	limited	[2000-5000]
15	GOLF_GLS	biofuel	average	<2000
16	GOLF_GTI	diesel_oil	average	[2000-5000]
17	R_5	motor_fuel	very_high	>5000
18	R_5_TL	biofuel	average	<2000
19	R_5_GTL	diesel_oil	average	[2000-5000]
20	R_5_TS	motor_fuel	high	>5000
21	R_5_Alpine	diesel_oil	average	[2000-5000]
22	A_112	motor_fuel	average	>5000
23	A_112_E	motor_fuel	high	>5000
24	A_112_Abarth	diesel_oil	average	[2000-5000]
25	FIAT_127	motor_fuel	average	[2000-5000]
26	FIAT_127_Confort	diesel_oil	high	[2000-5000]
27	OPEL_City	motor_fuel	very_high	>5000



# 6) Importation de fichiers de données (5)

Dans l'exemple suivant, on travaille un fichier csv qui recense, sous la forme d'un tableau, le nombre de morts sur les routes en 2012 dans les pays de l'OCDE en fonction d'un certain nombre de critères (tranche d'âge, type de véhicule...).

Country	Year	All_road_users	From_0_to_14_years	From_15_to_17_years	From_18_to_20_years	From_21_to_24_years	From_25_to_64_years	Above_64_years	Home_population_in_1000
Argentina	2012	5104	322	373	369	483	3118	439	41282
Australia	2012	1301	49	44	115	113	732	245	22710
Austria	2012	531	8	24	43	41	261	154	8408
Belgium	2012	767	19	18	50	78	428	177	11095
Cambodia	2012	1966	155	73	257	325	1075	77	14303
Canada	2012	2077							34754
Czech Republic	2012	742	15	17	40	56	455	157	10462
Denmark	2012	167	7	6	20	11	79	44	5581
Finland	2012	255	7	14	15	26	135	58	5401
France	2012	3653	115	131	334	419	1909	745	63379
Germany	2012	3600	73	113	262	349	1809	994	81844
Greece	2012	1027							11123
Hungary	2012	605	21	11	18	27	407	121	9932
Iceland	2012	9	0	0	0	2	3	4	320
Ireland	2012	162	3	7	12	23	81	36	4583
Israel	2012	263	23	15	18	21	123	55	7766
Italy	2012	3653	51	82	160	251	1994	1050	59394
Japan	2012	5237	98	97	201	166	1927	2748	127513
Korea	2012	5392	101	107	109	198	3011	1859	50004
Luxembourg	2012	34	1	2	3	3	16	9	525
Netherlands	2012	562	24	13	27	49	262	187	16730
New Zealand	2012	308	14	12	28	21	160	70	4433
Norway	2012	145	4	6	12	8	87	28	4986
Poland	2012	3571	89	86	250	335	2150	655	38538
Portugal	2012	718	13	10	18	47	429	195	10542
Slovenia	2012	130	3	1	8	11	81	26	2055
Spain	2012	1903	52	32	54	120	1122	507	46818
Sweden	2012	285	7	10	16	25	156	71	9556
Switzerland	2012	339	31	6	22	17	170	93	7955
United Kingdom	2012	1802	56	66	161	183	914	422	63705
USA	2012	33561	1168	1086	2333	3436	19917	5621	313914

# 5) Importation de fichiers de données (6)

Sous R, on nommera ce fichier *data*.

On se sert de la commande **read.csv2** en indiquant l'emplacement du fichier sur la machine:

```
> data<-read.csv2("D:/cohene/Desktop/STATNUM TRADD 2019-2020/mortalité_routes_OCDE_2012.csv")
> data
```

	Country	Year	All_road_users	From_0_to_14_years	From_15_to_17_years	From_18_to_20_years	From_21_to_24_years
1	Argentina	2012	5104	322	373	369	483
2	Australia	2012	1301	49	44	115	113
3	Austria	2012	531	8	24	43	41
4	Belgium	2012	767	19	18	50	78
5	Cambodia	2012	1966	155	73	257	325
6	Canada	2012	2077	NA	NA	NA	NA
7	Czech Republic	2012	742	15	17	40	56
8	Denmark	2012	167	7	6	20	11
9	Finland	2012	255	7	14	15	26
10	France	2012	3653	115	131	334	419
11	Germany	2012	3600	73	113	262	349
12	Greece	2012	1027	NA	NA	NA	NA
13	Hungary	2012	605	21	11	18	27
14	Iceland	2012	9	0	0	0	2
15	Ireland	2012	162	3	7	12	23
16	Israel	2012	263	23	15	18	21
17	Italy	2012	3653	51	82	160	251
18	Japan	2012	5237	98	97	201	166
19	Korea	2012	5392	101	107	109	198
20	Luxembourg	2012	34	1	2	3	3
21	Netherlands	2012	562	24	13	27	49
22	New Zealand	2012	308	14	12	28	21
23	Norway	2012	145	4	6	12	8
24	Poland	2012	3571	89	86	250	335
25	Portugal	2012	718	13	10	18	47
26	Slovenia	2012	130	3	1	8	11
27	Spain	2012	1903	52	32	54	120
28	Sweden	2012	285	7	10	16	25
29	Switzerland	2012	339	31	6	22	17
30	United Kingdom	2012	1802	56	66	161	183
31	USA	2012	33561	1168	1086	2333	3436

# 5) Importation de fichiers de données (7)

On peut également utiliser le répertoire, comme pour le fichier *car.txt*

```
> data<-read.csv2("mortalité_routes_OCDE_2012.csv")
```

```
> data
```

	Country	Year	All_road_users	From_0_to_14_years	From_15_to_17_years	From_18_to_20_years	From_21_to_24_years
1	Argentina	2012	5104	322	373	369	483
2	Australia	2012	1301	49	44	115	113
3	Austria	2012	531	8	24	43	41
4	Belgium	2012	767	19	18	50	78
5	Cambodia	2012	1966	155	73	257	325
6	Canada	2012	2077	NA	NA	NA	NA
7	Czech Republic	2012	742	15	17	40	56
8	Denmark	2012	167	7	6	20	11
9	Finland	2012	255	7	14	15	26
10	France	2012	3653	115	131	334	419
11	Germany	2012	3600	73	113	262	349
12	Greece	2012	1027	NA	NA	NA	NA
13	Hungary	2012	605	21	11	18	27
14	Iceland	2012	9	0	0	0	2
15	Ireland	2012	162	3	7	12	23
16	Israel	2012	263	23	15	18	21
17	Italy	2012	3653	51	82	160	251
18	Japan	2012	5237	98	97	201	166
19	Korea	2012	5392	101	107	109	198
20	Luxembourg	2012	34	1	2	3	3
21	Netherlands	2012	562	24	13	27	49
22	New Zealand	2012	308	14	12	28	21
23	Norway	2012	145	4	6	12	8
24	Poland	2012	3571	89	86	250	335
25	Portugal	2012	718	13	10	18	47
26	Slovenia	2012	130	3	1	8	11
27	Spain	2012	1903	52	32	54	120
28	Sweden	2012	285	7	10	16	25
29	Switzerland	2012	339	31	6	22	17
30	United Kingdom	2012	1802	56	66	161	183
31	USA	2012	33561	1168	1086	2333	3436

## 5) Importation de fichiers de données (8)

1 ligne = 1 individu

=> dans l'exemple suivant, un pays de l'OCDE, l'Argentine, pour lequel on observe le nombre de morts sur les routes pour l'année 2012 selon les différentes caractéristiques (variables) prises en comptes dans l'étude:

```
> data[1,]  
Country Year All_road_users From_0_to_14_years From_15_to_17_years From_18_to_20_years  
1 Argentina 2012 5104 322 373 369  
From_21_to_24_years From_25_to_64_years Above_64_years Home_population_in_1000  
1 483 3118 439 41282  
Network_length_of_all_public_roads_in_km Whereof_motorways Number_of_motor_vehicles_in_10000  
1 237849 2456 20645  
Whereof_powered_two_wheelers Whereof_passenger_cars Area_of_state_in_sqkm Pedestrians  
1 4774 15871 2736690 NA  
Bicyclists Powered_2wheelers Passenger_car_occupants Road_outside_urban_areas  
1 NA NA NA 2399  
Whereof_motorways.1  
1 NA
```

## 5) Importation de fichiers de données (9)

**1 colonne = 1 variable**

=> dans l'exemple suivant, il s'agit du nombre de morts sur les routes en 2012 pour la tranche d'âge 18-20 ans pour l'ensemble des pays de l'OCDE:

```
> data[,6]
 [1] 369 115 43 50 257 NA 40 20 15 334 262 NA 18 0 12
[16] 18 160 201 109 3 27 28 12 250 18 8 54 16 22 161
[31] 2333
```

On peut également faire un « zoom » sur cette variable en jouant sur son nom et non pas sa position dans le tableau de données comme

```
> data$From_18_to_20_years
 [1] 369 115 43 50 257 NA 40 20 15 334 262 NA 18 0 12
[16] 18 160 201 109 3 27 28 12 250 18 8 54 16 22 161
[31] 2333
```

# 6) Retraitement (1)

Difficulté: dans environ **80% des cas**, fichiers **bruts** et **pas toujours bien renseignés/construits** (valeurs manquantes, doublons, incohérences...).

Solution: travail de **retraitement** (souvent long et fastidieux) qui va conditionner par la suite la **qualité des analyses**.

➤ **Etape cruciale** avant d'avant **d'explorer les données**.

**i) descriptif du jeu de données importé**

`dim(data)`

*#dimension de **data**#*

`names(data)`

*#noms des variables de **data**#*

`summary(data)`

*#résumé statistique (non exhaustif) de chacune des variables de **data**#*

# 6) Retraitement (2)

## ii) troncatures, permutations et concaténations

```
data_v2<- data[3:15,]
```

*#sous-ensemble de **data** n'incluant que les observations allant de la 3<sup>ème</sup> à la 15<sup>ème</sup> ligne#*

```
data_v3<- data[,1:9]
```

*#sous-ensemble de **data** n'incluant que les variables allant de la 1<sup>ère</sup> à la 9<sup>ème</sup> colonne#*

```
data_v4<- data[8:13,5:11]
```

*#sous-ensemble de **data** n'incluant que les observations allant de la 8<sup>ème</sup> à la 13<sup>ème</sup> ligne **et** les variables allant de la 5<sup>ème</sup> à la 11<sup>ème</sup> colonne#*

```
data_v5<- data.frame(data[,1],data[,3],data[,7],data[,11],data[,15:19])
```

*#table de données formée à partir des 1<sup>ère</sup>, 3<sup>ème</sup>, 7<sup>ème</sup> et 11<sup>ème</sup> colonnes de **data** ainsi que du sous-ensemble de **data** n'incluant que les variables allant de la 15<sup>ème</sup> à la 19<sup>ème</sup> colonne#*

## 6) Retraitement (3)

```
> data_v5
data...1. data...3. data...7. data...11. Whereof_passenger_cars Area_of_state_in_sqkm Pedestrians Bicyclists Powered_2wheelers
1 Argentina 5104 483 237849 15871 2736690 NA NA NA
2 Australia 1301 113 900082 12714 7682300 174 33 221
3 Austria 531 41 124588 4513 82409 81 52 86
4 Belgium 767 78 154575 5444 30280 104 68 102
5 Cambodia 1966 325 NA 217 176520 207 77 1340
6 Canada 2077 NA 1408800 20680 9093510 NA NA NA
7 Czech Republic 742 56 55742 4582 77240 163 78 93
8 Denmark 167 11 73929 2198 42430 31 22 24
9 Finland 255 26 78109 3057 303890 29 19 28
10 France 3653 419 NA 31575 547660 489 164 843
11 Germany 3600 349 NA 42928 348570 520 406 679
```

On peut voir ci-dessus que la concaténation de certaines colonnes (les quatre premières) n'a pas permis de conserver les noms des attributs qu'elles représentent. Il nous faut donc les renommer:

```
names(data_v5)[1]<-c("Country") #on renomme la 1ère colonne de data_v5 de la même façon que la variable dans data#
names(data_v5)[2]<-c("All_road_users")
names(data_v5)[3]<-c("From_21_to_24_years")
names(data_v5)[4]<-c("Network_length_of_all_public_roads_in_km")
```

```
> data_v5
Country All_road_users From_21_to_24_years Network_length_of_all_public_roads_in_km Whereof_passenger_cars Area_of_state_in_sqkm Pedestrians Bicyclists Powered_2wheelers
1 Argentina 5104 483 237849 15871 2736690 NA NA NA
2 Australia 1301 113 900082 12714 7682300 174 33 221
3 Austria 531 41 124588 4513 82409 81 52 86
4 Belgium 767 78 154575 5444 30280 104 68 102
5 Cambodia 1966 325 NA 217 176520 207 77 1340
6 Canada 2077 NA 1408800 20680 9093510 NA NA NA
7 Czech Republic 742 56 55742 4582 77240 163 78 93
8 Denmark 167 11 73929 2198 42430 31 22 24
9 Finland 255 26 78109 3057 303890 29 19 28
10 France 3653 419 NA 31575 547660 489 164 843
11 Germany 3600 349 NA 42928 348570 520 406 679
```



# 6) Retraitement (4)

## iii) traitement des valeurs manquantes

```
which(is.na(data), arr.ind=TRUE)
```

*#indique l'emplacement (ligne, colonne) dans le jeu de données **data** de toutes les valeurs manquantes, notées NA#*

⇒ On peut voir que la valeur de la 6<sup>ème</sup> ligne et de la 4<sup>ème</sup> colonne de **data** est manquante. On peut donc lui imputer une valeur par défaut (par exemple la moyenne ou la médiane).

```
summary(data[,4])
```

*#on regarde d'abord le résumé statistique de la 4<sup>ème</sup> colonne#*

```
data[6,4]<- 87.21
```

*#on impute la moyenne de la variable à l'observation en question#*

```
data[,4]<-data[,4]
```

*#on actualise la colonne afin que la valeur précédente soit prise en compte dans **data**#*

```
data
```

*#la modification a bien été opérée#*

# 6) Retraitement (5)

## iv) conditions

***Pour ne conserver que les observations prenant une certaine valeur pour une variable numérique donnée***

```
data_v7<-data[ which(data$All_road_users<1000),]  
#on crée l'objet data_v7 qui est un sous-ensemble du tableau initial data et pour lequel les pays recensés enregistrent moins de 1000 morts sur les routes en 2012 tous usagers confondus#
```

***Pour ne conserver que les observations prenant une certaine valeur pour une variable qualitative donnée***

```
car_v2<-car[ which(car$Fuel_type=="diesel_oil"), ]  
#on crée l'objet car_v2 qui est un sous-ensemble du tableau initial car et pour lequel tous les véhicules recensés carburent au diesel#
```

```
car_v3<-car[ which((car$Fuel_type=="diesel_oil")&(car$Price_in_euros!="<2000")), ]  
#on crée l'objet car_v3 qui est un sous-ensemble du tableau initial car et pour lequel tous les véhicules recensés carburent au diesel et coûtent plus de 2000 euros#
```

## 6) Retraitement (6)

*Pour supprimer toutes les observations pour lesquelles on recense une valeur manquante pour une variable donnée*

```
data_v8<-data[ which(data$Pedestrians!="NA"),]
```

#on crée l'objet **data\_v8** qui est un sous-ensemble du tableau initial **data** et pour lequel on dispose de l'information concernant le nombre de piétons morts dans un accident de la route en 2012#

*Pour ne conserver que les observations pour lesquelles on recense une valeur manquante pour une variable donnée*

```
data_v9<-data[ is.na(data$Pedestrians),]
```

#on crée l'objet **data\_v9** qui est un sous-ensemble du tableau initial **data** et pour lequel on ne dispose pas de l'information concernant le nombre de piétons morts dans un accident de la route en 2012#