

Les éléments à télécharger sont disponibles à l'adresse suivante :  
**<http://www.editions-eni.fr>**  
Saisissez la référence ENI de l'ouvrage **RISTATR** dans la zone de recherche et validez. Cliquez sur le titre du livre puis sur le bouton de téléchargement.

## Avant-propos

### Chapitre 1 Introduction à R

- 1. Introduction. . . . . 15
- 2. R, un langage statistique. . . . . 15
  - 2.1 Histoire de R. . . . . 16
  - 2.2 Les caractéristiques de R. . . . . 16
  - 2.3 Installer une distribution R. . . . . 17
    - 2.3.1 R-Project de CRAN. . . . . 18
    - 2.3.2 Autres distributions R. . . . . 21
- 3. Quelques IDE pratiques . . . . . 22
  - 3.1 RStudio . . . . . 23
    - 3.1.1 Téléchargement et installation . . . . . 24
    - 3.1.2 Exploration de RStudio Desktop . . . . . 24
    - 3.1.3 Créer un projet avec RStudio . . . . . 33
  - 3.2 Outils R pour Visual Studio (RTVS). . . . . 37
    - 3.2.1 Installation de RTVS avec VS 2017 . . . . . 38
    - 3.2.2 Découverte de RTVS. . . . . 41

### Chapitre 2 Types de variables et opérations R

- 1. Introduction. . . . . 51
- 2. Session interactive R . . . . . 51
  - 2.1 La console R . . . . . 52
    - 2.1.1 Exécution d'une instruction . . . . . 52
    - 2.1.2 Commentaire R . . . . . 54
    - 2.1.3 Configuration des sorties ou affichage des résultats. . . . . 54
    - 2.1.4 Quitter la console . . . . . 56

2.2	L'aide interactive R	56
2.2.1	L'aide de base	56
2.2.2	Les vignettes et démonstrations	61
3.	Les variables R	63
3.1	Création de variables	63
3.2	Variables et espace de travail	66
3.2.1	Le cycle de vie des variables	66
3.2.2	Gestion de l'espace de travail	68
3.3	Les noms de variables et mots réservés	73
3.3.1	Les noms de variables	73
3.3.2	Les mots réservés	74
3.4	Les constantes natives R	75
4.	Types de vecteur atomique	76
4.1	Les nombres	76
4.1.1	Créer des objets numériques	76
4.1.2	Créer des nombres complexes	78
4.1.3	Opérateurs et opérations arithmétiques	80
4.2	Les chaînes de caractères	82
4.3	Les booléens	84
5.	Les dates et heures	87
5.1	La classe Date	88
5.1.1	Création des dates	88
5.1.2	Les opérations sur les dates	89
5.2	La classe POSIXt	90
5.2.1	Création d'objet POSIXct et POSIXcl	91
5.2.2	Opération sur les objets POSIXt	93
6.	Les valeurs ou données spéciales	94
6.1	Les valeurs manquantes ou Missing data	94
6.2	Inf ou infini	95
6.3	NaN ou « Not a Number »	96
6.4	NULL	96
7.	Les packages	97
7.1	Gestion des packages	97
7.1.1	Télécharger et installer un package	97
7.1.2	La mise à jour et la suppression des packages	99

- 7.2 Charger et télécharger un package . . . . . 100
  - 7.2.1 Charger le package dans l'environnement . . . . . 100
  - 7.2.2 Solliciter le package sans le charger . . . . . 102
  - 7.2.3 Télécharger un package . . . . . 103
- 7.3 Création de package . . . . . 103

**Chapitre 3**  
**Structure des données avec R**

- 1. Introduction . . . . . 105
- 2. Les vecteurs . . . . . 105
  - 2.1 Création de vecteur . . . . . 106
  - 2.2 Manipulations d'un vecteur . . . . . 107
    - 2.2.1 Indexation numérique d'un vecteur . . . . . 108
    - 2.2.2 Indexation booléenne d'un vecteur . . . . . 111
    - 2.2.3 Indexation nommée d'un vecteur . . . . . 112
    - 2.2.4 Quelques fonctions d'extraction ou d'indexation . . . . . 114
  - 2.3 Travailler avec des vecteurs numériques . . . . . 115
    - 2.3.1 Générer des vecteurs séquentiels . . . . . 115
    - 2.3.2 Les opérations sur les vecteurs numériques . . . . . 117
  - 2.4 Créer des séries de dates . . . . . 127
  - 2.5 Les fonctions de vecteur de chaînes de caractères . . . . . 128
    - 2.5.1 La fonction paste() . . . . . 128
    - 2.5.2 La fonction cat() . . . . . 130
    - 2.5.3 La fonction sprintf() . . . . . 131
    - 2.5.4 La fonction nchar() . . . . . 131
    - 2.5.5 Les fonctions substr() et substring() . . . . . 132
    - 2.5.6 Les fonctions toupper() et tolower() . . . . . 132
    - 2.5.7 Les fonctions chartr() et replace() . . . . . 133
    - 2.5.8 La fonction strsplit() . . . . . 133
    - 2.5.9 Les fonctions grep() et grepl() . . . . . 134
    - 2.5.10 Les fonctions regexpr() et gregexpr() . . . . . 135
    - 2.5.11 La fonction gsub() . . . . . 137
  - 2.6 Cas des vecteurs booléens . . . . . 139

2.7	Autres opérations génériques sur les vecteurs . . . . .	140
2.7.1	Suppression de doublons . . . . .	140
2.7.2	Les opérations ensemblistes . . . . .	141
2.7.3	La gestion de valeurs manquantes . . . . .	142
3.	Les matrices . . . . .	143
3.1	Construction d'une matrice . . . . .	143
3.1.1	Le constructeur <code>matrix()</code> . . . . .	143
3.1.2	Combiner des vecteurs en matrice . . . . .	145
3.1.3	Convertir un objet en matrice. . . . .	146
3.2	Dimension d'une matrice . . . . .	147
3.3	Nommer les lignes et colonnes d'une matrice. . . . .	148
3.4	Manipuler une matrice . . . . .	150
3.4.1	Indexation numérique des matrices . . . . .	150
3.4.2	Indexation booléenne des matrices. . . . .	155
3.4.3	Indexation par les noms des colonnes et des lignes. . . . .	156
3.5	Calculs matriciels . . . . .	158
3.5.1	Opérations arithmétiques . . . . .	159
3.5.2	Les multiplications matricielles. . . . .	160
3.5.3	Les matrices spéciales. . . . .	162
3.5.4	La transposée, le déterminant et l'inverse d'une matrice. . . . .	164
3.5.5	Valeurs et vecteurs propres d'une matrice . . . . .	166
3.5.6	Décomposition d'une matrice . . . . .	166
4.	Les tableaux multidimensionnels avec <code>array()</code> . . . . .	170
4.1	Création d'un tableau multidimensionnel. . . . .	171
4.2	Nommer un tableau multidimensionnel. . . . .	172
4.3	Extraction d'éléments d'un tableau multidimensionnel. . . . .	173
5.	Les facteurs ou variables catégoriels . . . . .	174
5.1	Création d'un facteur . . . . .	174
5.2	Modification des modalités d'un facteur. . . . .	178
5.3	Cas des variables ordinales . . . . .	179
5.4	Cas des variables continues . . . . .	180
5.5	Les variables catégorielles et la fonction <code>table()</code> . . . . .	181
5.6	Les dates et la fonction <code>factor()</code> . . . . .	184

- 6. Les data frame ..... 185
  - 6.1 Description d'un data frame ..... 185
  - 6.2 Construction d'un data frame ..... 192
  - 6.3 Dimension d'un data frame ..... 195
  - 6.4 Nommer les colonnes et lignes d'un data frame. .... 195
  - 6.5 Manipulation d'un data frame. .... 197
    - 6.5.1 Extraction de colonnes ou variables ..... 197
    - 6.5.2 Extraction des lignes ou des individus ..... 203
    - 6.5.3 Extraire un élément ou une valeur quelconque ..... 207
    - 6.5.4 Ajout de colonnes ou de lignes dans un data frame ..... 207
    - 6.5.5 Suppression des colonnes ou de lignes ..... 211
    - 6.5.6 Empiler et désempiler un data frame ..... 212
    - 6.5.7 La transposition de data frame avec reshape(). .... 214
    - 6.5.8 Le tri d'un data frame ..... 218
    - 6.5.9 La fonction aggregate() ..... 220
    - 6.5.10 La fonction summary() ..... 222
    - 6.5.11 La fonction by() ..... 224
    - 6.5.12 La famille des fonctions apply() ..... 224
- 7. Les listes ..... 225
  - 7.1 Construction d'une liste. .... 225
  - 7.2 Indexation des objets d'une liste ..... 227
    - 7.2.1 Extraction à l'intérieur des crochets ou doubles crochets .... 227
    - 7.2.2 Extraction avec l'opérateur dollar \$ ..... 230
    - 7.2.3 Extraction avec la structure attach() ... detach() ..... 230

**Chapitre 4**  
**La programmation avec R**

- 1. Introduction. .... 231
- 2. Les structures de contrôle ..... 231
  - 2.1 Les structures de groupage d'instructions ..... 232
    - 2.1.1 Le point-virgule ..... 232
    - 2.1.2 Les accolades ..... 232
  - 2.2 Les structures conditionnelles ..... 233
    - 2.2.1 La structure if...else ..... 233
    - 2.2.2 La structure switch() ..... 236

2.3	Les structures de répétition ou boucles	237
2.3.1	La boucle for et les commandes break, next	237
2.3.2	La boucle while	240
2.3.3	La boucle repeat ... break	241
3.	Les fonctions	241
3.1	Création d'une fonction	242
3.2	Les arguments par défaut	245
3.3	La commande return()	246
3.4	Les fonctions avec des arguments ... (indéfinis)	249
3.5	Les fonctions avec do.call()	250
3.6	Les fonctions anonymes	250
3.7	Les fonctions remplacement	251
3.8	Les fonctions infixes	252
3.9	Les fonctions récursives	253
4.	L'environnement R et la portée des objets	254
5.	Optimisation du code	257
5.1	La vectorisation	257
5.1.1	La fonction apply()	257
5.1.2	La fonction sapply() ou lapply() et mapply()	258
5.1.3	La fonction rapply()	261
5.1.4	La fonction tapply()	263
5.1.5	La fonction eapply()	264
5.1.6	La fonction replicate()	265
5.2	R et le langage C/C++	266
5.3	Gestion des exceptions et des erreurs	268
5.3.1	Alerter sur une erreur	269
5.3.2	Gérer les exceptions avec try()	271
5.3.3	Gérer les exceptions avec tryCatch()	272
6.	La programmation orientée objets (POO) avec R	274
6.1	Classe S3	274
6.1.1	Définition d'une classe	275
6.1.2	Définir un constructeur	277
6.1.3	Définir des méthodes membres d'une classe	279
6.1.4	Modèle UML de la classe stock()	283

- 6.2 Classe S4 ..... 289
  - 6.2.1 Définition de la classe et ses attributs ..... 290
  - 6.2.2 Constructeur new()..... 291
  - 6.2.3 Définir les méthodes membres de classes. .... 293
- 6.3 Classe R5 ou RC ..... 298
  - 6.3.1 Créer une classe R5 ou RC ..... 299
  - 6.3.2 Les méthodes membres de la classe ..... 301
- 7. Conception de package ou R. .... 305
  - 7.1 Structure d'un package ..... 305
    - 7.1.1 Installer les outils. .... 305
    - 7.1.2 Générer l'architecture du package. .... 306
  - 7.2 Le fichier DESCRIPTION. .... 307
  - 7.3 Doter le package de jeux de données. .... 310
    - 7.3.1 Créer le dossier data. .... 310
    - 7.3.2 Documenter les jeux de données ..... 312
  - 7.4 Ajouter les fonctions ..... 314
  - 7.5 Créer une vignette ..... 316
  - 7.6 Documenter le package ..... 316
  - 7.7 Installer et utiliser le package. .... 317
  - 7.8 Partager le package ..... 321

**Chapitre 5**

**Importer et exporter des données avec R**

- 1. Introduction. .... 323
- 2. Saisir/éditer les données depuis le clavier. .... 323
  - 2.1 Les fonctions readline(), scan() et readClipboard() ..... 324
    - 2.1.1 La fonction readline() ..... 324
    - 2.1.2 La fonction scan() ..... 325
    - 2.1.3 La fonction readClipboard() ..... 327
  - 2.2 La fonction textConnection() ..... 328
  - 2.3 Les fonctions edit()/fix() et la fonction data.entry() ..... 329
    - 2.3.1 La fonction edit()..... 329
    - 2.3.2 La fonction fix()..... 333
    - 2.3.3 La fonction data.entry() ..... 334

3.	Accéder aux données depuis un fichier de données . . . . .	334
3.1	Fichier texte délimité . . . . .	334
3.1.1	La fonction <code>read.table()</code> . . . . .	334
3.2	Fichier CSV . . . . .	340
3.3	Fichier Excel . . . . .	344
3.3.1	Package <code>xlsx</code> . . . . .	344
3.3.2	Package <code>XLConnect</code> . . . . .	346
3.3.3	Package <code>readxl</code> . . . . .	348
3.4	Fichier XML . . . . .	350
3.4.1	Analyse de la structure d'un fichier XML . . . . .	351
3.4.2	La conversion en data frame . . . . .	353
3.5	Fichier JSON . . . . .	356
4.	Importer des données d'autres logiciels d'analyse de données . . . . .	357
4.1	Logiciel IBM SPSS : <code>*.sav</code> . . . . .	357
4.1.1	Le package <code>foreign</code> . . . . .	357
4.1.2	Le package <code>haven</code> . . . . .	358
4.2	Logiciel SAS : <code>*.sas7bdat</code> . . . . .	359
4.2.1	Le package <code>sas7bdat</code> . . . . .	359
4.2.2	Le package <code>haven</code> . . . . .	359
4.3	Logiciel Stata : <code>*.dta</code> . . . . .	360
4.4	Le logiciel MATLAB : <code>*.mat</code> . . . . .	360
5.	Télécharger des données depuis le Web . . . . .	362
5.1	Télécharger un fichier de données grâce à l'URL . . . . .	363
5.2	Extraire les données d'un tableau HTML . . . . .	364
5.3	Importer spécialement des données financières . . . . .	369
6.	Accéder aux données depuis une base de données . . . . .	372
6.1	Base de données MySQL . . . . .	372
6.1.1	Le package <code>RMySQL</code> . . . . .	373
6.1.2	Le package <code>RODBC</code> . . . . .	375
6.2	Base de données Microsoft Access . . . . .	379
6.3	Importer de gros volumes de données . . . . .	381
6.3.1	Le package <code>data.table</code> . . . . .	382
6.3.2	Le package <code>readr</code> . . . . .	385
6.3.3	Le package <code>sqldf</code> . . . . .	386



- 7. Exporter des données. . . . . 387
  - 7.1 Fichier texte délimité . . . . . 387
  - 7.2 Fichier CSV . . . . . 388
  - 7.3 Fichier Microsoft Excel . . . . . 389
  - 7.4 Exporter les données sous les formats de données R . . . . . 392
    - 7.4.1 Les fonctions `save()` et `save.image()` . . . . . 393
    - 7.4.2 Les fonctions `dput()` et `dump()` . . . . . 394
  - 7.5 Exporter des données sous des formats lisibles  
par d'autres logiciels statistiques . . . . . 395
    - 7.5.1 Le package `haven` . . . . . 395
    - 7.5.2 Le package `foreign` . . . . . 395

**Chapitre 6**  
**Introduction à l'analyse de données**

- 1. Introduction. . . . . 397
- 2. Préparation des jeux de données. . . . . 398
  - 2.1 Charger et comprendre les données. . . . . 398
  - 2.2 La gestion des données manquantes . . . . . 399
    - 2.2.1 Inspecter l'existence des données manquantes . . . . . 399
    - 2.2.2 Le traitement des données manquantes. . . . . 403
- 3. Analyse descriptive des variables quantitatives. . . . . 410
  - 3.1 Cas univarié. . . . . 411
    - 3.1.1 Mesures de tendance centrale. . . . . 411
    - 3.1.2 Mesures de dispersion . . . . . 418
    - 3.1.3 Mesures de forme. . . . . 424
    - 3.1.4 Indice de Gini et Courbe de Lorenz . . . . . 425
    - 3.1.5 Les fonctions sommaires des indicateurs . . . . . 427
  - 3.2 Cas bivarié. . . . . 430
    - 3.2.1 La covariance . . . . . 430
    - 3.2.2 Le coefficient de corrélation . . . . . 433
- 4. Analyse descriptive des variables catégorielles . . . . . 436
  - 4.1 Tableaux statistiques pour les variables catégorielles . . . . . 436
    - 4.1.1 Les tableaux de fréquences . . . . . 437
    - 4.1.2 Les tableaux de contingence . . . . . 441

4.2	Statistique des tableaux de contingence . . . . .	445
4.2.1	Les tableaux de fréquences et les statistiques marginales . . . . .	445
4.2.2	Le package prettyR. . . . .	447
4.2.3	Sommaire statistique avec les variables catégorielles . . . . .	452
5.	Visualisation graphique des données . . . . .	453
5.1	Explorer le package ggplot2 . . . . .	457
5.1.1	Aperçu graphique rapide avec qplot(). . . . .	457
5.1.2	Graphique avancé avec ggplot() . . . . .	460
5.2	Visualisation graphique avec ggplot2 . . . . .	472
5.2.1	Diagramme en barres. . . . .	473
5.2.2	Diagramme en secteurs . . . . .	477
5.2.3	Histogramme . . . . .	478
5.2.4	Boxplot ou diagramme en boîte à moustache . . . . .	479
5.2.5	Graphique à lignes . . . . .	481
5.2.6	Graphique à nuage de points. . . . .	482
5.2.7	Matrice de corrélation avec GGally. . . . .	484
5.2.8	Autres fonctions statistiques . . . . .	486

## Chapitre 7

### Probabilité et inférence paramétrique

1.	Introduction . . . . .	489
2.	Simulation . . . . .	490
2.1	Probabilité et dénombrement. . . . .	490
2.1.1	Arrangement . . . . .	490
2.1.2	Permutation . . . . .	494
2.1.3	Combinaison . . . . .	496
2.2	Tirage aléatoire d'échantillon . . . . .	500
2.2.1	La fonction sample() . . . . .	500
2.2.2	La reproductibilité des valeurs aléatoires . . . . .	503
2.2.3	Répétition d'une expérience avec replicate() . . . . .	504
2.3	Notion de variable aléatoire . . . . .	505
2.3.1	Cas d'une variable aléatoire discrète . . . . .	505
2.3.2	Cas d'une variable aléatoire continue . . . . .	511

- 2.4 Les fonctions R des distributions de probabilité. . . . . 521
  - 2.4.1 Vocabulaire des fonctions de distribution R . . . . . 521
  - 2.4.2 Tables des lois et leur équivalent R . . . . . 522
- 2.5 Quelques lois de densité univariées . . . . . 523
  - 2.5.1 La loi Normale . . . . . 523
  - 2.5.2 La loi de Student ou T- distribution. . . . . 525
  - 2.5.3 La loi de Khi-deux . . . . . 527
  - 2.5.4 La loi de Fisher . . . . . 528
- 2.6 Le théorème central limite . . . . . 529
- 3. Estimation par intervalle de confiance. . . . . 532
  - 3.1 Intervalle de confiance d'une moyenne . . . . . 532
    - 3.1.1 Formulation . . . . . 532
    - 3.1.2 Application avec R. . . . . 533
  - 3.2 Intervalle de confiance d'une médiane . . . . . 536
    - 3.2.1 Formulation . . . . . 536
    - 3.2.2 Application avec R. . . . . 537
  - 3.3 Intervalle de confiance d'une proportion . . . . . 538
    - 3.3.1 Formulation . . . . . 538
    - 3.3.2 Application avec R. . . . . 539
  - 3.4 Intervalle de confiance d'une variance. . . . . 541
    - 3.4.1 Formulation . . . . . 541
    - 3.4.2 Application avec R. . . . . 541
- 4. Tests d'hypothèses. . . . . 542
  - 4.1 Terminologie. . . . . 542
    - 4.1.1 Formulation d'un test . . . . . 542
    - 4.1.2 Les types d'erreurs et puissance d'un test. . . . . 543
    - 4.1.3 La statistique du test et la p-value . . . . . 544
  - 4.2 Tests d'indépendance . . . . . 544
    - 4.2.1 Test de Khi-deux . . . . . 544
    - 4.2.2 Test de Khi-deux avec correction de Yates . . . . . 547
    - 4.2.3 Simulation de Monte-Carlo . . . . . 548
    - 4.2.4 Test de Fisher exact . . . . . 549
  - 4.3 Tests d'hypothèses de la moyenne . . . . . 550
    - 4.3.1 Test de conformité . . . . . 550
    - 4.3.2 Test de comparaison de deux échantillons indépendants. . . . . 552
    - 4.3.3 Test de comparaison de deux échantillons appariés . . . . . 555
    - 4.3.4 ANOVA . . . . . 558

4.4	Tests d'hypothèses de la proportion	559
4.4.1	Test de conformité	559
4.4.2	Test de comparaison à échantillons indépendants	561
4.5	Tests d'hypothèses de la variance	563
4.5.1	Test de conformité	563
4.5.2	Test de comparaison pour deux échantillons	564
4.6	Tests d'hypothèse de la corrélation	565
4.6.1	Test de conformité	565
4.6.2	Test de comparaison	568
4.7	Analyse de puissance	570
4.7.1	Les fonctions R	571
4.7.2	Les fonctions du package pwr	571
4.8	Tests non paramétriques	574
4.8.1	Test de normalité Shapiro-Wilk	574
4.8.2	Test de distribution de Kolmogorov-Simionov	576
4.8.3	Test d'ajustement de $X^2$	578
4.8.4	Test de corrélation de Kendall et de Spearman	579
4.8.5	Test d'égalité de variance de Bartlett	580
4.8.6	Le test U de Mann-Whitney ou test de somme de rangs de Wilcoxon	580
4.8.7	Test de Kruskal-Wallis	581
5.	Méthodes de rééchantillonnage	583
5.1	Bootstrap	583
5.1.1	Démarche	583
5.1.2	Illustration avec le package boot	584
5.2	Jackknife	587
5.2.1	Démarche	587
5.2.2	Illustration avec le package bootstrap	588

**Chapitre 8**  
**Modèles statistiques classiques**

- 1. Introduction . . . . . 589
- 2. Analyse de la variance (ANOVA) . . . . . 590
  - 2.1 Expression de modèles avec les objets formula . . . . . 591
    - 2.1.1 Formulation d'un modèle . . . . . 591
    - 2.1.2 Les opérateurs R pour l'expression de relations . . . . . 593
    - 2.1.3 Quelques opérations et fonctions utiles aux objets formula . . . . . 594
  - 2.2 ANOVA à un facteur ou simple entrée . . . . . 595
    - 2.2.1 Modèle et hypothèses . . . . . 595
    - 2.2.2 ANOVA à un facteur avec R . . . . . 597
    - 2.2.3 Tests « post-hoc » . . . . . 605
  - 2.3 ANOVA à double facteur ou entrée . . . . . 606
    - 2.3.1 ANOVA à deux facteurs avec R . . . . . 608
    - 2.3.2 Tests « post-hoc » . . . . . 619
- 3. Modèles de régression linéaire . . . . . 620
  - 3.1 Régression linéaire simple . . . . . 621
    - 3.1.1 Formalisation et estimation . . . . . 621
    - 3.1.2 Modèle de régression linéaire simple avec R . . . . . 622
    - 3.1.3 Sommaire du résultat de lm() . . . . . 631
    - 3.1.4 Diagnostic de la régression . . . . . 637
    - 3.1.5 Détection des points ou observations atypiques . . . . . 644
    - 3.1.6 Prédiction . . . . . 652
  - 3.2 Régression linéaire multiple . . . . . 655
    - 3.2.1 Formalisation et estimation . . . . . 655
    - 3.2.2 Régression linéaire multiple avec R . . . . . 656
  - 3.3 Sélection de variables et choix du modèle . . . . . 671
    - 3.3.1 Les indicateurs ou critères de sélection . . . . . 672
    - 3.3.2 Algorithmes de sélection . . . . . 675
    - 3.3.3 Application avec R . . . . . 676
  - 3.4 Problème de colinéarité et régressions robustes . . . . . 690
    - 3.4.1 Définition et détection de la colinéarité . . . . . 690
    - 3.4.2 Les régressions régularisées . . . . . 694
    - 3.4.3 Les régressions régularisées en pratique . . . . . 695
    - 3.4.4 La régression sur composantes principales . . . . . 706

# 14 \_\_\_\_\_ Langage R

Prise en main des statistiques

**Conclusion** ..... 715

Index ..... 717

Les éléments à télécharger sont disponibles à l'adresse suivante :  
**<http://www.editions-eni.fr>**  
Saisissez la référence de l'ouvrage **EP3DASR** dans la zone de recherche  
et validez. Cliquez sur le titre du livre puis sur le bouton de téléchargement.

## Chapitre 1 Introduction

1. Data scientist, un métier à la mode . . . . .	19
2. Un nouveau métier : citizen data scientist . . . . .	20
2.1 Un objectif réaliste, devenir citizen data scientist . . . . .	20
2.2 Analyse et valorisation . . . . .	20
2.3 Animation et spécification . . . . .	21
2.4 Modélisation et inférence . . . . .	21
2.5 Déploiement et maintien en conditions opérationnelles . . . . .	22
3. Les data sciences . . . . .	23
4. Le Big Data . . . . .	24
5. La dynamique de cet ouvrage . . . . .	25
5.1 Nos objectifs . . . . .	25
5.2 La structure de l'ouvrage . . . . .	26
5.2.1 Les deux parcours complémentaires . . . . .	27
5.2.2 Ressources complémentaires à utiliser . . . . .	29
6. Petit bestiaire des data sciences . . . . .	30
6.1 Les fondamentaux . . . . .	30
6.1.1 Apprentissage et classification . . . . .	30
6.1.2 Petit vocabulaire graphique du machine learning . . . . .	31
6.1.3 Régression . . . . .	43
6.1.4 Régression linéaire généralisée . . . . .	45
6.1.5 Arbres de décision, prune, élagage . . . . .	45
6.1.6 Clustering, k-means . . . . .	47
6.1.7 k-NN . . . . .	47
6.1.8 Modèles paramétriques . . . . .	48
6.1.9 Lazy algorithm (algorithme paresseux) . . . . .	48
6.1.10 Overfitting : sur-détermination, sur-apprentissage . . . . .	49
6.1.11 Validation croisée, régularisation, bagging . . . . .	49
6.1.12 Optimisation, méthode du gradient . . . . .	50

# 2 \_\_\_\_\_ Data Scientist et langage R

Bases de l'intelligence artificielle dans l'univers de la data

6.1.13	Algorithme glouton (greedy algorithm) . . . . .	52
6.1.14	Programmation linéaire, simplexe, point intérieur . . . . .	53
6.1.15	Estimation via la méthode de Monte-Carlo . . . . .	54
6.1.16	Entropie, indépendance et information mutuelle . . . . .	55
6.1.17	Discrétisation . . . . .	57
6.2	Méthodes "ensemble" . . . . .	58
6.2.1	Random forest . . . . .	58
6.2.2	AdaBoost (adaptative boosting) . . . . .	59
6.3	Lois de probabilité et distribution . . . . .	60
6.3.1	Généralités . . . . .	60
6.3.2	Petit bestiaire des lois de probabilité . . . . .	62
6.4	Les graphes. . . . .	64
6.4.1	Vocabulaire de base . . . . .	65
6.4.2	Conversion d'un tableau d'observations en graphe, similarité . . . . .	67
7.	Informatique professionnelle et data sciences . . . . .	67
7.1	La technologie . . . . .	67
7.2	Business Intelligence versus Big Data . . . . .	70
7.2.1	Différences en termes d'architecture . . . . .	70
7.2.2	Différences en termes d'usage . . . . .	71
7.2.3	Synthèse . . . . .	72
8.	Notations . . . . .	73
8.1	Notations des paramètres. . . . .	73
8.2	Autres notations . . . . .	74
8.2.1	Fonctions et applications ... $f(x)$ , $d(x,y)$ . . . . .	74
8.2.2	Quelques confusions possibles . . . . .	75
9.	À vous de jouer ! . . . . .	76

## Chapitre 2

### Premiers pas avec R

1.	Installation des composants . . . . .	77
1.1	Installation et lancement de R. . . . .	77
1.2	Installation et lancement de RStudio . . . . .	78
1.3	Installation de nouveaux packages . . . . .	79
1.4	Installation de packages : compléments . . . . .	81



- 2. Prise en main de R . . . . . 83
  - 2.1 R, une calculatrice efficace . . . . . 83
  - 2.2 R, un langage vectorisé . . . . . 85
  - 2.3 Fonctions agissant sur des vectors . . . . . 87
    - 2.3.1 Une première analyse rapide des données . . . . . 88
    - 2.3.2 Quelques statistiques simples sur les vectors . . . . . 89
    - 2.3.3 Trier un vector . . . . . 90
    - 2.3.4 Diverses fonctions avec somme, produit, min et max . . . . . 91
  - 2.4 Types de données simples . . . . . 92
    - 2.4.1 Les booléens . . . . . 92
    - 2.4.2 Les ensembles . . . . . 95
    - 2.4.3 Les listes . . . . . 96
    - 2.4.4 Les facteurs . . . . . 98
    - 2.4.5 Les tableaux . . . . . 105
  - 2.5 Les fonctions . . . . . 114
    - 2.5.1 Création et utilisation d'une fonction simple . . . . . 114
    - 2.5.2 Création d'un opérateur
      - à partir d'une fonction de deux variables . . . . . 115
    - 2.5.3 Usage des fonctions et portée des variables . . . . . 116
    - 2.5.4 Application des fonctions sur les matrices : apply . . . . . 119
    - 2.5.5 Compléments utiles . . . . . 120
  - 2.6 Structures de contrôle . . . . . 125
    - 2.6.1 Des instructions communes avec d'autres langages . . . . . 125
    - 2.6.2 Parcourir une matrice via des boucles for . . . . . 127
  - 2.7 Les chaînes de caractères . . . . . 130
  - 2.8 Le formatage des nombres . . . . . 133
  - 2.9 Les dates et les temps . . . . . 134
  - 2.10 Mesure de la durée d'un algorithme . . . . . 134
  - 2.11 Les nombres complexes . . . . . 140
    - 2.11.1 Manipulations de base des nombres complexes . . . . . 140
    - 2.11.2 Visualisation de nombres complexes . . . . . 140
  - 2.12 Programmation orientée objet . . . . . 142
    - 2.12.1 Classes et objets, en bref . . . . . 143
    - 2.12.2 Constructeurs . . . . . 148
    - 2.12.3 Héritage . . . . . 150
    - 2.12.4 Objets mutables . . . . . 153
    - 2.12.5 Gestion de pile : implémentation orientée objet avec RC . . . 154

# 4 \_\_\_\_\_ Data Scientist et langage R

Bases de l'intelligence artificielle dans l'univers de la data

3.	Manipulation des données	159
3.1	Lecture des données : les bases	159
3.2	Manipulation des colonnes d'un data.frame	162
3.3	Calculs simples sur un data.frame	163
3.3.1	Calculs sur les colonnes et les lignes	163
3.3.2	Manipulation des lignes	164
3.3.3	Application : comparaison effectifs de classes et Khi-2	167
3.3.4	Création de colonnes calculées	169
3.3.5	Tri d'un data.frame via order()	173
3.4	Analyse visuelle des données	175
3.4.1	Visualisation simple des données	175
3.4.2	Visualisation des variables numériques 2 à 2 avec mention des classes	177
3.4.3	Corrélations entre variables numériques	180
3.4.4	Séparation par classe, ggplot2, qplot	181
3.4.5	Visualisation 3D, relation entre trois variables numériques	182
3.4.6	Graphiques par paires	183
3.4.7	Boîtes à moustaches et tentative d'élimination des outliers	184
3.4.8	Création d'un modèle par arbre de décision	187

## Chapitre 3

### Maîtriser les bases

1.	Se mettre en harmonie avec les données	191
1.1	Quelques notions fondatrices	192
1.1.1	Phénomène aléatoire	192
1.1.2	Probabilité, variable aléatoire et distribution	193
1.1.3	Un peu de mathématiques : notations et définitions utiles	196
1.1.4	Moments d'une variable aléatoire discrète X	202
1.1.5	Premières considérations sur les erreurs et estimations	209
1.2	Se familiariser avec ses données	210
1.2.1	R Commander	211
1.2.2	Rattle	213

- 2. Matrices et vecteurs . . . . . 218
  - 2.1 Conventions, notations, utilisations basiques . . . . . 218
  - 2.2 Matrices, vecteurs : vers une introduction  
à la notion d'apprentissage supervisé . . . . . 228
  - 2.3 Plus loin dans la manipulation des matrices avec R . . . . . 230
    - 2.3.1 Opérations basiques . . . . . 230
    - 2.3.2 Quelques savoir-faire utiles sur les matrices de R . . . . . 231
    - 2.3.3 Normes de vecteurs et normes de matrices . . . . . 235
    - 2.3.4 Matrices et vecteurs : diverses syntaxes utiles. . . . . 238
- 3. Estimations . . . . . 243
  - 3.1 Positionnement du problème d'estimation . . . . . 243
    - 3.1.1 Formulation générale du problème. . . . . 243
    - 3.1.2 Application et reformulation du problème d'estimation . . . . . 245
  - 3.2 Les indicateurs d'écart utilisés en machine learning . . . . . 253
    - 3.2.1 MSE, RMSE, SSE, SST. . . . . 253
    - 3.2.2 MAE, ME . . . . . 255
    - 3.2.3 NRMSE/NRMSD, CV\_RMSE . . . . . 256
    - 3.2.4 SDR. . . . . 257
    - 3.2.5 Accuracy, R2 . . . . . 257
- 4. Mise en pratique : apprentissage supervisé . . . . . 260
  - 4.1 Préparation . . . . . 260
  - 4.2 Tester des hypothèses, p\_value . . . . . 262
    - 4.2.1 Analyse graphique interactive avec iplots . . . . . 262
    - 4.2.2 Test de Breusch-Pagan et zoom sur p\_value . . . . . 263
  - 4.3 Création d'un modèle (régression linéaire multiple) . . . . . 264
  - 4.4 Établissement d'une prédiction . . . . . 265
  - 4.5 Étude des résultats et représentations graphiques . . . . . 266
  - 4.6 Indicateurs courants - calculs. . . . . 269
  - 4.7 Étude du modèle linéaire généré . . . . . 270
  - 4.8 Conclusion sur le modèle linéaire . . . . . 276
  - 4.9 Utilisation d'un modèle "Random Forest". . . . . 277

# 6 \_\_\_\_\_ Data Scientist et langage R

Bases de l'intelligence artificielle dans l'univers de la data

## Chapitre 4

### Techniques et algorithmes incontournables

1. Constituer sa boîte à outils . . . . .	281
2. Représentation graphique des données . . . . .	282
2.1 Un graphique "simple" . . . . .	282
2.2 Histogrammes évolués . . . . .	285
2.2.1 Distribution multiclasse . . . . .	285
2.2.2 Mixture de plusieurs distributions par classe. . . . .	286
2.2.3 Visualisation de la densité d'une distribution . . . . .	287
2.2.4 Une autre mixture par classe. . . . .	288
2.2.5 Une variable, mais un histogramme pour chaque classe. . . . .	289
2.2.6 Graphique avec une densité par classe . . . . .	291
2.3 Diagramme par paires et à facettes . . . . .	293
2.3.1 Diagramme par paires, version simple . . . . .	293
2.3.2 Classes en configuration XOR . . . . .	294
2.3.3 Diagramme par paires avec "factors". . . . .	296
2.3.4 Facettes et échelle logarithmique . . . . .	298
3. Machine learning : pratiques courantes . . . . .	300
3.1 Parcours théorique accéléré . . . . .	300
3.1.1 Linéarité . . . . .	300
3.1.2 Erreurs in et out, notion de VC dimension . . . . .	302
3.1.3 Hyperplans, séparabilité avec marges . . . . .	305
3.1.4 Kernel Trick, noyaux, transformations, feature space . . . . .	307
3.1.5 Problèmes de la régression : introduction à la régularisation . . . . .	310
3.2 Pratique par pratique . . . . .	315
3.2.1 Cross validation : k-fold CV . . . . .	315
3.2.2 Naive Bayes . . . . .	316
3.2.3 C4.5 et C5.0 . . . . .	322
3.2.4 Support Vector Machines (SVM) . . . . .	324
3.2.5 Clusterisation, k-means. . . . .	328
4. Où en sommes-nous dans notre apprentissage ? . . . . .	337
4.1 Vos acquis opérationnels . . . . .	337
4.2 Les lacunes éventuelles à combler maintenant . . . . .	338

**Chapitre 5**  
**Cadre méthodologique du data scientist**

- 1. Le problème méthodologique au niveau du projet . . . . . 339
  - 1.1 L'expression du besoin . . . . . 339
  - 1.2 La gestion du projet . . . . . 340
- 2. Le cycle interne des data sciences . . . . . 343
  - 2.1 Revue de détail du problème posé . . . . . 343
  - 2.2 Travaux préliminaires sur les données . . . . . 344
    - 2.2.1 Exigences sur les données . . . . . 344
    - 2.2.2 Collecte, nettoyage et compréhension des données . . . . . 344
  - 2.3 Le cycle de modélisation . . . . . 346
    - 2.3.1 Feature engineering . . . . . 346
    - 2.3.2 Modélisation et évaluation . . . . . 347
    - 2.3.3 Choix du meilleur modèle . . . . . 348
    - 2.3.4 Test, interprétation et confrontation avec le métier . . . . . 348
  - 2.4 Préparation de l'industrialisation et déploiement . . . . . 350
  - 2.5 Préparation des itérations suivantes . . . . . 350
    - 2.5.1 Éléments à prendre en compte . . . . . 350
    - 2.5.2 Documentation gérée par les data scientists . . . . . 351
- 3. Compléments méthodologiques . . . . . 352
  - 3.1 Classifier vos objectifs . . . . . 352
  - 3.2 Trucs et astuces . . . . . 353

**Chapitre 6**  
**Traitement du langage naturel**

- 1. Positionnement du problème . . . . . 355
- 2. Analyse sémantique latente et SVD . . . . . 356
  - 2.1 Aspects théoriques . . . . . 356
    - 2.1.1 SVD : généralités . . . . . 357
    - 2.1.2 Une justification de la décomposition SVD . . . . . 357
    - 2.1.3 SVD dans le contexte LSA . . . . . 358
    - 2.1.4 Interprétation . . . . . 358
    - 2.1.5 Alternative non linéaire, Isomap  
(MDS, géodésique, variété, manifold) . . . . . 359

# 8 \_\_\_\_\_ Data Scientist et langage R

Bases de l'intelligence artificielle dans l'univers de la data

2.2	Mise en pratique . . . . .	360
2.2.1	Initialisation . . . . .	360
2.2.2	Au cœur de LSA . . . . .	362
2.2.3	Résultats . . . . .	363
2.2.4	Manipulations, interprétations récréatives et non fondées. . .	366

## Chapitre 7

### Graphes et réseaux

1.	Introduction . . . . .	369
2.	Premiers pas . . . . .	369
2.1	Quelques notions et notations complémentaires basiques . . . . .	370
2.2	Manipulations simples de graphes avec R . . . . .	371
2.3	Structure des graphes . . . . .	382
3.	Graphes et réseaux (sociaux) . . . . .	384
3.1	Analyse des réseaux sociaux : concepts de base . . . . .	385
3.2	Mise en pratique . . . . .	388
3.3	Détection de communautés . . . . .	390

## Chapitre 8

### Autres problèmes, autres solutions

1.	Séries temporelles . . . . .	399
1.1	Introduction . . . . .	399
1.2	Modèle stationnaire . . . . .	400
1.2.1	Processus stationnaire : les bases . . . . .	400
1.2.2	Processus autorégressif AR : aller plus loin . . . . .	402
1.2.3	Considérations (très) utiles . . . . .	404
1.3	Processus non stationnaires . . . . .	405
1.3.1	Le modèle ARIMA . . . . .	405
1.3.2	Processus saisonniers : SARIMA . . . . .	406
1.3.3	Modèles ARCH et GARCH . . . . .	407
1.3.4	Convolution et filtres linéaires . . . . .	409

- 1.4 Mise en pratique . . . . . 411
  - 1.4.1 Les bases de la manipulation des séries temporelles en R . . . . 412
  - 1.4.2 Étude de séries temporelles . . . . . 416
  - 1.4.3 Prédications sur ARIMA (AR MA SARIMA) . . . . . 421
- 1.5 Mini-bestiaire ARIMA . . . . . 424
- 2. Systèmes flous . . . . . 427
  - 2.1 Introduction . . . . . 427
  - 2.2 La logique floue en pratique (systèmes experts) . . . . . 435
- 3. Essaim (swarm) . . . . . 441
  - 3.1 Swarm et optimisation : l'algorithme PSO . . . . . 441
    - 3.1.1 Présentation de PSO . . . . . 442
    - 3.1.2 Description de PSO . . . . . 442
  - 3.2 Mise en pratique de PSO . . . . . 444

**Chapitre 9**  
**Feature Engineering**

- 1. Feature Engineering, les bases . . . . . 447
  - 1.1 Positionnement du problème . . . . . 447
  - 1.2 Ce sur quoi il faut être très attentif . . . . . 448
    - 1.2.1 La qualité de la distribution . . . . . 448
    - 1.2.2 La nature des features . . . . . 449
  - 1.3 Maîtriser la dimensionnalité . . . . . 451
  - 1.4 Une solution éprouvée : la PCA . . . . . 453
  - 1.5 Un exemple simple d'utilisation de la PCA . . . . . 454
  - 1.6 Les valeurs inconnues et les features mal conditionnées . . . . . 461
  - 1.7 Création de nouvelles features . . . . . 464
  - 1.8 En guise de conclusion . . . . . 466
- 2. PCA classique, éléments mathématiques . . . . . 467
- 3. Réduction des données (data reduction) . . . . . 469
- 4. Réduction de la dimensionnalité et entropie . . . . . 470
  - 4.1 Description théorique du problème . . . . . 471
  - 4.2 Implémentation en R et discussion . . . . . 472

# 10 \_\_\_\_\_ Data Scientist et langage R

Bases de l'intelligence artificielle dans l'univers de la data

## Chapitre 10

### Compléments utiles

1. GAM : généralisation de LM/GLM . . . . .	481
2. Manipulation d'images . . . . .	483
2.1 Création, visualisation, lecture et écriture d'images . . . . .	484
2.2 Transformation des images . . . . .	489
2.2.1 Exemples de manipulation de la couleur et des intensités . . . . .	489
2.2.2 Exemples de manipulation de la géométrie de l'image . . . . .	492
2.2.3 Application de filtres sur les images . . . . .	495
3. Comment créer un échantillon : LHS (hypercube latin) . . . . .	496
4. Travailler sur des données spatiales . . . . .	499
4.1 Variogramme . . . . .	499
4.1.1 Champ et variable régionalisée . . . . .	499
4.1.2 Détermination du variogramme . . . . .	504
4.2 Krigeage (kriging) . . . . .	506
4.2.1 La théorie, en bref . . . . .	506
4.2.2 Implémentation en R . . . . .	507
5. Savoir-faire utiles . . . . .	509
5.1 Tracer une courbe ROC . . . . .	509
5.2 Un réseau neuronal (premiers pas vers le deep learning) . . . . .	511
6. Gradient Boosting et Generalized Boosted Regression . . . . .	516
6.1 Les grands principes . . . . .	516
6.2 Les paramètres et les usages (package GBM) . . . . .	517
6.2.1 Covariance . . . . .	517
6.2.2 Loss . . . . .	517
6.2.3 Optimisation de l'algorithme . . . . .	518
6.3 Mise en pratique . . . . .	518

## Chapitre 11

### Full Stack R

1. Pourquoi ce chapitre ? . . . . .	521
2. Programmation fonctionnelle et/ou défensive . . . . .	521
3. Persistance, bases de données et R . . . . .	538



- 4. Parallélisme . . . . . 548
- 5. Collecter des données externes . . . . . 549
- 6. Créer une API avec R . . . . . 551

**Chapitre 12**  
**Partager ses analyses**

- 1. Rédiger en Markdown . . . . . 553
  - 1.1 Introduction . . . . . 553
  - 1.2 Syntaxe Markdown . . . . . 555
- 2. Création d'un fichier R Markdown . . . . . 564
  - 2.1 En-tête R Markdown . . . . . 566
  - 2.2 Mise en forme du code . . . . . 570
  - 2.3 Insertion et exécution de code R . . . . . 571
  - 2.4 Exécuter du code R . . . . . 576
  - 2.5 Générer le rapport . . . . . 577
  - 2.6 Publier son rapport avec RPubS . . . . . 578
- 3. Créer sa première application Shiny . . . . . 579
  - 3.1 Introduction . . . . . 579
  - 3.2 Installation dans RStudio et préambule . . . . . 580
  - 3.3 Organisation des fichiers de l'application . . . . . 581
  - 3.4 Syntaxe des inputs et outputs . . . . . 584
  - 3.5 Mise en page et organisation visuelle du contenu . . . . . 591
  - 3.6 Modifier le thème de l'application . . . . . 602
  - 3.7 Générer l'application . . . . . 604
  - 3.8 Déployer l'application sur le Web . . . . . 604
  - 3.9 Shiny Dashboard . . . . . 606

**Chapitre 13**  
**Cartographie**

- 1. Pourquoi étudier les représentations cartographiques ? . . . . . 615
- 2. Accéder à l'information géographique . . . . . 616
- 3. Création de cartes statiques avec R . . . . . 620
- 4. Création de cartes dynamiques avec R . . . . . 627

# 12 \_\_\_\_\_ Data Scientist et langage R

Bases de l'intelligence artificielle dans l'univers de la data

## Chapitre 14

### Autour des données

1. Présentation du chapitre . . . . .	629
2. Appréhension directe et rapide d'un dataset . . . . .	630
3. Analyse de la conformation des distributions par rapport à la loi normale. . . . .	635
4. Dépendance linéaire entre variables . . . . .	636
5. Mise en évidence de différences sur les distributions . . . . .	639
6. Points étrangers . . . . .	642
7. Tris et agrégats. . . . .	643
7.1 Tris automatisés . . . . .	643
7.2 Calculs sur les agrégats . . . . .	648
7.3 Extraction, pivot et tri rapide de données d'un dataframe. . . . .	649
7.4 Utilisation des jointures . . . . .	650
7.4.1 Enrichissement d'un dataframe par une jointure . . . . .	650
7.4.2 Jointures pour décodage ou transformation - ex. : factor to integer . . . . .	652
7.5 Applications de fonctions variables sur les données . . . . .	653
7.6 Compléments en matière de métaprogrammation. . . . .	655
7.6.1 Quoting appliqué aux graphiques. . . . .	655
7.6.2 Définition d'une macro . . . . .	658

## Chapitre 15

### Analyse numérique et maths opérationnelles

1. Calculs numériques dans l'esprit Matlab . . . . .	661
1.1 Considérations pratiques . . . . .	661
1.2 Un aperçu de la précision de notre machine . . . . .	662
1.3 Un package possédant les mêmes instructions que Matlab. . . . .	664
1.4 Quelques compléments sur les complexes . . . . .	665
1.5 Quelques compléments sur les matrices . . . . .	666
1.5.1 Création de matrices basiques. . . . .	666
1.5.2 Création de matrices de nombres aléatoires. . . . .	668
1.5.3 Calculs sur les matrices d'entiers . . . . .	669

2.	Un peu d'algèbre linéaire. . . . .	670
2.1	Exemple de création d'une base orthonormale. . . . .	670
2.2	Projection sur un sous-espace vectoriel. . . . .	671
3.	Fonctions et systèmes d'équations, 1 à n variables. . . . .	673
3.1	Fonctions d'une variable . . . . .	673
3.1.1	Exploration de divers tracés . . . . .	673
3.1.2	Autour des racines d'une fonction . . . . .	676
3.2	Fonctions de plusieurs variables . . . . .	678
3.2.1	Représentation d'une fonction réelle de deux variables réelles . . . . .	678
3.2.2	Recherche de minima d'une fonction en deux dimensions . .	679
3.3	Systèmes d'équations non différentielles, linéaires ou non . . . . .	681
3.3.1	Système d'équations linéaires . . . . .	681
3.3.2	Système d'équations non linéaires non différentielles, $n = m$ . . . . .	684
4.	Dérivation de fonctions. . . . .	685
4.1	Dérivée symbolique et numérique avec R-base . . . . .	685
4.2	Dérivées en utilisant des packages spécifiques . . . . .	687
4.3	Dérivée de fonctions spéciales . . . . .	687
5.	Autour de l'intégration . . . . .	689
5.1	Calcul d'une intégrale multiple . . . . .	689
5.2	Autres intégrales, dont intégrale impropre . . . . .	690
6.	Fonctions spéciales et équations différentielles. . . . .	691
6.1	Autres fonctions spéciales . . . . .	691
6.1.1	Fonctions d'Airy et de Bessel . . . . .	691
6.1.2	La fonction d'erreur erf . . . . .	693
6.2	Équation différentielle ordinaire normalisée d'ordre 1 . . . . .	693
6.3	Équation différentielle du second degré (problème aux limites) . . .	694
6.3.1	EQD linéaire (2nd degré) . . . . .	694
6.3.2	EQD non linéaire (2nd degré) . . . . .	696
7.	Éléments pratiques de calcul différentiel . . . . .	699
7.1	Fonctions réelles d'un réel . . . . .	699
7.2	Fonctions réelles d'un vecteur . . . . .	701
7.2.1	Calcul des dérivées partielles. . . . .	701
7.2.2	Calculs d'opérateurs utilisant les dérivées partielles . . . . .	703
7.3	Fonctions vectorielles d'un vecteur . . . . .	704

# 14 \_\_\_\_\_ Data Scientist et langage R

Bases de l'intelligence artificielle dans l'univers de la data

8. Calcul symbolique avec SymPy . . . . .	706
8.1 Documentation à consulter . . . . .	707
8.2 Accéder aux résultats dans Rmarkdown . . . . .	707

## Chapitre 16 R et Python

1. Une polémique stérile, mais une vraie question : R ou Python ? . . . . .	709
2. Exemples de codes R et Python comparables. . . . .	712
2.1 Codes triviaux . . . . .	713
2.2 Packages . . . . .	713
2.3 Fonctions simples . . . . .	714
2.4 Structures de données. . . . .	715
2.5 Booléens . . . . .	717
2.6 Programmation de base . . . . .	718
2.7 Création de fonctions . . . . .	718
2.8 Blocs et indentation . . . . .	719
2.9 Datasets . . . . .	720
3. Accéder à Python à partir de R . . . . .	722
3.1 Installation et initialisation du contexte technique . . . . .	722
3.2 Interactions R et Python, du point de vue R. . . . .	723
3.2.1 Partage de variables et exécution de code Python dans le biotope R. . . . .	723
3.2.2 Bibliothèques Python dans du code R, exemples NLP . . . . .	724
4. Considérations sur les cas difficiles. . . . .	730

## Chapitre 17 Deep learning avec TensorFlow et Keras

1. Deep learning selon Google . . . . .	733
2. Installation et initialisation de votre contexte technique. . . . .	735
3. Tenseurs TensorFlow/Keras . . . . .	738
3.1 Forme des tenseurs . . . . .	738
3.2 Manipulation des tenseurs . . . . .	740
3.3 Autre usage du framework . . . . .	751

- 4. Mise au point d'un modèle de référence en utilisant caret . . . . . 752
- 5. Création d'un modèle avec TensorFlow 2 et Keras . . . . . 757
  - 5.1 Initialisation et préparation des données . . . . . 758
  - 5.2 Étape 1 : décrire la structure du réseau . . . . . 760
  - 5.3 Étape 2 : compiler le modèle . . . . . 763
  - 5.4 Étape 3 : training . . . . . 763

**Chapitre 18**  
**BigML, le machine learning pour tous**

- 1. Introduction . . . . . 769
- 2. Pour qui ? . . . . . 770
- 3. Présentation de la démarche . . . . . 772
- 4. Manipulation des sources de données . . . . . 773
- 5. Création de projets . . . . . 777
- 6. Manipulation de datasets . . . . . 778
- 7. Division du jeu de données . . . . . 779
- 8. Création d'un modèle de machine learning . . . . . 781
  - 8.1 Paramétrages des modèles . . . . . 781
  - 8.2 Visualisation du modèle . . . . . 782
  - 8.3 Filtrage et élagage . . . . . 783
  - 8.4 Autres visualisations . . . . . 784
- 9. Évaluation du modèle . . . . . 785
- 10. Partage de ses modèles . . . . . 787
- 11. Réalisation d'une première prédiction . . . . . 787
- 12. Utilisation du modèle dans R . . . . . 790

# 16 \_\_\_\_\_ Data Scientist et langage R

Bases de l'intelligence artificielle dans l'univers de la data

## Chapitre 19

### Déploiement Shiny sur votre propre Cloud

1. Introduction . . . . .	793
2. Location d'un serveur . . . . .	793
3. Création d'une machine virtuelle . . . . .	794
4. Se connecter à distance à votre machine . . . . .	795
4.1 Installation du client SSH . . . . .	795
4.1.1 OpenSSH . . . . .	795
4.1.2 GitBash . . . . .	796
4.2 Génération des clés publique et privée . . . . .	797
4.3 Gestion des paquets . . . . .	798
5. Ajout des composants logiciels de base . . . . .	799
6. Installation de R et de Shiny. . . . .	800
7. Paramétrage réseau et Linux . . . . .	802
7.1 Paramétrer le VCN (Virtual Cloud Network). . . . .	802
7.2 Gestion du firewall . . . . .	803
7.3 Se doter d'une application de monitoring : htop . . . . .	806
7.4 Se doter d'un serveur HTTP (Apache). . . . .	806
7.5 Installation de Webmin . . . . .	807

## Annexes

1. De l'utilité de ces annexes . . . . .	811
2. Formulas . . . . .	811
3. Stratégies suivant la nature des données . . . . .	813
3.1 Comptages . . . . .	813
3.2 Proportions . . . . .	813
3.3 Variable de réponse binaire . . . . .	813
3.4 Données induisant un modèle mixte (mixed effect) . . . . .	814
3.5 Graphes . . . . .	815
3.6 Analyse de survie (survival analysis) . . . . .	815
4. Filtres (sur images) . . . . .	816
5. Distances . . . . .	818

6. Astuces et petits conseils .....	821
6.1 Sur les tests .....	821
6.2 Gestion des variables .....	821
6.3 Analyse et manipulation de résultats .....	822
6.3.1 Résidus .....	822
6.3.2 Manipulation des modèles .....	822
7. Packages et thèmes à étudier .....	823
7.1 Créer son propre package .....	823
7.2 Règles d'association .....	824
7.3 Exporter un modèle .....	824
7.4 Tenseurs .....	824
7.5 SVM pour la détection de nouveautés (novelty detection) .....	824
8. Vocabulaire et "tricks of the trade" .....	825
8.1 Compléments sur les bases du machine learning .....	825
8.2 Compléments sur les aspects bayésiens .....	826
8.3 Vocabulaire (dont anglais) des modèles gaussiens .....	828
9. Algorithmes à étudier .....	828
10. Quelques formulations d'algèbre linéaire .....	829
11. Bibliographie .....	830
Conclusion .....	833
Index .....	835