

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

Chapitre 1

Avant-propos et introduction

1. Avant-propos	11
2. Python et l'analyse de données	12
2.1 L'explosion des données	12
2.2 L'analyse de données	13
2.3 R et Python pour l'analyse de données	14
3. Connaître les sources de données libres	15
3.1 Kaggle	15
3.2 Les données gouvernementales	17
4. Déroulement du livre	18

Chapitre 2

Mise en place de l'environnement de travail

1. Introduction : pourquoi utiliser Python pour la Data Science ?	19
2. Introduction à IPython et Jupyter	20
2.1 Introduction à IPython	20
2.2 Le projet Jupyter	22
3. Qu'est-ce qu'Anaconda ?	23
4. Installation d'Anaconda	24
4.1 Installation sur Windows	24
4.2 Installation sur MacOS	26
4.3 Installation sur Linux	26

2 — Python pour la Data Science

Analysez vos données par la pratique

5.	Découverte d'Anaconda Navigator	28
5.1	Applications disponibles dans Anaconda Navigator	28
5.2	Gestion des packages et environnements	29
6.	Prise en main de Jupyter Notebook	32
6.1	Tableau de bord de Jupyter Notebook	32
6.2	Premiers pas avec les notebooks	34
6.3	Comprendre l'interface des notebooks	38
6.3.1	La barre de menus	38
6.3.2	La barre d'outils	40
6.3.3	Les cellules	41
6.3.4	Les modes Commande et Edition et les raccourcis-clavier	42
6.3.5	Les bases du langage Markdown pour écrire dans un notebook	45
6.3.6	Partager son notebook	51
7.	Les packages Python essentiels pour la Data Science	52
7.1	NumPy	52
7.2	Pandas	53
7.3	Matplotlib	53
7.4	Seaborn	53

Chapitre 3

Rappels sur le langage Python

1.	Introduction sur le langage de programmation Python	55
2.	Les variables	56
3.	Les différents types de données (int, float, bool, str)	58
3.1	Les nombres réels et entiers	58
3.2	Les booléens	59
3.3	Les chaînes de caractères	60

4. Les structures de données basiques (listes, tuples et dictionnaires) .	62
4.1 Les listes	62
4.1.1 Créer une liste	62
4.1.2 Accéder aux éléments d'une liste	63
4.1.3 Ajouter et supprimer des éléments à une liste.	66
4.2 Les tuples.	70
4.3 Les dictionnaires	72
4.3.1 Introduction aux dictionnaires.	72
4.3.2 Ajouter, modifier et supprimer des éléments d'un dictionnaire	74
4.3.3 Parcourir un dictionnaire	76
5. Les opérateurs arithmétiques, relationnels et logiques	78
5.1 Les opérateurs arithmétiques.	78
5.2 Les opérateurs relationnels et logiques	80
5.2.1 Les opérateurs relationnels	80
5.2.2 Les opérateurs logiques	81
6. Vocabulaire en Python : fonctions, méthodes, attributs, modules et librairies (packages)	82
6.1 Fonctions.	82
6.2 Méthodes.	84
6.3 Attributs	85
6.4 Modules.	85
6.5 Librairies (packages)	87
7. Instructions de condition if et boucles for	88
7.1 Instruction de condition if.	88
7.2 Boucle for	90

4 Python pour la Data Science

Analysez vos données par la pratique

Chapitre 4 Maîtriser la librairie NumPy

1. Introduction à NumPy	93
2. Les tableaux NumPy	94
2.1 Créer un ndarray	94
2.1.1 Créer un ndarray à partir de listes	94
2.1.2 Créer un ndarray grâce à des fonctions NumPy	97
2.1.3 Créer un ndarray à partir d'un fichier	100
2.2 Indexation	103
2.2.1 Indexation simple	103
2.2.2 Indexation booléenne	105
2.2.3 Fancy indexing	108
2.3 Accéder aux éléments par tranche (slicing)	110
2.3.1 Slicing sur un tableau NumPy à 1 dimension	110
2.3.2 Slicing sur un tableau NumPy à 2 dimensions	112
2.4 Notion de vue et copie	114
3. Les opérations mathématiques avec NumPy	116
3.1 Les opérations arithmétiques	116
3.2 Les fonctions d'agrégations	119
4. Inspecter un tableau grâce aux attributs de NumPy	122
5. Manipuler des tableaux NumPy	124
5.1 Ajouter et supprimer des éléments dans un tableau	124
5.1.1 Ajouter des éléments dans un tableau	124
5.1.2 Supprimer des éléments d'un tableau	127
5.2 Diviser un tableau NumPy (split, hsplit et vsplit)	128
5.2.1 Sur un tableau à une dimension	129
5.2.2 Sur un tableau à deux dimensions	130
5.3 Concaténer/combiner des tableaux	131
5.3.1 La fonction concatenate()	131
5.3.2 Les fonctions vstack() et hstack()	133
6. Introduction aux matrices avec NumPy	135

Chapitre 5
Maîtriser la librairie Pandas

- 1. Introduction 137
 - 1.1 Introduction à la librairie Pandas. 137
 - 1.2 Introduction au jeu de données utilisé pour les exemples ... 140
- 2. Lire et écrire des fichiers avec Pandas 143
 - 2.1 Lecture de fichiers texte (CSV ou TXT) 143
 - 2.1.1 Lecture basique d'un fichier 143
 - 2.1.2 Gestion de l'en-tête 146
 - 2.1.3 Gestion des index. 148
 - 2.1.4 Création d'un tableau à une dimension
à partir du fichier. 150
 - 2.1.5 Filtrage des colonnes lors de la lecture du fichier 150
 - 2.1.6 Les types des différentes colonnes 152
 - 2.1.7 Gestion des dates lors de la lecture du fichier 152
 - 2.2 Lecture de fichiers Excel. 153
 - 2.3 Importation des données à partir d'une base de données. 156
 - 2.4 Lecture de fichiers au format JSON. 158
 - 2.5 Écriture de fichiers ou exportation de données. 160
- 3. Structure de données Pandas : les Series (Séries) 162
 - 3.1 Introduction 162
 - 3.2 Créer des séries 163
 - 3.2.1 À partir de valeurs aléatoires. 163
 - 3.2.2 À partir d'une liste Python 165
 - 3.2.3 À partir d'un tableau NumPy (ndarray) 168
 - 3.2.4 À partir d'un fichier texte 168
 - 3.3 Choisir l'index d'une série. 169
 - 3.4 Accéder aux valeurs d'une série 171
 - 3.4.1 Indexing via la position des valeurs 171
 - 3.4.2 Indexing via l'étiquette des valeurs 172
 - 3.4.3 Les indexeurs loc et iloc. 173
 - 3.4.4 Indexing via une expression booléenne 175

6 — Python pour la Data Science

Analysez vos données par la pratique

3.4.5	Slicing : découpage de valeurs successives	178
3.5	Les attributs et les méthodes des objets de classe Series	184
3.5.1	Les attributs des objets de classe Series	184
3.5.2	Les méthodes des objets de classe Series	185
3.6	Ajouter, supprimer et modifier les valeurs d'une série	187
3.6.1	Ajouter des valeurs à une série	187
3.6.2	Supprimer une valeur d'une série	188
3.6.3	Modifier les valeurs d'une série	189
4.	Structure de données Pandas : les objets de type DataFrame	191
4.1	Introduction	191
4.2	Indexing : sélectionner des valeurs d'un dataframe	193
4.2.1	Indexing et slicing avec l'attribut loc	194
4.2.2	Indexing et slicing avec l'attribut iloc	198
4.2.3	Indexing avec une expression booléenne	199
4.3	Ajout, suppression et modification sur un dataframe	201
4.3.1	Ajouter une ou plusieurs colonnes à un dataframe	201
4.3.2	Ajouter une ligne à un dataframe	202
4.3.3	Supprimer des lignes ou colonnes d'un dataframe	205
4.3.4	Modifier des valeurs dans un dataframe	207
4.4	Nettoyage et préparation des données avec Pandas	209
4.4.1	Gestion des données manquantes	210
4.4.2	Gestion des données dupliquées	215
4.5	Exploration préliminaire d'un dataframe	219
4.5.1	Principaux attributs	219
4.5.2	Définition des termes variable, variable quantitative et variable qualitative et découverte de la méthode describe()	222
4.5.3	Méthodes de tri d'un dataframe	225
5.	Structure de données Pandas : les panels	227
6.	Manipulation avancée des données avec Pandas	228
6.1	Les opérations groupby	228
6.1.1	groupby sur une colonne	228
6.1.2	groupby sur plusieurs colonnes	231

6.1.3 Appliquer plusieurs fonctions avec la méthode groupby et la méthode aggregate	232
6.2 Appliquer une fonction à un dataframe avec la méthode apply	234
6.3 Remodeler/réorganiser des dataframes	236
6.3.1 Pivotage : la méthode pivot_table	236
6.3.2 Les méthodes stack (empiler) et unstack (désempiler) .	238

Chapitre 6

Maîtriser la librairie Matplotlib

1. Introduction	241
2. Le fonctionnement de Matplotlib	242
2.1 Architecture de Matplotlib	242
2.2 Organisation des figures avec Matplotlib	244
3. La création d'un premier graphique simple	246
3.1 Préparer son jeu de données	246
3.2 Créer un nuage de points	250
3.3 Ajouter un titre principal et des labels aux axes du nuage de points	254
3.4 Enregistrer son graphique	256
3.5 Changer la taille de la fenêtre graphique et la résolution de son graphique	258
3.6 Tracer plusieurs courbes sur un même graphique (sur un même objet axes)	260
3.7 Ajouter une légende à son graphique	263
3.8 Annoter son graphique avec du texte	265
3.9 Combiner plusieurs graphiques grâce à subplot et subplots .	268
3.9.1 Tracer des sous-graphiques (subplot) sur une ligne ou une colonne	268
3.9.2 Tracer des sous-graphiques sur plusieurs lignes et plusieurs colonnes	271
3.9.3 Incruster un objet axes dans un autre	274

8 _____ Python pour la Data Science

Analysez vos données par la pratique

4. Les différents types de graphes	278
4.1 Types de graphiques selon les types de variables (quantitatives et qualitatives)	278
4.2 Scatterplot.	279
4.3 Graphique à barres (bargraph).	282
4.3.1 Graphique à barres simple.	282
4.3.2 Graphique à barres groupées.	287
4.3.3 Graphique à barres empilées	294
4.4 Boxplots	296

Chapitre 7

Maîtriser la librairie Seaborn

1. Introduction	301
2. L'esthétique des figures avec Seaborn (Aesthetic)	303
2.1 Paramétrer les styles Seaborn (thèmes).	304
2.2 Supprimer les axes	306
2.3 Paramétrer les contextes avec Seaborn	309
2.4 Les palettes de couleur avec Seaborn.	314
2.4.1 Choisir une palette de couleurs existante	314
2.4.2 Créer sa propre palette de couleurs	317
3. Les différents types de graphiques.	318
3.1 Préparation du jeu de données.	318
3.2 Nuage de points (scatterplot)	318
3.3 Graphiques de régression	323
3.4 Pointplot	327
3.5 Nuage de points avec une variable qualitative : stripplot	331
3.6 Boxplots	334
3.7 Graphique à barres : countplot	336
3.8 Histogrammes.	339
3.9 Jointplot	344
3.10 Pairplot	346
3.11 Heatmap	349

- 4. Les graphiques multi-grilles 352
 - 4.1 FacetGrid 352
 - 4.2 PairGrid 354
 - 4.3 JointGrid 359
- 5. Conclusion 361

Chapitre 8

Exercice complet sur jeu de données réel

- 1. Introduction 363
- 2. Présentation du jeu de données 365
- 3. Énoncé de l'exercice 366
 - 3.1 Lire le fichier 366
 - 3.2 Afficher les dimensions du dataframe 367
 - 3.3 Compter les films et les séries 367
 - 3.4 Générer le résumé statistique du dataframe 367
 - 3.5 Compter les valeurs manquantes 368
 - 3.6 Explorer les valeurs manquantes 368
 - 3.6.1 Sur la colonne des directeurs de production 368
 - 3.6.2 Sur la colonne des acteurs 368
 - 3.7 Supprimer les lignes dupliquées 369
 - 3.8 Compter les films/séries produits
par les États-Unis et par la France 369
 - 3.9 Afficher le contenu le plus vieux disponible sur Netflix 370
 - 3.10 Afficher le film avec la durée la plus longue sur Netflix 370
 - 3.10.1 Nouvelle notion : les méthodes str 370
 - 3.10.2 Énoncé 371
 - 3.11 Étudier les catégories avec le plus de contenu 372
 - 3.12 Afficher les directeurs qui ont produit
le plus de films/séries disponibles sur Netflix 375
 - 3.13 Voir si Jan Suter travaille souvent avec les mêmes acteurs ... 375

10 _____ Python pour la Data Science

Analysez vos données par la pratique

3.14 Représenter les dix pays qui ont produit le plus de contenus disponibles sur Netflix, avec le nombre de contenus par pays	376
3.15 Tracer un graphe à barres du nombre de films/séries par classement de contenu (rating)	377
3.16 Afficher l'évolution du nombre de films/séries disponibles sur Netflix au cours du temps	377
3.16.1 Notions supplémentaires sur les dates	377
3.16.2 Énoncé	377
3.17 Afficher la distribution de la durée des films disponibles sur Netflix	378
3.18 Tracer un graphique représentant le nombre de séries par modalité de nombre de saisons	379
Index	381