

Les éléments à télécharger sont disponibles à l'adresse suivante :

**<http://www.editions-eni.fr>**

Saisissez la référence de l'ouvrage **EIPYTDATA** dans la zone de recherche et validez. Cliquez sur le titre du livre puis sur le bouton de téléchargement.

## Avant-propos

### Chapitre 1 Introduction

1.	Des données partout . . . . .	15
1.1	Provenance des données . . . . .	16
1.1.1	Le Web . . . . .	16
1.1.2	Les données privées . . . . .	17
1.1.3	Créons nos propres données . . . . .	18
1.2	Forme des données . . . . .	19
1.3	Volumétrie . . . . .	20
2.	La data science . . . . .	21
2.1	Feature engineering . . . . .	21
2.1.1	La collecte des données . . . . .	22
2.1.2	Le nettoyage . . . . .	22
2.1.3	L'exploration . . . . .	23
2.1.4	L'analyse . . . . .	24
2.2	La modélisation . . . . .	25
2.2.1	La sélection et la préparation des données . . . . .	25
2.2.2	La séparation des données . . . . .	26
2.2.3	La phase d'expérimentation et d'évaluation . . . . .	27
2.2.4	La finalisation . . . . .	28
2.2.5	La présentation des résultats . . . . .	28
2.2.6	La maintenance . . . . .	28

# 2 Maîtrisez la Data Science

avec Python

3. Python .....	29
3.1 Les atouts naturels de Python .....	29
3.2 Les librairies spécialisées .....	30
3.3 Plus encore .....	31

## Chapitre 2

### Bases de Python et environnements

1. Les notebooks .....	33
1.1 Principe du notebook .....	33
1.1.1 Fonctionnement par cellule .....	34
1.1.2 Possibilité d'annoter le code .....	34
1.1.3 Affichage de contenu interactif .....	34
1.2 Comment créer un notebook .....	36
1.2.1 Installation directe du module Jupyter .....	36
1.2.2 Installation de la suite Anaconda .....	36
1.2.3 Google Colaboratory .....	37
2. Commandes de base .....	39
2.1 Acquisition des données .....	39
2.1.1 Définition du dossier de travail .....	40
2.1.2 Accès aux données .....	40
2.2 Définition des données .....	42
2.2.1 Changement du type .....	42
2.2.2 Gestion des dates .....	43
2.2.3 Taille du stockage par type .....	44
2.3 Structuration du code .....	46
2.3.1 PEP8 .....	46
2.3.2 Optimisation du code .....	48
3. Utilisation avancée .....	49
3.1 Gestion des librairies .....	49
3.1.1 Installation .....	50
3.1.2 Mise à jour .....	50
3.1.3 Suppression .....	50

3.2 L'environnement virtuel .....	51
3.2.1 Déploiement d'un environnement virtuel .....	51
3.2.2 Utilisation d'un environnement virtuel dans un notebook .....	52
3.3 Les notions utiles pour la data science .....	53
3.3.1 Le pipeline .....	54
3.3.2 La programmation orientée objet (POO) .....	55
3.3.3 Les décorateurs .....	56
3.3.4 La gestion des erreurs .....	58

## Chapitre 3

### Préparer les données avec Pandas et Numpy

1. Pandas, la bibliothèque Python incontournable pour manipuler les données .....	61
1.1 Installation .....	61
1.2 Structure et type de données .....	62
1.3 Possibilités offertes .....	63
2. Numpy, le pilier du calcul numérique .....	64
2.1 La structure ndarray .....	64
2.1.1 Une structure homogène .....	65
2.1.2 L'indexation .....	68
2.1.3 La modification des structures .....	69
2.1.4 La vectorisation .....	73
2.2 La puissance au service du calcul scientifique .....	74
2.3 Les possibilités offertes par Numpy .....	75
2.3.1 Opérations mathématiques de base .....	75
2.3.2 Algèbre linéaire et calculs statistiques .....	76
2.3.3 Création d'images .....	78
3. Collecte des données .....	79
3.1 Acquisition et contrôle des données .....	81
3.1.1 Les formats classiques des fichiers de données .....	81
3.1.2 L'acquisition de données en pratique .....	82

# 4 Maîtrisez la Data Science

avec Python

3.2	Manipulations avancées des données . . . . .	87
3.2.1	Concaténation . . . . .	87
3.2.2	Fusion . . . . .	89
3.2.3	Agrégation. . . . .	90
3.2.4	Export des données. . . . .	93
4.	Nettoyage des données. . . . .	96
4.1	Sélection des données. . . . .	97
4.2	Contrôle de la qualité des données . . . . .	99
4.2.1	Définition du bon type de données. . . . .	99
4.2.2	Gestion des problèmes d'encodage . . . . .	100
4.3	Identification des valeurs atypiques ou aberrantes . . . . .	100
4.3.1	Z-score et méthode des quartiles. . . . .	101
4.3.2	Local Outlier Factor . . . . .	104
4.4	Gestion des outliers . . . . .	106
4.4.1	Suppression des valeurs . . . . .	106
4.4.2	Changement de la distribution . . . . .	107
4.4.3	Conservation des valeurs aberrantes. . . . .	107
4.5	Imputations . . . . .	108
4.5.1	Imputation par la valeur la plus fréquente (modale) . . . . .	108
4.5.2	Imputation par la moyenne ou la médiane. . . . .	109
4.5.3	Imputation par régression . . . . .	110
4.5.4	Imputation basée sur les plus proches voisins (KNN) . . . . .	111
4.5.5	Autres types d'imputations . . . . .	112

## Chapitre 4

### DataViz avec Matplotlib, Seaborn, Plotly

1.	Introduction à la visualisation des données . . . . .	113
1.1	La visualisation au service de la compréhension. . . . .	114
1.2	La méthodologie . . . . .	114
1.2.1	Contextualisation des recherches . . . . .	114
1.2.2	Public concerné. . . . .	115
1.2.3	Les nombreuses possibilités de graphiques . . . . .	115

1.2.4 Règles à respecter concernant les graphiques . . . . .	116
2. Les principales bibliothèques pour la visualisation : Matplotlib, Seaborn et Plotly-Express. . . . .	117
2.1 Matplotlib . . . . .	117
2.1.1 Présentation de Matplotlib . . . . .	117
2.1.2 Premiers pas avec Matplotlib . . . . .	118
2.1.3 Personnalisation et options avancées . . . . .	120
2.2 Seaborn . . . . .	124
2.2.1 Présentation de Seaborn . . . . .	124
2.2.2 Simplification de l'exploration des relations complexes	124
2.3 Plotly.express . . . . .	127
2.3.1 La version simplifiée de Plotly . . . . .	127
2.3.2 L'interactivité de Plotly-Express . . . . .	128
2.3.3 L'avenir de Plotly-Express . . . . .	129
3. Les différents types de graphiques . . . . .	129
3.1 Les enjeux . . . . .	129
3.1.1 Le cheminement vers le bon graphique . . . . .	129
3.1.2 Les postes importants . . . . .	130
3.1.3 Les contraintes . . . . .	130
3.2 Les graphiques univariés . . . . .	133
3.2.1 Graphiques univariés pour les données numériques	133
3.2.2 Graphiques univariés pour les données catégorielles	140
3.2.3 Récapitulatif . . . . .	152
3.3 Les graphiques bivariés et multivariés . . . . .	152
3.3.1 Graphiques bivariés portant sur des variables de même nature . . . . .	153
3.3.2 Graphiques bivariés portant sur des variables de natures différentes . . . . .	159
3.3.3 Graphiques multivariés . . . . .	166
3.4 Les autres types de graphiques . . . . .	172
3.4.1 La cartographie . . . . .	172
3.4.2 Les données temporelles . . . . .	178
3.4.3 Les autres solutions graphiques . . . . .	182

# 6 \_\_\_\_\_ Maîtrisez la Data Science

avec Python

## Chapitre 5

### Analyse des données

1.	Introduction à l'analyse des données . . . . .	185
1.1	Définition et rôle de l'analyse de données . . . . .	186
1.2	Enjeux . . . . .	186
1.2.1	Innovation et créativité . . . . .	187
1.2.2	Prise de conscience des contraintes spécifiques . . . . .	188
1.2.3	Amélioration de la prise de décision . . . . .	189
2.	Statistiques descriptives et inférentielles . . . . .	191
2.1	Description des variables quantitatives . . . . .	192
2.1.1	Mesures de tendance centrale . . . . .	192
2.1.2	Mesures de dispersion . . . . .	198
2.1.3	La distribution . . . . .	203
2.2	Description des variables catégorielles . . . . .	207
2.2.1	Fréquence, proportion et gestion des modalités rares . . . . .	207
2.2.2	Tableau de contingence . . . . .	209
2.2.3	Indices de diversité . . . . .	210
2.3	Statistiques inférentielles . . . . .	215
2.3.1	Concepts de base . . . . .	215
2.3.2	Hypothèses nulles et alternatives . . . . .	215
2.3.3	P-value . . . . .	216
2.3.4	Significativité . . . . .	216
2.3.5	Marge d'erreur et impact des effectifs sur l'intervalle de confiance . . . . .	217
3.	Modules Python pour l'analyse de données . . . . .	219
3.1	Les capacités limitées des modules classiques . . . . .	219
3.2	Les modules spécialisés en statistiques . . . . .	220
3.2.1	Scipy . . . . .	220
3.2.2	Statmodels . . . . .	221
4.	Tests statistiques de normalité . . . . .	221
4.1	Contexte et objectif . . . . .	221
4.2	Les Q-Q plots . . . . .	222

4.2.1	Définition et tracé du graphique . . . . .	222
4.2.2	Interprétation . . . . .	223
4.3	Principe de fonctionnement général des tests de normalité . .	224
4.3.1	Principe de fonctionnement . . . . .	224
4.3.2	Les différents tests de normalité . . . . .	225
5.	Tests statistiques bivariés . . . . .	228
5.1	Tests bivariés entre des variables de même nature. . . . .	229
5.1.1	Corrélations entre variables numériques . . . . .	229
5.1.2	Tests d'indépendance entre variables catégorielles . .	235
5.2	Tests bivariés entre des variables de nature différente. . . .	241
5.2.1	Tests de comparaison à deux modalités . . . . .	241
5.2.2	Tests de comparaison à trois modalités ou plus. . . . .	243
5.2.3	Conclusions sur les tests bivariés . . . . .	249
6.	Analyse multivariée . . . . .	249
6.1	Analyse de la variance multivariée (MANOVA) . . . . .	250
6.1.1	Présentation et champs d'applications . . . . .	250
6.1.2	Cas pratique d'utilisation. . . . .	250
6.2	Analyse en composantes multiples (ACM) . . . . .	252
6.3	Analyse en composantes principales (ACP) . . . . .	255
6.3.1	Un des piliers de la data science. . . . .	255
6.3.2	Utilisation sur un cas pratique . . . . .	256
6.3.3	L'éboulis des valeurs propres . . . . .	257
6.3.4	Le cercle des corrélations . . . . .	258
6.3.5	Le graphique des individus. . . . .	259

## Chapitre 6

### Le Machine Learning avec Scikit-Learn

1.	Introduction au Machine Learning : concepts et types de modèles. . . . .	263
1.1	L'apprentissage non supervisé . . . . .	264
1.1.1	Définition . . . . .	264
1.1.2	La réduction dimensionnelle . . . . .	265

1.1.3	Le clustering . . . . .	267
1.2	L'apprentissage supervisé . . . . .	269
1.2.1	Introduction . . . . .	269
1.2.2	Régression . . . . .	270
1.2.3	Classification . . . . .	271
1.3	Le texte et l'image . . . . .	273
1.3.1	Définitions des concepts . . . . .	273
1.3.2	Le texte et le NLP . . . . .	273
1.3.3	Le traitement des images . . . . .	274
2.	Présentation de Scikit-Learn, la bibliothèque Python pour la data science . . . . .	276
2.1	Une offre simple et complète de fonctionnalités . . . . .	276
2.2	Des méthodes communes aux différentes fonctions . . . . .	277
2.2.1	La méthode fit() . . . . .	278
2.2.2	Les méthodes transform et fit_transform . . . . .	279
2.2.3	La méthode predict . . . . .	280
2.2.4	La méthode score() . . . . .	280
2.2.5	Les méthodes get_params et set_params . . . . .	281
2.3	Le soutien de la licence BSD et d'une communauté active . . . . .	282
3.	Les grandes étapes d'un projet de Machine Learning . . . . .	282
3.1	La préparation des données . . . . .	282
3.1.1	La séparation des variables explicatives de la variable cible . . . . .	282
3.1.2	La séparation entre données d'entraînement et données de test . . . . .	283
3.1.3	Les transformations des variables . . . . .	284
3.1.4	La mise en œuvre ciblée des transformations . . . . .	287
3.1.5	Finalisation de la préparation des données . . . . .	290
3.2	L'expérimentation . . . . .	291
3.2.1	Définition des métriques pour l'évaluation . . . . .	292
3.2.2	Les algorithmes d'optimisation d'hyperparamètres . . . . .	295
3.2.3	Le modèle de base (DummyRegressor et DummyClassifier) . . . . .	295

3.2.4	Tests des divers algorithmes avec différentes combinaisons de paramètres . . . . .	297
3.2.5	L'évaluation et le choix final . . . . .	299
4.	Conclusions sur la modélisation . . . . .	301

## Chapitre 7

### L'apprentissage supervisé

1.	Introduction . . . . .	303
2.	Les familles d'algorithmes . . . . .	303
2.1	Les algorithmes linéaires . . . . .	304
2.1.1	Les régressions . . . . .	304
2.1.2	Les régressions régularisées . . . . .	307
2.1.3	Les machines à vecteur de support (SVM) . . . . .	310
2.2	Les algorithmes semi-linéaires (modèles à noyau) . . . . .	313
2.3	Les algorithmes non linéaires . . . . .	317
2.3.1	Les plus proches voisins (KNN) . . . . .	317
2.3.2	L'arbre de décision . . . . .	319
2.3.3	Les méthodes ensemblistes . . . . .	321
2.3.4	Les réseaux de neurones . . . . .	327
3.	La régression en pratique . . . . .	330
3.1	Préparation des données . . . . .	331
3.1.1	Import des données . . . . .	331
3.1.2	Séparation des variables explicatives de la variable cible . . . . .	332
3.1.3	Séparation entre données d'entraînement et de test . . . . .	332
3.1.4	Les transformations des variables . . . . .	333
3.1.5	Finalisation de la préparation des données . . . . .	333
3.2	Fonction de calcul et d'affichage des régressions . . . . .	335
3.3	La modélisation d'une régression . . . . .	337
3.3.1	Modèle de base (DummyRegressor) . . . . .	337
3.3.2	Test des algorithmes concurrents . . . . .	338
3.3.3	Le pipeline . . . . .	343

# 10 Maîtrisez la Data Science

## avec Python

4. La classification en pratique.....	347
4.1 Préparation des données.....	347
4.1.1 Import des données .....	347
4.1.2 Séparation entre les variables explicatives et la variable cible .....	347
4.1.3 Séparation entre données d'entraînement et de test ..	347
4.1.4 Transformation des colonnes .....	348
4.1.5 Remise en forme des noms .....	348
4.1.6 Ajustement du type des variables .....	349
4.2 Fonction de calcul et d'affichage des classifications.....	349
4.3 Expérimentations .....	352
4.3.1 Modèle de base (DummyClassifier) .....	352
4.3.2 Algorithmes concurrents .....	354
5. Conclusion .....	359

## Chapitre 8

### L'apprentissage non supervisé

1. Introduction .....	363
2. La réduction dimensionnelle .....	364
2.1 L'ACP en pratique pour analyser.....	364
2.1.1 Préparation des données.....	364
2.1.2 L'éboulis des valeurs propres .....	367
2.1.3 Le cercle des corrélations .....	370
2.1.4 Le graphique des individus.....	373
2.2 L'ACP en pratique pour modéliser.....	376
2.3 Les autres algorithmes de réduction dimensionnelle .....	378
3. Le clustering .....	383
3.1 La pratique du clustering avec le K-means .....	383
3.1.1 Acquisition et préparation des données .....	383
3.1.2 Les tests pour déterminer le nombre de clusters .....	386
3.1.3 Choix du clustering .....	389
3.1.4 Le score ARI .....	391

3.2	Les autres algorithmes de clustering .....	392
3.2.1	GMM .....	392
3.2.2	Meanshift .....	394
3.2.3	DBSCAN .....	396

## Chapitre 9

### Modéliser le texte et l'image

1.	La modélisation du texte .....	401
1.1	Les modules du NLP .....	402
1.1.1	NLTK .....	402
1.1.2	TextBlob .....	404
1.1.3	spaCy .....	405
1.2	Mise en pratique de la NLP .....	407
1.2.1	Prétraitement des données .....	407
1.2.2	Les extracteurs de caractéristiques .....	411
1.2.3	La modélisation .....	412
1.3	Introduction aux modèles avancés en NLP .....	418
1.3.1	Les représentations de mots .....	418
1.3.2	L'encodage des phrases .....	420
1.3.3	Transformers et modèles contextuels .....	420
1.3.4	Les Larges Languages Models (LLM) .....	421
2.	La modélisation des images .....	421
2.1	Les solutions de Machine Learning destinées aux images .....	422
2.1.1	Pillow pour s'initier au prétraitement .....	422
2.1.2	Scikit-image .....	426
2.1.3	OpenCV .....	431
2.2	Méthodes de modélisation des images .....	433
2.2.1	Segmenter .....	434
2.2.2	Détecter .....	438
2.2.3	Classifier .....	441

# 12 \_\_\_\_\_ Maîtrisez la Data Science

avec Python

2.3	Aller plus loin avec les CNN . . . . .	443
2.3.1	Principe de fonctionnement du CNN . . . . .	443
2.3.2	Transfer learning . . . . .	444
2.3.3	Initiation à Tensorflow et Keras . . . . .	445
2.3.4	Exemples d'utilisation des CNN . . . . .	446

## Chapitre 10

### Mener un projet de data science avec Python

1.	Introduction . . . . .	455
2.	Le sujet : déterminer le prix des véhicules d'occasion . . . . .	455
2.1	Les données . . . . .	455
2.2	Les étapes du projet . . . . .	456
2.2.1	Le notebook de l'EDA . . . . .	456
2.2.2	Le notebook de modélisation . . . . .	456
2.2.3	Les aléas des données . . . . .	457
3.	La modélisation en pratique . . . . .	457
3.1	Notebook 1 : EDA . . . . .	457
3.1.1	Acquisition et premiers contrôles des données . . . . .	457
3.1.2	Nettoyage des données . . . . .	460
3.1.3	Exploration et analyse . . . . .	467
3.2	Notebook 2 : modélisation simple . . . . .	480
3.2.1	Acquisition et sélection des données . . . . .	480
3.2.2	Modélisation . . . . .	482
3.2.3	Résultats . . . . .	484
3.3	Notebook 3 : modélisation mixte . . . . .	491
3.3.1	Acquisition et sélection des données . . . . .	491
3.3.2	Modélisation . . . . .	493
3.3.3	Résultats . . . . .	494
4.	Conclusion . . . . .	496

**Conclusion**

1. Le rôle central des données et de leur compréhension .....	497
2. Des évolutions qui transforment et accélèrent tout .....	498
2.1 L'évolution du matériel technologique .....	498
2.2 L'amélioration des modèles .....	499
2.3 La diffusion dans le grand public et la prise en compte progressive des enjeux .....	499
3. Importance de la théorie et invitation à l'exploration .....	500
Index .....	501