

Cours d'analyse de données en géographie
Niveau Master 1 - GEANDO
Support pour les étudiants

Maxime Forriez^{1,a}

¹ Institut de géographie, 191, rue Saint-Jacques, Bureau 105, 75 005 Paris,
^amaxime.forriez@sorbonne-universite.fr

18 septembre 2025

Première partie

Pour commencer avec de bonnes bases...

Séance 1

Présentation du contenu du cours et mise en route

Objectifs de la séance

- Présentation des objectifs pédagogiques
- Prise en main de `GitHub`
- Prise en main de `Docker`
- Établissement d'une architecture propre
- Prise en main de `Python` avec `Docker`
- Découverte de `Python`

1.1 Présentation des objectifs pédagogiques

1.1.1 Règles générales du fonctionnement du cours

Avant chaque séance, vous **devez lire** son contenu se trouvant dans ce livret **et** tous les éléments présents sur mon compte couvert sur la plateforme `GitHub`. L'organisation du cours suit la **logique de la pédagogie inversée**. Chaque séance proposera un exercice d'application du contenu que vous aurez lu. Elle ne résumera **jamais** ce que vous n'avez pas lu. Bien entendu, si vous avez des questions, je reste à votre disposition en cours ou par mon courriel professionnel que ce soit **pendant** la durée du module ou **après** lorsque vous manipulerez les données de votre mémoire dans la limite de mes compétences et de mon temps disponible.

Chaque séance se déroulera de la même manière.

1. Vos questions sur ce que vous avez lu
2. Présentation de l'analyse de données du jour
3. Téléchargement des ressources sur `GitHub`
4. Travail en local avec `Python` sous `Docker`

5. Remise des résultats que vous avez obtenus en fin de cours dans un portfolio sur votre compte `GitHub`

Une explication de la plateforme `GitHub` est disponible sur `GitHub`.

Un contenu détaillé de l'utilisation de `Python` est disponible sur `GitHub`.

Un formulaire mathématique est à votre disposition sur `GitHub`.

1.1.2 Calendriers

Débutants

7 séances :

Séance 1. Mercredi 17 septembre 2025 – 12h00-14h00 – Clignancourt : Salle 426

Séance 2. Mercredi 8 octobre 2025 – 12h00-14h00 – Clignancourt : Salle 426

Séance 3. Mercredi 22 octobre 2025 – 12h00-14h00 – Clignancourt : Salle 426

Séance 4. Mercredi 12 novembre 2025 – 12h00-14h00 – Clignancourt : Salle 426

Séance 5. Mercredi 19 novembre 2025 – 12h00-14h00 – Clignancourt : Salle 426

Séance 6. Mercredi 26 novembre 2025 – 12h00-14h00 – Clignancourt : Salle 426

Séance 7. Mercredi 3 décembre 2025 – 12h00-14h00 – Clignancourt : Salle 426

Intermédiaires

6 séances :

Séance 1. Mercredi 1^{er} octobre 2025 – 12h00-14h00 – Clignancourt : Salle 426

Séance 2. Mercredi 15 octobre 2025 – 12h00-14h00 – Clignancourt : Salle 426

Séance 3. Mercredi 5 novembre 2025 – 12h00-14h00 – Clignancourt : Salle 426

Attention ! Ne pas oublier de faire une copie des données sur votre P.C. de l'université pour les faire basculer sur votre nouveau P.C. dans la nouvelle salle.

Séance 4. Mercredi 19 novembre 2025 – 10h00-12h00 – Clignancourt : Salle 217

Séance 5. Mercredi 26 novembre 2025 – 10h00-12h00 – Clignancourt : Salle 217

Séance 6. Mercredi 3 décembre 2025 – 10h00-12h00 – Clignancourt : Salle 217

Confirmés (GEOINT)

7 séances :

Séance 1. Mercredi 17 septembre 2025 – 10h00-12h00 – Clignancourt : Salle 217

Séance 2. Mercredi 1^{er} octobre 2025 – 10h00-12h00 – Clignancourt : Salle 217

Séance 3. Mercredi 8 octobre 2025 – 10h00-12h00 – Clignancourt : Salle 217

Séance 4. Mercredi 15 octobre 2025 – 10h00-12h00 – Clignancourt : Salle 217

Séance 5. Mercredi 22 octobre 2025 – 10h00-12h00 – Clignancourt : Salle 217

Séance 6. Mercredi 5 novembre 2025 – 10h00-12h00 – Clignancourt : Salle 217

Séance 7. Mercredi 12 novembre 2025 – 10h00-12h00 – Clignancourt : Salle 217

1.1.3 Parcours

En fonction de votre parcours, vous ne ferez pas les mêmes séances numérotées dans ce livret.

Débutants. 1, 2, 3, 4, 5, 6

N.B. La septième séance reste à déterminer.

Intermédiaires. 1, 4, 5, 6, 7, 8

Confirmés. 1, 5, 6, 7, 8, 9, 10

1.1.4 Introduction générale à l'analyse des données

Ce cours d'analyse de données est à la croisée de trois grands champs :

- la géographie, évidemment ;
- les données ;
- l'informatique, notamment grâce au langage de programmation `Python`.

En quelques heures de cours, il est difficile d'être exhaustif sur l'ensemble de ces champs. C'est pour cela qu'il sera orienté sur la manipulation des données avec le langage de programmation `Python`, que vous devrez, à force le pratiquer, un minimum maîtriser à la fin des séances.

Objectifs à atteindre

- Connaître et savoir manipuler différents formats de données, notamment géographiques
- Apprendre à lire des équations mathématiques et à savoir les utiliser dans un contexte de mathématiques appliquées à la géographie
- Connaître les principales méthodes statistiques (univariées, bivariées et multivariées)
- Savoir utiliser les bonnes méthodes en fonction des données géographiques à analyser, mais également des objectifs à atteindre
- Apprendre utiliser des outils informatiques professionnels (`Git` et `Docker`)
- Connaître et apprendre à utiliser les outils de calcul informatique grâce à un langage de programmation (`Python`)
- Réviser ou apprendre les notions mathématiques associées aux outils d'analyse (cf. le dossier « Formulaire de mathématiques » sur `GitHub`)

Les données en géographie

Les **sciences des données** sont appelés *data science*. Les étapes pour traiter les données sont :

1. trouver les données ;
2. nettoyer les données ;

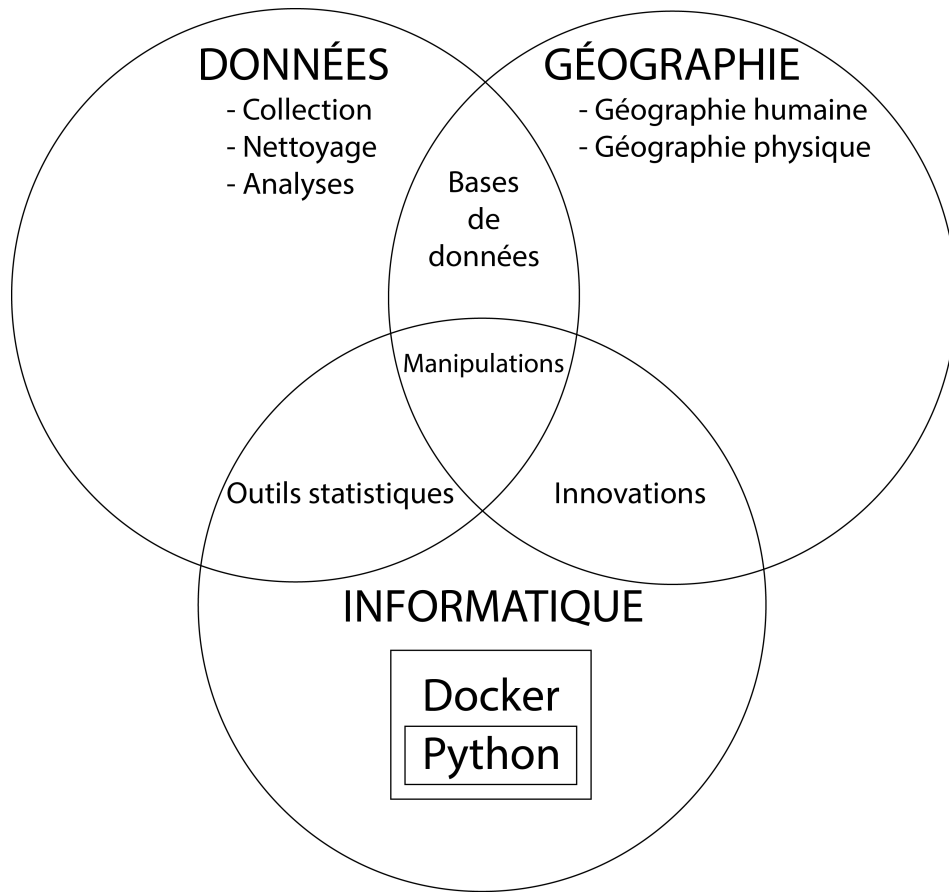


FIGURE 1.1 – Positionnement du cours

Source : Maxime Forriez (2025)

3. explorer les données ;
4. identifier les analyses nécessaires ;
5. produire les sorties finales et les visualisations diffusables.

La géographie à l'ère des humanités numériques

Il existe un grand mouvement de digitalisation de toutes les données, parmi elles les données géographiques. Ce mouvement est appelé les **humanités numériques**. Les spécialistes des données sont très recherchés de nos jours. si vous n'apprenez pas à coder, ce seront les autres sciences qui analyseront les données géographiques.

La géographie et les mathématiques, entre rapprochement et éloignement

Géographie et mathématiques ont un lien très fort. La géographie est l'origine de la géométrie par exemple. Avec la mathématisation de la cartographie aux XVII^e-XVIII^e siècles, les mathématiques et la géographie se sont peu à peu éloigné l'une de l'autre. À partir du milieu du XX^e siècle, les statistiques ont peu à peu réintégré la géographie sous la forme des analyses spatiales (géostatistique, géotraitement, *etc.*). Cela a été accompagné par un développement des outils informatiques qui remplacent peu à peu les méthodes traditionnelles de

cartographie. Pour faire des analyses de données géographiques, il faut maîtriser deux éléments : 1. l'informatique ; 2. les statistiques. Les données massives ne font que renforcer les liens entre ces outils et la géographie grâce à la géomatique.




L'objectif n'est pas de faire de vous des experts absolus dans ces deux outils, mais de vous faire acquérir un **niveau intermédiaire**, c'est-à-dire autonome.

Je vous informe que, dans le cadre de ce cours, nous ferons des **mathématiques appliquées**. L'objectif n'est pas de faire de vous des mathématiciens émérites, mais simplement de vous apprendre à comprendre les enjeux statistiques de tel ou tel jeu de données et d'appliquer les bonnes formules en ayant les bons réflexes d'interprétation. C'est l'ordinateur qui s'occupe des calculs grâce au langage `Python`. Vous devez juste savoir manipuler du code `Python` et interpréter les résultats des analyses, et non d'être capable de les faire à la main. Bien entendu, si vous êtes curieux, je vous invite fortement à tenter de le faire. Cela vous permettra de comprendre en profondeur les formules. Le contenu de ce livret vous y aidera.

Problème classique. Les données collectées sont mal organisées et compliquées, c'est-à-dire qu'elles ont des formats différents, une architecture inexistante, *etc.* `Python` servira à résoudre ce problème.

1.1.5 Aperçu des métiers

Data Analyst vs Data Scientist vs Data Engineer

Data Analyst	Data Scientist	Data Engineer
		
Analyzes data to extract insights	Builds models to predict outcomes	Develops data pipelines and infrastructure
Tools • Excel, SQL Tableau	Tools • Python, R, scikit-learn	Tools • SQL, Python, Spark

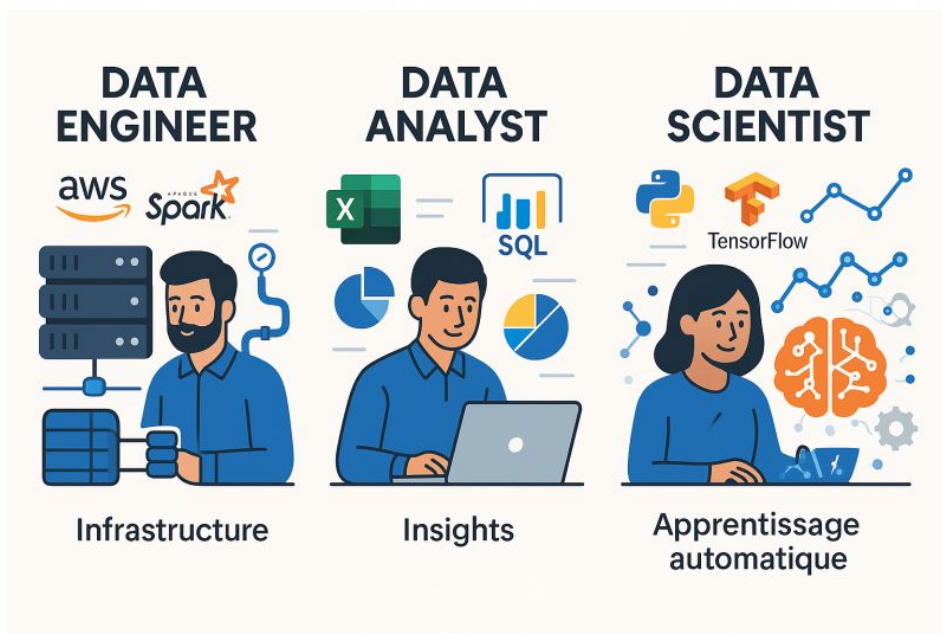


FIGURE 1.2 – Les métiers autour de l'analyse de données

Source : LinkedIn






Data Analyst	Data Scientist	Business Analyst	ML Engineer	GenAI Engineer
				
Analyze existing data to generate insights and support data-driven decision-making	Develop and implement statistical models and machine learning algorithms to derive insights and make data driven predictions	Analyze and document business processes to identify opportunities, requirements, and recommendsl for improvement	Design, develop, and deploy machine learning systems to ensure scalability, performance, and reliability in production	Develop and deploy generative AI models and applications for content generation, automation, and personalized experiences
Skills:	Skills:	Skills:	Skills:	Tools:
SQL Excel Statistics Data Visualization Python	Python / R Machine Learning Statistics and Probability Data Wrangling and Feature Eng Data Visualization	Business Process Modeling Communication & Requirements Gathering Data Analysis Problem-Solving Basic SQL & Excel	Machine Learning Data Engineering (ETL, Pipelines) Python / Java SQL Big Data Tools	Python (Transformers PyTorch, TensorFlow) HuggingFace LangChain
Tools:	Tools	Tools	Software Engineering	Tools
SQL Excel Tableau Jira	Excel Tableau	Microsoft-Office Suite	Tools Spark, Hadoop	LangGAPis LLAMs

FIGURE 1.3 – Les métiers autour de l'analyse de données

Source : LinkedIn



FIGURE 1.4 – Comment devenir un spécialiste des données avec Python ?

Source : LinkedIn

1.1.6 Évaluation

Toutes les séances sont notées sous la forme d'un portfolio à déposer sur votre compte GitHub.

Attention ! Toute absence est rarement rattrapable.

Attention ! Soyez à l'heure à chaque séance. Plus vous êtes en retard, moins vous aurez de temps pour faire l'analyse du jour.

1.2 Prise en main de GitHub

GitHub est une plateforme collaborative de versionnage. Le dossier du cours se localise à l'adresse suivante : <https://github.com/MaximeForriez/Sorbonne-M1-Analyse-de-donnees>.

Nous utiliserons GitHub simplement pour son dépôt. Nous ne l'utiliserons pas à son plein potentiel. Si vous êtes intéressé par les métiers de l'analyse de données, je ne saurais que trop vous conseiller d'apprendre à l'utiliser.

Vous devez ouvrir un compte sur GitHub. C'est sur votre dépôt que vous déposerez les éléments qui seront évalués à la fin du semestre. Votre utilisation de la plateforme fait partie intégrante de votre notation.

— Nom d'utilisateur :

— Mot de passe :

Si vous perdez votre mot de passe, il n'existe aucun moyen de le récupérer. Notez le bien.

Une explication du fonctionnement basique de Git est disponible au lien suivant : <https://github.com/MaximeForriez/Sorbonne-M1-Analyse-de-donnees/Git/Git.md>.

Si vous souhaitez approfondir la question, vous pouvez le faire au lien suivant : <https://github.com/MaximeForriez/Sorbonne-M1-Analyse-de-donnees/Git/Git-Avance.md>.



FIGURE 1.5 – Git Cheat Sheet

Source : LinkedIn

1.3 Prise en main de Docker

Vous devez ouvrir un compte sur Docker.

1. Aller sur le site `https://www.docker.com`
2. Sur la page d'accueil, cliquer sur le bouton `Sign In`
3. Une page d'identification s'ouvre. Comme vous n'avez pas de compte, aller en bas de la page, et cliquer sur le lien `Sign Up`
4. Une page de création de compte s'ouvre. Par défaut, elle s'ouvre sur un compte professionnel `Work`. Cliquer sur l'onglet `Personal`
5. Saisissez votre courriel, votre nom d'utilisateur et votre mot de passe
 - Courriel :
 - Nom d'utilisateur :
 - Mot de passe :

Si vous perdez votre mot de passe, il n'y a aucun moyen de le récupérer. Notez le bien.

Normalement, un logiciel appelé `Docker Desktop` est installé sur votre machine. Pour vérifier sa présence, ouvrez un terminal de type `Invite de commande`.

1. Cliquer sur le bouton `Windows` dans votre barre de menu.
2. Laisser le menu affiché tout en tapant `cmd`
3. L'`Invite de commande` est disponible. Cliquer dessus pour l'ouvrir ou faites `Entrée`
4. Une fenêtre noire s'affiche. Pour ceux qui n'ont pas l'habitude, pas de panique !
5. Taper (le contenu après le symbole `>`) :

```
C:\...\> docker -v
```

Si un numéro de version avec trois nombres séparés par un point apparaît, `Docker Engine` est bien installé sur votre machine

```
Docker version 26.0.0, build 2ae903e
```

Si vous avez une erreur, vous m'appellez immédiatement.

Ouvrez le logiciel `Docker Engine` en le cherchant de la même manière que l'`Invite de commande`. Nous allons utiliser `Docker` pour installer la même version de `Python` sur toutes les machines, mais, pour cela, il faut créer un pont entre votre machine locale et le serveur `Docker Hub`.

1. Cliquer sur l'onglet `Hub`
2. Enregistrer les identifiants que vous avez créés précédemment avec le bouton `Sign In`
3. Votre navigateur par défaut s'ouvrira et vous pourrez entrer vos données.

N.B. À la fin de chaque séance, il vous faudra :

1. Fermer votre compte personnel pour la personne suivante, sinon elle aura accès à toutes vos données personnelles

2. Éteindre le Docker Engine. En bas, à gauche, un onglet vert indique l'exécution du programme. Pour l'éteindre, il suffit de cliquer sur le bouton d'arrêt Quit Docker Desktop. C'est très important sinon le Docker Engine s'ouvre à chaque ouverture de l'ordinateur, ce qui le ralentira considérablement. Pensez à vos camarades !
3. Sauvegarder votre travail sur une clé U.S.B.
4. Même s'il y a un groupe après vous, éteignez votre P.C. Cela permet de vider la mémoire vive et de donner un ordinateur propre à l'étudiant suivant.

Désormais, vous êtes attaché à un numéro de poste. Vous ne pourrez plus en changer.

— Numéro de poste en salle _____ :

Courage ! C'est presque fini.

1.4 Établissement d'une architecture propre

Sur votre machine, créer un dossier avec votre nom, l'année universitaire et le titre du cours dans le lecteur C : \.

```
Forriez-2025-2026-Analyse-de-donnees
```

N.B. Mettez bien les -.

Dans ce dossier, créer autant de sous-dossiers que de séances :

```
Seance-01
Seance-02
Seance-03
Seance-04
Seance-05
Seance-06
Seance-07
```

N.B. N'oubliez pas de sauvegarder votre travail sur une clé U.S.B. et sur votre dépôt GitHub en fin de cours.

1.5 Prise en main de Python avec Docker

Pour exécuter Python, sous Docker, il faut que vous alliez récupérer sur GitHub trois fichiers :

1. Dockerfile
2. docker-compose.yml
3. requirements.txt

Chaque fichier doit être téléchargé et placé dans le dossier de la séance du jour.

Attention ! Il est strictement interdit de renommer ces fichiers. Si vous opérez plusieurs téléchargements, et qu'un nombre se glisse dans le nom du fichier, il faut le retirer en renommant le fichier.

Ces trois fichiers sont des fichiers de configuration. Ils sont conçus pour que vous puissiez réaliser toutes les séances. En principe, sauf si, entre temps, vous devenez des experts de Docker, **vous n'écrirez rien dans ces trois fichiers.**

Vous devez également récupérer le contenu du dossier `src` (source), le fichier `main.py`. **Attention !** Vous devez reconstituer l'architecture `src\main.py` dans le dossier de chacune des séances, sinon ça ne marchera pas.

Placez-vous dans le dossier `Seance-01`, et non dans `Seance-01\src`, c'est super important.

1. Aller dans la barre d'adresse et y taper `cmd`, puis Entrée pour ouvrir l'Invite de commande dans ce dossier

N.B. Si vous savez naviguer dans un terminal Windows, vous pouvez utiliser vos connaissances pour vous mettre au bon emplacement.

2. Taper la commande en respectant la casse (minuscules et majuscules) :

```
C:\Forriez-2025-2026-Analyse-de-donnees\Seance-01 > docker-
compose up -d
```

Félicitations ! Vous venez de créer et de lancer une image Docker. Elle est configurée par les fichiers `Dockerfile` et `docker-compose.yml`

N.B. Si vous avez déjà créé un conteneur Docker, et que vous en modifiez les paramètres initiaux, c'est-à-dire que vous modifiez les trois fichiers de configuration, il faudra taper :

```
C:\Forriez-2025-2026-Analyse-de-donnees\Seance-01 > docker-
compose up -d --build
```

ce qui lancera un nouveau conteneur qui prendra en compte vos modifications.

3. Un téléchargement démarre. La première installation peut durer quelques minutes.

Ouvrez le logiciel Notepad++ et y ouvrir le fichier `main.py` du jour. Il contient le code « très utile » :

```
print('Bienvenue au cours d'analyse de données en géographie !')
```

Chaque programme Python s'exécute en ligne de commande. Avec Docker, il faut taper la commande suivante pour exécuter votre code :

```
C:\Forriez-2025-2026-Analyse-de-donnees\Seance-01 > docker-compose
run python
```

À chaque modification de votre code, il faudra effectuer cette commande. Vous finirez par la connaître par cœur.

N.B. Sans Docker, placer dans le dossier contenant les fichiers, il suffirait de taper : `python nom du fichier à exécuter`. Cela suppose une installation manuelle de Python qui prend beaucoup de temps, mais vous pouvez vous reporter aux tutoriels en ligne proposés sur GitHub dans le dossier consacré à Python.

Attention ! Sauf exception clairement précisée dans l'exercice du jour, le point d'entrée de tous les exercices sera toujours un fichier `main.py`, même si plusieurs fichiers Python existent.

Attention ! Vous ne devez jamais fermer sauvagement la fenêtre de l'Invite de commande. Vous devez d'abord éteindre le programme Docker Compose avec la commande :

```
C:\Forriez-2025-2026-Analyse-de-donnees\Seance-01 > docker-compose
down
```

Une fois Python arrêté, vous pourrez fermer la fenêtre.

1.6 Découverte de Python

Docker peut apparaître superflu, mais cette couche supplémentaire va permettre de gérer beaucoup plus simplement avec Python, surtout si vous n'avez jamais codé de votre vie.

Python fonctionne avec un système de paquets¹. Lorsque vous avez besoin d'un outil spécifique, il faut systématiquement le télécharger, ce qui peut être fastidieux, notamment parce qu'il faut régulièrement mettre les paquets à jour manuellement. Grâce à Docker, tous les paquets sont regroupés dans un fichier unique : `requirements.txt`, pris en charge par les deux autres fichiers de configuration.

Sur le GitHub, un dossier documentaire regroupant l'essentiel de ce qu'il faut connaître du langage est mis à votre disposition. Bien entendu, cela ne doit pas vous empêcher de lire les documentations officielles en ligne.

Parmi les paquets, vous pouvez télécharger des **bibliothèques**. Pour faire vos analyses de données, vous allez utiliser principalement :

- Pandas
- NumPy
- Matplotlib
- *etc.*

Il faudra utiliser des paquets pour lire des données Excel, C.S.V., S.Q.L., JSON, *etc.*

N.B. En principe, la totalité des paquets utile pour ce cours est contenu dans le fichier `requirements.txt`. Toutefois, il se peut que vous en ayez besoin d'autres. Dans ce cas, il faudra l'ajouter dans la liste des paquets contenus dans ce fichier, puis arrêter `docker-compose`, et le relancer avec la commande `--build` pour télécharger le nouveau paquet dans votre conteneur Docker.

```
C:\Forriez-2025-2026-Analyse-de-donnees\Seance-01 > docker-compose
down
C:\Forriez-2025-2026-Analyse-de-donnees\Seance-01 > docker-compose
up -d --build
```

Pourquoi apprendre le Python ? R n'est-il pas suffisant ?

1. En anglais : *package*

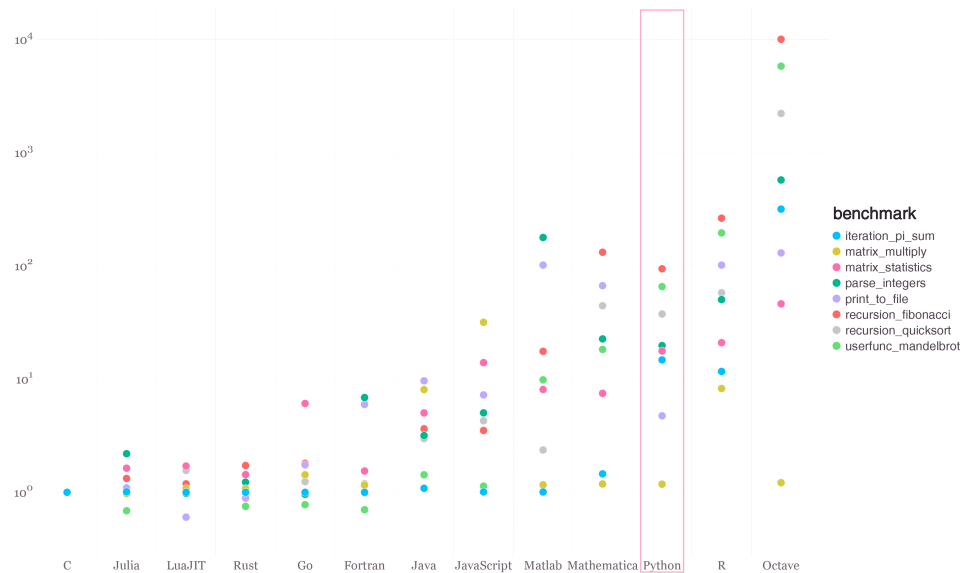


FIGURE 1.6 – Comparaison des temps de calcul entre le Python et d’autres langages de programmation avec différents algorithmes classiques

Source : <https://julialang.org/benchmarks>

Mettez la liste des bibliothèques complémentaires que vous identifierez tout le semestre :

Important Python Libraries and Frameworks

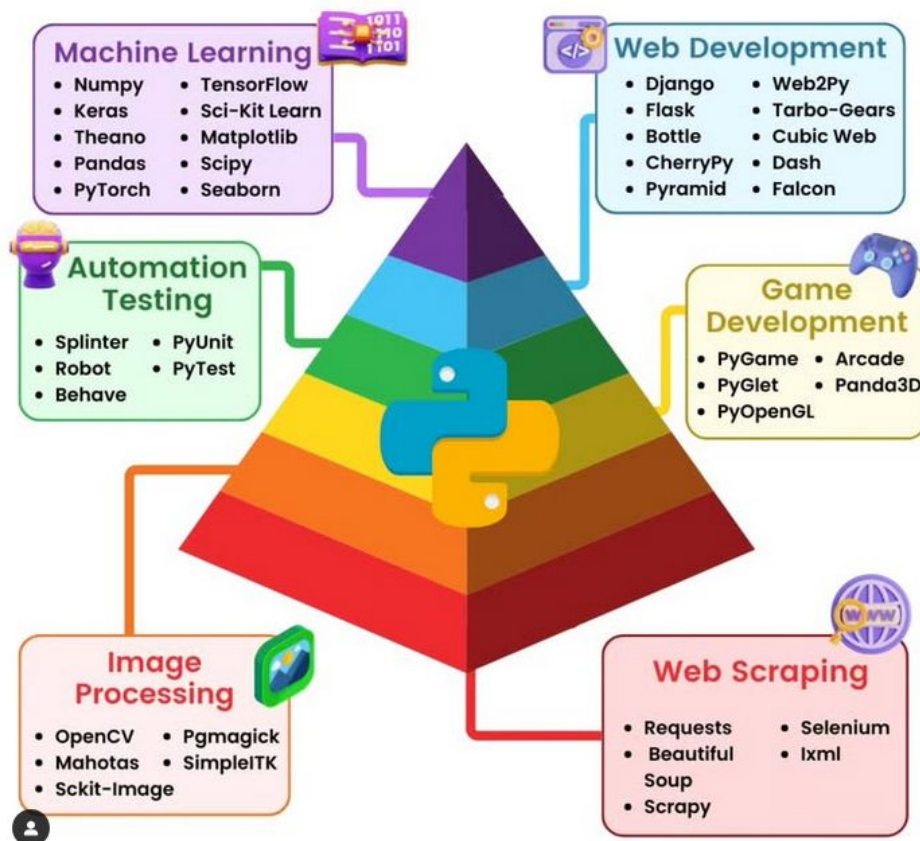


FIGURE 1.7 – Les principales bibliothèques de Python 1

Source : LinkedIn



FIGURE 1.8 – Les principales bibliothèques de Python 2

Source : LinkedIn



FIGURE 1.9 – Les principales bibliothèques de Python 3
Source : LinkedIn

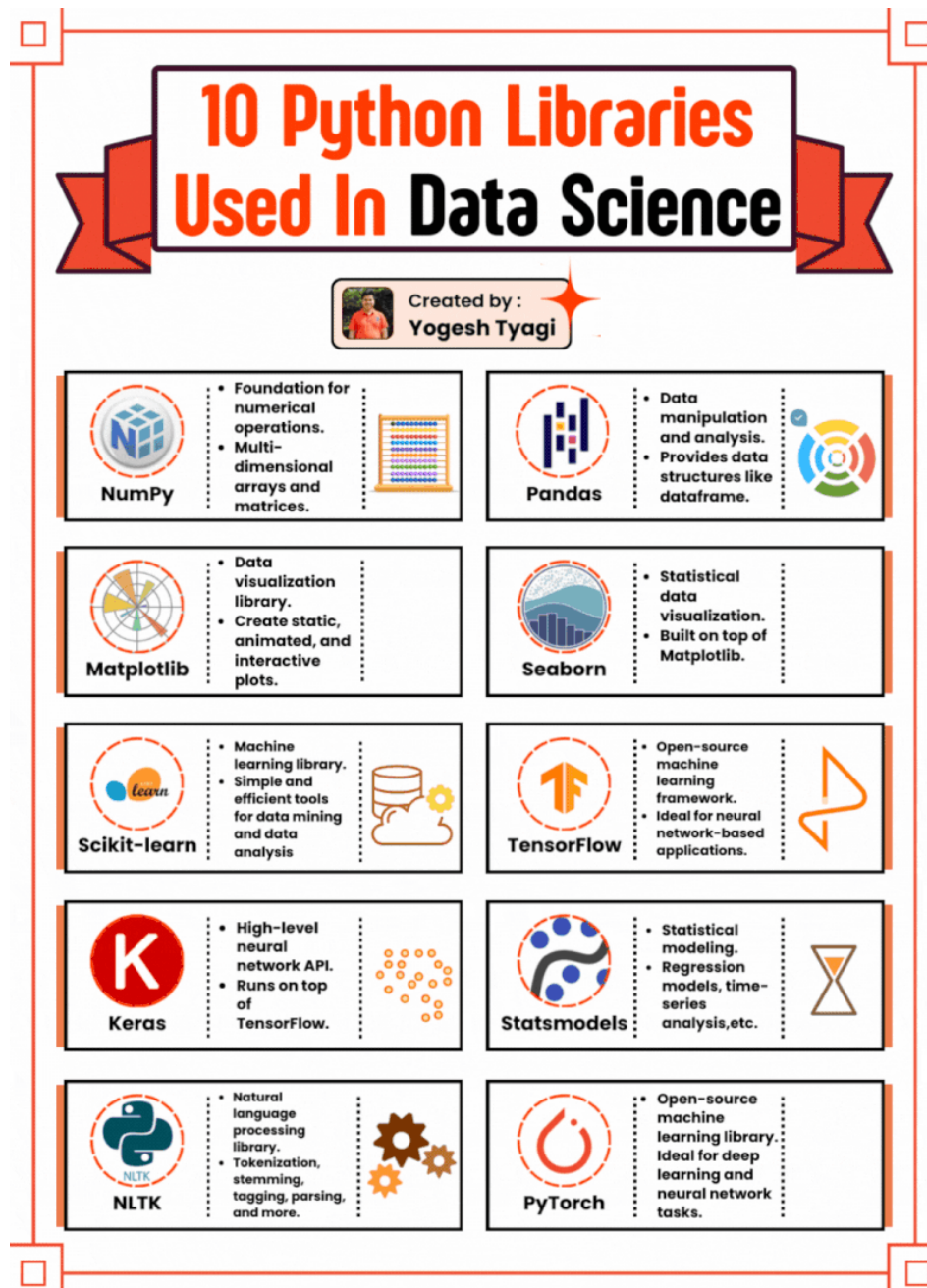


FIGURE 1.10 – Les principales bibliothèques de Python 4

Source : LinkedIn

1.6.1 Objectifs à atteindre

- Comprendre les bases de la programmation avec un outil de plus en plus utilisé, y compris dans les logiciels de géomatique comme ArcGIS ou Q-GIS
- Pouvoir utiliser ces bases dans des routines permettant des analyses statistiques univa-

riées, bivariées et multivariées

- Se familiariser avec la programmation afin d’obtenir une culture minimale de ce que c’est pour vos recherches, mais également votre citoyenneté

1.6.2 Syntaxe de base

1.6.3 Élément d’algorithmique

Les variables

Les opérateurs

Les instructions

Les conditions

Les boucles

Les boucles conditionnelles

Les fonctions et les procédures

Les objets

Conclusion

Si vous maîtrisez la manière d'utiliser ces éléments **presque** universels à chaque langage de programmation, vous pourrez apprendre n'importe quel autre langage.

PYTHON CHEAT SHEET

Basic Commands:

- `print("Hello, World!")`: Display output
- `type(x)`: Get data type of x
- `id(x)`: Get memory address of x
- `help()`: Show documentation
- `dir(obj)`: List object attributes and methods

Conditional Statements

- `for`: Iterates over a sequence.
- `while`: Runs while a condition is true.
- `break`: Exits the loop.
- `continue`: Skips to the next iteration.
- `pass`: Does nothing (placeholder).

File Handling

- `open("file.txt", "r")`: Read file
- `open("file.txt", "w")`: Write file
- `open("file.txt", "a")`: Append file
- `file.read()`: Read file content
- `file.readline()`: Read one line
- `file.write("text")`: Write to file
- `file.close()`: Close file
- Use with `open("file.txt", "w") as f:` for safe handling

List Methods

- `append(x)`: Add element to end
- `insert(i, x)`: Insert at index i
- `remove(x)`: Remove first occurrence of x

Variables & Data Types:

- `int`: Whole numbers.
- `float`: Decimal numbers.
- `str`: Text values.
- `bool`: True or False values.
- `list`: Ordered, changeable collection.
- `tuple`: Ordered, unchangeable collection.
- `set`: Unordered, unique collection.
- `dict`: Key-value pair collection.

Functions

- `def`: Define a new function
- `return`: Return value from a function
- `lambda`: Create a anonymous function

Built-in Functions

- `len(x)`: Length of x
- `max(x)`: Maximum value in x
- `min(x)`: Minimum value in x
- `sum(x)`: Sum of elements
- `sorted(x)`: Sorted list
- `range(start, stop, step)`: Generate sequence
- `map(func, iterable)`: Apply function to elements
- `filter(func, iterable)`: Filter elements
- `zip(iter1, iter2)`: Combine iterables
- `pop(i)`: Remove element at index i
- `reverse()`: Reverse list order
- `sort()`: Sort list

FIGURE 1.11 – Python Cheat Sheet

Source : LinkedIn

Méthodologie générale de chaque séance

Le cours est un mélange subtil entre analyse de données en géographie, apprentissage de Python et connaissances statistiques. Il est **progressif**. Si vous loupez une séance sans la rattraper, vous prenez le risque de ne rien comprendre à la séance suivante.

Vous êtes autonome dans votre apprentissage. C'est l'unique manière d'acquérir un savoir-faire. Je ne répèterais pas en boucle ce qui a déjà été vu, surtout pour les absents.

Que faire en arrivant en cours ? (environ 10 min)

1. Allumer votre poste attribué
2. Aller sur `GitHub` télécharger l'exercice du jour dans le bon dossier de l'architecture
3. Ouvrir le `Docker Engine`
4. Se connecter à son compte `Docker`
5. Lancer Python avec la commande `docker-compose up -d`
6. Poser des questions si des points dans le topo n'ont pas été compris

Que faire pendant le cours ? (environ 1h45min)

- Faire l'exercice du jour en n'oubliant pas de rédiger le rendu de la séance
- Poser des questions à l'enseignant sans le monopoliser (merci de penser à vos camarades)
- Même si le rendu est individuel, ne pas hésiter à travailler en groupe, à vous poser des questions entre vous, la séance, sans être un bazar total, doit être animée pour rendre les statistiques joyeuses

Que faire en quittant en cours ? (environ 5 min)

1. Éteindre le `Docker Compose` avec la commande `docker-compose down`
2. Se déconnecter de son compte `Docker`
3. Éteindre le `Docker Engine`
4. Éteindre votre poste attribué

Que faire entre les séances de cours ?

1. Lire la documentation du cours suivant
2. Faire les exercices de compréhension
3. Ne pas hésiter à se documenter
4. Ne pas hésiter à contacter, dans la limite du raisonnable, l'enseignant pour qu'il puisse vous aider en cas de blocage dans votre apprentissage

5. Prendre en main le langage de programmation `Python` en codant régulièrement de petits programmes en fonction de vos besoins pour vos mémoires

Séance 2

Principes généraux de la statistique

C'est la séance qui vous demandera le plus d'investissement.

Objectifs de la séance

- Comprendre la différence entre statistique et statistiques
- Comprendre ce qu'est le hasard
- Comprendre d'où viennent les données en géographie
- Comprendre la notion de probabilités et son lien avec les statistiques
- Connaître les types de statistique
- Bien apprendre ou réviser le vocabulaire statistique
- Bien distinguer les deux grands types de variable
- Comprendre ce qu'est une amplitude, une densité, une classe, un effectif et une fréquence

Ressources GitHub à lire avant la séance

Afin de limiter la consommation de papier, il faut vous rendre à la page suivante :
<https://github.com/MaximeForrieux/Sorbonne-M1-Analyse-de-donnees/Seance-02>.

Exercice du jour

Il s'agit d'apprendre à utiliser les bibliothèques `NumPy` et `Pandas` afin de calculer : une amplitude, une densité, une classe, un effectif et une fréquence.



Data Cleaning in Python (Pandas)

`import pandas as pd`

Step 1: Load the Data

```
df = pd.read_csv('your_file.csv') # For CSV files
df = pd.read_excel('your_file.xlsx') # For Excel files
df = pd.read_json('your_file.json') # For JSON files
```

Step 3: Check Missing Values (Numerically + Visually)

```
df.isnull().sum() # Count of m. values in each column
import missingno as msno # This shows a bar chart of
missing values. msno.bar(df)
```

Step 5: Remove Duplicates

```
df.drop_duplicates(inplace=True)
# Delete duplicate rows
```

Step 7: Strip Whitespaces from Column N.

```
df.columns = df.columns.str.strip()
# Remove leading/trailing spaces in column names
```

Step 9: Fix Inconsistent Labels (Spelling or Format Issues)

```
df['Payment_method'] =
df['Payment_method'].replace({
    'cc': 'Credit Card',
    'cash': 'Cash'})
```

Step 11: Drop Unnecessary Columns

```
df.drop(columns=['Unnecessary_column'],
inplace=True)
# Remove columns you don't need
```

Step 13: Split Date into Year and Month

```
df['year'] = df['date'].dt.year # Extract year
df['month'] = df['date'].dt.month # Extract month
```

Step 2: Understand the Dataset

```
df.info() # Show column, n, non-null count & data.type
df.describe() # Summary statistics for numerical columns
df.head() # Displays the first 5 rows of the dataset
df.shape # Shows the number of (rows, cols)
df.columns # Lists all column names
```

Step 4: Handle Missing Values

```
df['col'] = df['col'].fillna(df['col'].mean())
# Fill missing values with the mean, median, mode.
df.dropna(inplace=True)
# Remove rows with missing values.
```

Step 6: Fix Data Types

```
df['date'] = pd.to_datetime(df['date'])
# Convert to datetime format
df['amount'] = df['amount'].astype(float)
# Convert to float
```

Step 8: Rename Columns for Consistency

```
df.columns = df.columns.str.lower().str.replace(' ', '_')
# Make all column names simple
'Customer Name' becomes 'Customer_name'
df.rename(columns={'Cust_id': 'Customer_id'},
inplace=True) # Rename Specific Columns Only
```

Step 10: Remove Outliers

```
Q1 = df['amount'].quantile(0.25)
Q3 = df['amount'].quantile(0.75)
IQR = Q3 - Q1
df = df[(df['amount'] >= Q1 - 1.5 * IQR) & (df['amount'] <=
Q3 + 1.5 * IQR)]
If most sales are around ₹500, and one row says ₹50,000
```

Step 12: Create New Columns

```
df['Total_price'] = df['quantity'] * df['Unit_price']
# Add a new useful column
```

Step 14: Save the Cleaned Data

```
df.to_csv('cleaned_data.csv', index=False) # Save to a new CSV file without row numbers
from google.colab import files # Download the cleaned CSV file to your computer
files.download('cleaned_data.csv')
```



Pranav Borge



IF YOU FOUND THIS USEFUL, FOLLOW ME



FIGURE 2.1 – Python : Pandas, les commandes de base

Source : LinkedIn

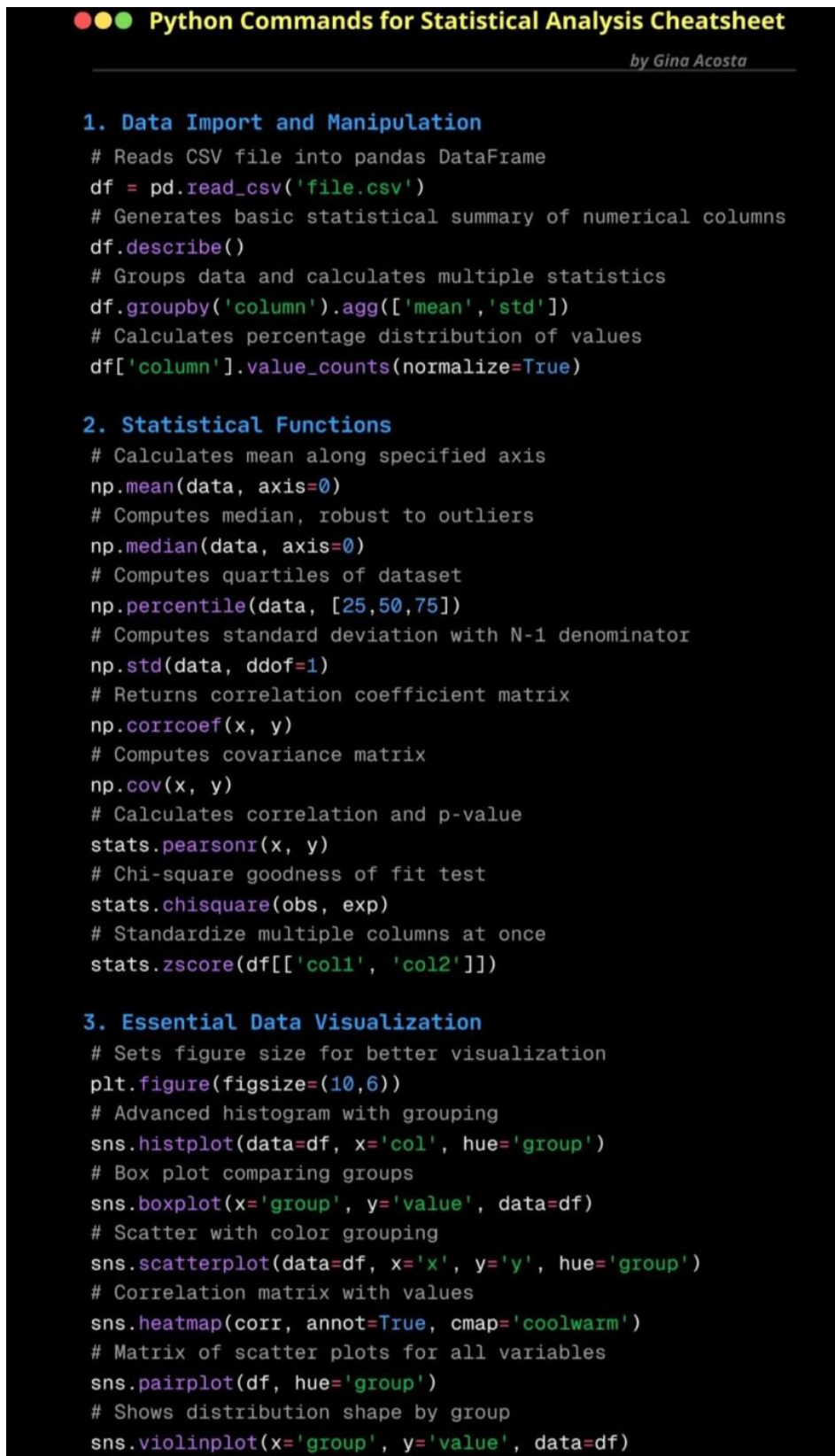


FIGURE 2.2 – Python : Commandes statistiques de base

Source : LinkedIn

9 Must-Know Python-Pandas Operations for Working with Data

By Brij Kishore Pandey

Data Import

```
pd.read_csv('file.csv')
pd.read_excel('file.xlsx', sheet_name='Sheet1')
pd.read_sql(query, connection)
pd.read_json('file.json')
pd.read_parquet('file.parquet')
```

Data Selection

```
df['column'] # Single column
df.loc['row', 'col'] # Label based
df.iloc[0:5, 0:2] # Integer based
df.query('col > 5') # SQL-like filtering
df[df['col'].isin(['A', 'B'])] # Multiple values
```

Data Manipulation

```
df.groupby('col').agg({'col2': ['mean', 'sum']})
df.merge(df2, on='key', how='left')
df.pivot_table(values='val', index='idx')
df.sort_values(['col1', 'col2'])
df.melt(id_vars='id', value_vars=['A', 'B'])
df.apply(lambda x: x*2) # Apply function
```

Statistics

```
df.describe() # Summary statistics
df['col'].agg(['mean', 'median', 'std'])
df['col'].value_counts(normalize=True)
df.corr(method='pearson')
df.cov() # Covariance matrix
df.quantile([0.25, 0.5, 0.75])
```

Data Cleaning

```
df.dropna(subset=['col'], how='any')
df.fillna(method='ffill') # Forward fill
df.drop_duplicates(subset=['col'])
df['col'].replace({'old': 'new'})
df['col'].astype('category')
df.interpolate(method='linear')
```

Time Series

```
df.resample('M').mean() # Monthly average
df.rolling(window=7).mean()
df.shift(periods=1) # Shift values
pd.date_range('2024', periods=12, freq='M')
df.asfreq('D', method='ffill')
df['date'].dt.strftime('%Y-%m-%d')
```

String Operations

```
df['col'].str.contains('pattern')
df['col'].str.extract('(\d+)')
df['col'].str.split('_').str[0]
df['col'].str.lower()
df['col'].str.strip()
df['col'].str.replace(r'\s+', ' ')
```

Advanced Features

```
df.pipe(func) # Method chaining
pd.eval('df1 + df2') # Expression eval
df.memory_usage(deep=True)
df.select_dtypes(include=['number'])
df.nlargest(5, 'col') # Top N values
df.explode('col') # Expand list column
```

Data Export

```
df.to_csv('output.csv', index=False)
df.to_excel('output.xlsx', sheet_name='Sheet1')
df.to_parquet('output.parquet')
df.to_json('output.json', orient='records')
```

Tips & Best Practices

- Use `.copy()` when creating DataFrame views
- Chain operations with method chaining
- Set `dtype='category'` for categorical columns
- Use `inplace=True` carefully, prefer reassignment

FIGURE 2.3 – Python : Pandas, la gestion des fichiers

Source : LinkedIn

Data Storytelling

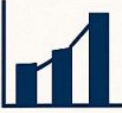






Chart	Types of Chart	When to Use it	Example of Use Case
	Bar Chart	Compare sales across categories	Displays growth of different products
	Line Chart	Show trends over time	Display the growth of website traffic over a budget
	Pie Chart	Highlight proportions percentages	Display the growth of expenses in budget
	Scatter plot	Represent relationships between variables	Identify correlations between marketing spend and ROI
	Histogram	Visualize the distribution of data	Analyze aggregate distribution of survey respondents
	Map	Visualize geospatial data	Display regional sales performance on a map
	Heatmap	Visualize data density and patterns in large datasets	Identify hotspots of customer activity in a shopping mall

FIGURE 2.4 – Les principales manières de visualiser les données avec Python

Source : LinkedIn

Deuxième partie

Statistique univariée

Séance 3

Paramètres statistiques élémentaires

Objectifs de la séance

- Connaître et comprendre les paramètres de position : espérance, moyenne, médiane, mode, médiale
- Connaître et comprendre les paramètres de dispersion : écart type, variance, coefficient de variation, étendue, écart interquantile, écart moyen, boîte de dispersion
- Connaître et comprendre les paramètres de forme : moments, coefficients β_1 et β_2 de Pearson et de Fisher

Ressources GitHub à lire avant la séance

<https://github.com/MaximeForriez/Sorbonne-M1-Analyse-de-donnees/Seance-03>

Exercice du jour

Approfondissement

- Connaître les propriétés de l'espérance, de la variance et de l'écart type
- Apprendre ou réviser les notions élémentaires sur les probabilités :
 - le dénombrement ;
 - les axiomes de Kolmogorov ;
 - les variables aléatoires ;
 - les lois de probabilité.

Séance 4

Distributions statistiques

Objectifs de la séance

- Connaître, et surtout reconnaître, les distributions statistiques des variables discrètes
- Connaître, et surtout reconnaître, les distributions statistiques des variables continues
- Bien comprendre le lien entre probabilités et distributions statistiques

Ressources GitHub à lire avant la séance

<https://github.com/MaximeForriez/Sorbonne-M1-Analyse-de-donnees/Seance-04>

Exercice du jour

Séance 5

Statistiques inférentielles

Objectifs de la séance

- Comprendre la notion d'échantillonnage et ses principales méthodes
- Comprendre les notions d'estimateur et d'estimation
- Comprendre la notion de statistique exhaustive
- Comprendre comment choisir un estimateur
- Comprendre comment estimer un paramètre par intervalle de confiance
- Comprendre les différentes méthodes d'estimation d'un paramètre
- Comprendre ce qu'est un test statistique

Ressources GitHub à lire avant la séance

<https://github.com/MaximeForriez/Sorbonne-M1-Analyse-de-donnees/Seance-05>

Exercice du jour

Séance 6

Statistique d'ordre des variables qualitatives

Objectifs de la séance

- Savoir définir une statistique d'ordre
- Comprendre ce qu'est la corrélation des rangs

Ressources GitHub à lire avant la séance

<https://github.com/MaximeForriez/Sorbonne-M1-Analyse-de-donnees/Seance-06>

Exercice du jour

Troisième partie

Statistique bivariée

Séance 7

Relations entre deux variables

Objectifs de la séance

- Comprendre comment choisir deux variables quantitatives
- Comprendre ce qu'est l'indépendance statistique
- Comprendre à quoi sert la covariance
- Comprendre la différence de traitement entre variables quantitatives et qualitatives

Ressources GitHub à lire avant la séance

<https://github.com/MaximeForrieux/Sorbonne-M1-Analyse-de-donnees/Seance-07>

Exercice du jour

Séance 8

Régression et corrélation statistique

Objectifs de la séance

- Comprendre la méthode des moindres carrées
- Comprendre ce qu'est une corrélation simple
- Comprendre ce qu'est une régression linéaire
- Comprendre ce qu'est un intervalle de confiance pour une droite de régression

Ressources GitHub à lire avant la séance

<https://github.com/MaximeForriez/Sorbonne-M1-Analyse-de-donnees/Seance-08>

Exercice du jour

Approfondissement

- Analyse de la variance à simple entrée
- Analyse factorielle de correspondance (A.F.C.)

Quatrième partie

Statistique multivariée

Séance 9

Analyses multivariées : les méthodes descriptives

Objectifs de la séance

- Comprendre ce qu'est une analyse factorielle en composantes principales (A.C.P.)
- Comprendre ce qu'est une analyse factorielle des correspondances multiples (A.C.M.)
- Comprendre ce qu'est une classification ascendante hiérarchique (C.A.H.)
- Comprendre ce qu'est une analyse factorielle discriminante (A.F.D.)
- Comprendre ce qu'est une analyse factorielle multiple (A.F.M.)

Ressources GitHub à lire avant la séance

<https://github.com/MaximeForriez/Sorbonne-M1-Analyse-de-donnees/Seance-09>

Exercice du jour

Approfondissement

- Comprendre l'analyse canonique

Séance 10

Analyses multivariées : les méthodes explicatives

Objectifs de la séance

- Comprendre l'usage des méthodes de régression multivariées et le sens des corrélations multiple et partielle

Ressources GitHub à lire avant la séance

<https://github.com/MaximeForrieux/Sorbonne-M1-Analyse-de-donnees/Seance-10>

Exercice du jour

Approfondissement

- Comprendre l'usage de l'analyse de la variance à double entrée
- Comprendre l'analyse de la variance orthogonale à entrées multiples

Bibliographie

- [Abdessemed et Escofier, 1996] ABDESSEMED, L. et ESCOFIER, B. (1996). Analyse factorielle multiple de tableaux de fréquences. comparaison avec l'analyse canonique des correspondances. Journal de la société statistique de Paris, 137(2):3–18.
- [Abitoul et Hachez-Leroy, 2015] ABITOUL, S. et HACHEZ-LEROY, F. (2015). Humanités numériques, pages 43–57.
- [Bardos, 2001] BARDOS, M. (2001). Analyse discriminante. Application au risque et scoring financier. Dunod, Paris.
- [Bécue-Bertaut et Pagès, 2001] BÉCUE-BERTAUT, M. et PAGÈS, J. (2001). Analyse simultanée de questions ouvertes et de questions fermées. méthodologie, exemple. Journal de la société française de statistique, 142(4):91–104.
- [Béguin, 1979] BÉGUIN, H. (1979). Méthodes d'analyse géographique quantitative. Litec, Paris.
- [Béguin et Pumain, 1994a] BÉGUIN, M. et PUMAIN, D. (1994a). La représentation des données géographiques. Statistique et cartographie. Cours. Armand Colin, Paris. réédition de 2007.
- [Béguin et Pumain, 1994b] BÉGUIN, M. et PUMAIN, D. (1994b). La représentation des données géographiques. Statistique et cartographie. Cours. Armand Colin, Paris.
- [Benali et Escofier, 1987] BENALI, H. et ESCOFIER, B. (1987). Stabilité de l'analyse factorielle des correspondances multiples en cas de données manquantes et de modalités à faibles effectifs. Revue de statistique appliquée, 35(1):41–51.
- [Benali et Escofier, 1990] BENALI, H. et ESCOFIER, B. (1990). Analyse factorielle lissée et analyse factorielle des différences locales. Revue de statistique appliquée, 38(2):55–76.
- [Benammou et al., 2007] BENAMMOU, S., SAPORTA, G. et SOUISSI, B. (2007). Une procédure de réduction du nombre de paires en analyse conjointe. Journal de la société française de statistique, 148(4).
- [Benzécri, 1960] BENZÉCRI, J.-P. (1960). Sur les variétés localement affines et localement projectives. Bulletin de la société mathématique de France, 88:229–332.
- [Benzécri, 1973a] BENZÉCRI, J.-P. (1973a). L'analyse de données, t. 1, La taxinomie. Dunod, Paris.
- [Benzécri, 1973b] BENZÉCRI, J.-P. (1973b). L'analyse de données, t. 2, L'analyse de correspondances. Dunod, Paris.

- [Benzécri, 1980] BENZÉCRI, J.-P. (1980). Pratique de l'analyse des données. Dunod, Paris.
- [Benzécri, 1982a] BENZÉCRI, J.-P. (1982a). L'analyse des données. Leçons sur l'analyse factorielle et la reconnaissance des formes et travaux. Dunod, Paris.
- [Benzécri, 1982b] BENZÉCRI, J.-P. (1982b). Sur la généralisation du tableau de burt et son analyse par bandes. Les cahiers de l'analyse des données, 7(1):33–43.
- [Berry, 2011] BERRY, D. M. (2011). The computational turn : Thinking about the digital humanities. Culture Machine, 12:1–22.
- [Bobee, 1978] BOBEE, B. (1978). Éléments de statistiques. Université du Québec, Montreal.
- [Bonnet et al., 1996] BONNET, P., LE ROUX, B. et LEMAINÉ, G. (1996). Analyse géométrique des données : une enquête sur le racisme. Mathématiques et sciences humaines, (136):5–28.
- [Bouroche et Saporta, 2002] BOUROCHE, J.-M. et SAPORTA, G. (2002). L'analyse de données. Que sais-je ? P.U.F., Paris.
- [Bouroche et Tenenhaus, 1970] BOUROCHE, J.-M. et TENENHAUS, M. (1970). Quelques méthodes de segmentation. Revue française d'informatique et de recherche opérationnelle, (4):29–42.
- [Brown, 1962] BROWN, R. G. (1962). Smoothing forecasting and prediction of discrete time series. Quantitative Methods Series. Prentice-Hall, London.
- [Celeux et Nakache, 1994] CELEUX, G. et NAKACHE, J.-P. (1994). Analyse discriminante sur variables qualitatives. Polytechnica, Paris.
- [Chadule, 1997] CHADULE, G. (1997). Initiation aux pratiques statistiques en géographie. Masson, Paris.
- [Charre, 1995] CHARRE, J. (1995). Statistique et territoire. Espaces modes d'emploi. G.I.P. RECLUS, Montpellier.
- [Cibois, 1991] CIBOIS, P. (1991). L'analyse factorielle. Analyse en composantes principales et analyse factorielle des correspondances. Que sais-je ? PUF, Paris.
- [Cibois, 2000] CIBOIS, P. (2000). L'analyse factorielle. Analyse en composantes principales et analyse factorielle des correspondances. Que sais-je ? P.U.F., Paris.
- [Cicéri et al., 1977] CICÉRI, M.-F., MARCHAND, B. et RIMBERT, S. (1977). Introduction à l'analyse spatiale. U. Armand Colin, Paris.
- [Cloux et al., 2016] CLOUX, P.-Y., GARLOT, T. et KOHLER, J. (2016). Docker. Pratique des architectures à base de conteneurs. Études, développement & intégration. Dunod, Paris.
- [Cloux et al., 2022] CLOUX, P.-Y., GARLOT, T. et KOHLER, J. (2022). Docker et conteneurs. Architectures, développement, usages et outils. Études, développement, intégration. Dunod, Malakoff. 3e édition.
- [Couty et al., 2007] COUTY, F., DEBORD, J. et FREDON, D. (2007). Mini-manuel de probabilités et statistique. Dunod, Paris.
- [Dacos, 2011] DACOS, M. (2011). Manifeste des Digital Humanities.

- [Dacos et Mounier, 2015] DACOS, M. et MOUNIER, P. (2015). Humanité numérique. État des lieux et positionnement de la recherche française dans le contexte international.
- [Dagnelie, 1975] DAGNELIE, P. (1975). Analyse statistique à plusieurs variables. Les presses agronomiques de Gembloux, Paris.
- [Deheuvels, 2021] DEHEUVELS, P. (2021). La probabilité, le hasard et la certitude. Que sais-je ? PUF, Paris.
- [Delsart et Vaneecloo, 2010] DELSART, V. et VANEECLOO, N. (2010). Probabilités, variables aléatoires, lois classiques, tome 1. Guides pratiques. Presses universitaires du Septentrion, Villeneuve-d'Ascq.
- [Dervin, 1992] DERVIN, C. (1992). Comment interpréter les résultats d'une analyse factorielle des correspondances ? STAT - ITCF. Institut technique des céréales et des fourrages (I.C.T.F.), Paris.
- [Diday, 1971] DIDAY, E. (1971). Une nouvelle méthode en classification automatique et reconnaissance des formes la méthode des nuées dynamiques. Revue de statistique appliquée, 19(2):19–33.
- [Dozo, 2011] DOZO, B.-O. (2011). Mesures de l'écrivain. Profil socio-littérature et capital relationnel dans l'entre-deux-guerres en Belgique francophone. Presses universitaires de Liège, Liège.
- [Dumolard, 2011] DUMOLARD, P. (2011). Données géographiques. Analyse statistique multivariée. Lavoisier - Hermès, Paris.
- [Dumolard et al.,] DUMOLARD, P., ALLIGNOL, F., PAUL, E. et QUESSEVEUR, E. L'outil informatique en géographie. En ligne : [https ://dspace.msh-alpes.prd.fr/bitstream/1801/19/2/book_geoinfo.pdf](https://dspace.msh-alpes.prd.fr/bitstream/1801/19/2/book_geoinfo.pdf).
- [Dumolard et al., 2003] DUMOLARD, P., DUBUS, N. et CHARLEUX, L. (2003). Les statistiques en géographie. Atout géographie. Belin, Paris.
- [Durand, 1998] DURAND, J.-L. (1998). Taux de dispersion des valeurs propres en a.c.p., a.c. et a.c.m. Mathématiques et sciences humaines, (144):15–28.
- [Escofier, 1979] ESCOFIER, B. (1979). Une représentation des variables dans l'analyse des correspondances multiples. Revue de statistique appliquée, 27(4):37–47.
- [Escofier, 1981] ESCOFIER, B. (1981). Quelques indices pour comparer des tableaux de contingence. Statistique et analyse de données, 6(1):39–51.
- [Escofier, 1984] ESCOFIER, B. (1984). Analyse factorielle en référence à un modèle. application à l'analyse de tableaux d'échanges. Revue de statistique appliquée, 32(4):25–36.
- [Escofier et al., 1990] ESCOFIER, B., BENALI, H. et BACHAR, K. (1990). Comment introduire la contiguïté en analyse des correspondances ? application en segmentation d'image. Statistique et analyse de données, 15(3):61–92.
- [Escofier et Pagès, 2016a] ESCOFIER, B. et PAGÈS, J. (2016a). Analyses factorielles simples et multiples. Cours et études de cas. Sciences sup. Dunod, Paris.
- [Escofier et Pagès, 2016b] ESCOFIER, B. et PAGÈS, J. (2016b). Analyses factorielles simples et multiples. Cours et études de cas. Sciences sup. Dunod, Paris.

- [Escofier-Cordier, 1969] ESCOFIER-CORDIER, B. (1969). L'analyse factorielle des correspondances. Cahiers du Bureau universitaire de recherche opérationnelle. Série Recherche, (13):25–59.
- [Flechy, 2001] FLECHY, P. (2001). L'imaginaire d'internet. La Découverte, Paris.
- [Fleurant et Fleurant, 2015] FLEURANT, S. et FLEURANT, C. (2015). Bases de mathématiques pour la géologie et la géographie. Cours. Dunod, Paris.
- [Foucart, 1984] FOUCART, T. (1984). Analyse factorielle de tableaux multiples. Méthode + Programmes. Masson, Paris.
- [Foucart, 1997] FOUCART, T. (1997). L'analyse de données. Mode d'emploi. Méthodes et études de cas. Didact Statistique. Presses universitaires de Rennes, Rennes.
- [Gassara, 2022] GASSARA, E. (2022). Docker / Kubernetes. Pour optimiser et accélérer les développements d'applications conteneurisées. Eyrolles, Paris.
- [Girschig, 1969] GIRSCHIG, M. (1969). Analyse statistique.
- [Gold, 2012] GOLD, M., éditeur (2012). Debates in the Digit Humanities, pages 1–532. University of Minnesota Press, Minneapolis.
- [Goldfarb et Pardoux, 2011] GOLDFARB, B. et PARDOUX, C. (2011). Introduction à la méthode statistique. Manuel et exercices corrigés. Éco sup. Dunod, Paris.
- [Gouigoux, 2015] GOUIGOUX, J.-P. (2015). Docker. Prise en main et mise en pratique sur une architecture micro-services. Collection Epsilon. Éditions ENI, Paris.
- [Grandjacquot, 1999] GRANDJACQUOT, M.-P. (1999). Outils statistiques. Théorie et pratique du management - Méthodes quantitatives. Eska, Paris.
- [Guillaud, 2010] GUILLAUD, H. (2010). Qu'apportent les digital humanities ? La Feuille.
- [Hotelling, 1936] HOTELLING, H. (1936). Relation between two sets of variates. Biometrika, 28(3/4):321–377.
- [Howell, 2024] HOWELL, D. C. (2024). Méthodes statistiques en sciences humaines. Ouvertures psychologiques. DeBoeck Supérieur, Louvain-la-Neuve. 3e édition corrigée.
- [Husson et Pagès, 2006] HUSSON, F. et PAGÈS, J. (2006). Aspects méthodologiques du modèle indscal. Revue de statistique appliquée, 54(2):83–100.
- [Jacquard, 2000] JACQUARD, A. (2000). Les probabilités. Que sais-je ? n°1571. PUF, Paris. réédition de 1974.
- [Jelensperger et Moreau, 1997] JELENSPERGER, C. et MOREAU, Y. (1997). Nouveau mémento de mathématiques. Algèbre. Analyse. Probabilités. Statistiques. Informatique. Série "Usuels". Vuibert, Paris.
- [Josse et al., 2009] JOSSE, J., HUSSON, F. et PAGÈS, J. (2009). Gestion des données manquantes en analyse en composantes principales. Journal de la société française de statistique, 150(2):28–51.
- [Kohonen, 1982] KOHONEN, T. (1982). Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43:59–69.

- [Lahousse et Piedanna, 1998a] LAHOUSSE, P. et PIEDANNA, V. (1998a). L'outil statistique en géographie, t. 1, Les distributions spatiales. Synthèse 43. Armand Colin, Paris.
- [Lahousse et Piedanna, 1998b] LAHOUSSE, P. et PIEDANNA, V. (1998b). L'outil statistique en géographie, t. 2, L'analyse bivariable. Synthèse 96. Armand Colin, Paris.
- [Le Dien et Pagès, 2003] LE DIEN, S. et PAGÈS, J. (2003). Analyse factorielle multiple hiérarchique. Revue de statistique appliquée, 51(2):47–73.
- [Le Roux, 1991] LE ROUX, B. (1991). Sur la construction d'un protocole additif. Mathématiques et sciences humaines, (114):57–62.
- [Le Roux, 1998] LE ROUX, B. (1998). Inférence combinatoire en analyse géométrique des données. Mathématiques et sciences humaines, (144):5–14.
- [Le Roux, 1999] LE ROUX, B. (1999). Analyse spécifique d'un nuage euclidien : application à l'étude des questionnaires. Mathématiques et sciences humaines, (146):65–83.
- [Le Roux, 2014] LE ROUX, B. (2014). Analyse géométrique des données multidimensionnelles. Dunod, Paris.
- [Le Roux et Rouanet, 1984] LE ROUX, B. et ROUANET, H. (1984). L'analyse multidimensionnelle des données structurées. Mathématiques et sciences humaines, (85):5–18.
- [Lebart et al., 2006] LEBART, L., PIRON, M. et MORINEAU, A. (2006). Statistique exploratoire multidimensionnelle. Visualisation et inférence en fouilles de données. Sciences sup. Dunod, Paris. 4e édition augmentée.
- [Lecoutre, 2008] LECOUTRE, J.-P. (2008). Statistique et probabilités. T.D. Dunod, Paris.
- [Lerman, 1966] LERMAN, I. C. (1966). Essai sur l'analyse hiérarchique. Mathématiques et sciences humaines, (17):37–46.
- [Lerman, 1970] LERMAN, I. C. (1970). Sur l'analyse des données préalable à une classification automatique (proposition d'une nouvelle mesure de similarité). Mathématiques et sciences humaines, (32):5–15.
- [Lerman, 1973] LERMAN, I. C. (1973). Introduction à une méthode de classification automatique illustrée par la recherche d'une typologie des personnages enfants à travers la littérature enfantine. Revue de statistique appliquée, 21(3):23–49.
- [Lerman, 1981] LERMAN, I. C. (1981). La classification : concepts et caractéristiques d'une méthodologie d'analyse des données. Journal de la société statistique de Paris, 122(2):70–90.
- [Lerman, 1982] LERMAN, I. C. (1982). Programmes d'analyse des résultats d'une classification automatique.
- [Lerman, 2008] LERMAN, I. C. (2008). Analyse logique, combinatoire et statistique de la construction d'une hiérarchie binaire implicative. niveaux et nœuds significatifs. Mathematics and Social Sciences, (184):47–107.
- [Lethielleux, 2010] LETHIELLEUX, M. (2010). Statistique descriptive en 27 fiches. Express sup. Dunod, Paris.
- [Lipschutz, 1987] LIPSCHUTZ, S. (1987). Probabilités. Série Schaum. McGraw-Hill, Paris.

- [Marchand, 1972] MARCHAND, B. (1972). L'usage des statistiques en géographie. L'espace géographique, 1(2):79–100.
- [Martin, 2020] MARTIN, O. (2020). L'analyse quantitative des données. 128. Armand Colin, Paris.
- [McCarty, 2002] MCCARTY, W. (2002). Humanities computing : Essential problems, experimental practice. Literary and Linguistic Computing, 17(1):103–125.
- [McCarty, 2005] MCCARTY, W. (2005). Humanities Computing. Palgrave-Macmillan, London.
- [Monino et al., 2010] MONINO, J.-L., KOSIANSKI, J.-M. et LE CORNU, F. (2010). Statistique descriptive. Dunod, Paris.
- [Morand et Pagès, 2007] MORAND, E. et PAGÈS, J. (2007). Analyse factorielle multiple procustéenne. Journal de la société française de statistique, 148(2):65–97.
- [Moreau et al., 2000] MOREAU, J., DOUDIN, P.-A. et CAZES, P. (2000). L'analyse des correspondances et les techniques connexes. Approches nouvelles pour l'analyse statistique des données. Mathématiques & applications n°32. Springer, Paris.
- [Morgenthaler, 2007] MORGENTHALER, S. (2007). Introduction à la statistique. Enseignement des mathématiques. Presses polytechniques et universitaires romandes, Lausanne. 3e édition augmentée.
- [Nakache et Confais, 2000] NAKACHE, J.-P. et CONFAIS, J. (2000). Méthodes de classification avec illustrations SPAD et SAS. Centre international de statistique et d'informatique appliquées (C.I.S.I.A.), Montreuil.
- [Pagès et Périnel, 2007] PAGÈS, J. et PÉRINEL, E. (2007). Blocs incomplets équilibrés versus plans optimaux. Journal de la société française de statistique, 148(2):100–112.
- [Pagès et Tenenhaus, 2002] PAGÈS, J. et TENENHAUS, M. (2002). Analyse factorielle multiple et approche p.l.s. Revue de statistique appliquée, 50(1):5–33.
- [Paquienséguy et Pélissier, 2021] PAQUIENSÉGUY, F. et PÉLISSIER, N., éditeurs (2021). Questionner les humanités numériques, pages 1–298. Société française des sciences de l'information et de la communication & Conférence permanente des directeurs de laboratoires en sciences de l'information et de la communication, Paris.
- [Petit, 2018] PETIT, T. (2018). Tous les algorithmes. Programmation. Python. Prépa scientifique 1ère et 2e années. Ellipses, Paris.
- [Philippeau, 1992] PHILIPPEAU, G. (1992). Comment interpréter les résultats d'une analyse en composantes principales ? STAT - ITCF. Institut technique des céréales et des fourrages (I.C.T.F.), Paris.
- [Poinsot, 2004] POINSOT, D. (2004). Statistiques pour statophobes. Une introduction au monde des tests statistiques à l'intention des étudiants qui n'y entravent que pouic et qui détestent les maths par-dessus le marché. <https://perso.univ-rennes1.fr/denis.poinsot>.
- [Pontier et Normand, 1992] PONTIER, J. et NORMAND, M. (1992). À propos de généralisation de l'analyse canonique. Revue de statistique appliquée, 40(1):57–75.

- [Preda et Saporta, 2002] PREDA, C. et SAPORTA, G. (2002). Régression p.l.s. sur un processus stochastique. Revue de statistique appliquée, 50(2):27–45.
- [Randriamihamison, 2021] RANDRIAMIHAMISON, N. (2021). Classification ascendante hiérarchique sous contrainte de contiguïté pour l'analyse de données Hi-C. Thèse de doctorat, Université Paul Sabatier - Toulouse III, Toulouse.
- [Rateau, 2001] RATEAU, P. (2001). Méthode et statistiques expérimentales en sciences humaines. Université. Ellipses, Paris.
- [Réau et Chauvat, 1988] RÉAU, J.-P. et CHAUVAT, G. (1988). Probabilités & statistiques. Résumé des cours. Exercices et problèmes corrigés. Q.C.M. Cursus. Armand Colin, Paris.
- [Rodriguez Herrera et Salles-Le Gac, 2002] RODRIGUEZ HERRERA, R. et SALLES-LE GAC, D. (2002). Initiation à l'analyse factorielle des données. Fondements mathématiques et interprétations. Ellipses, Paris.
- [Rouanet et Le Roux, 1993] ROUANET, H. et LE ROUX, B. (1993). Analyse des données multidimensionnelles. Statistiques en sciences humaines. Dunod, Paris.
- [Rouanet et Le Roux, 1997] ROUANET, H. et LE ROUX, B. (1997). Analyse des données multidimensionnelles. Dunod, Paris.
- [Rousselet, 2020] ROUSSELET, M. (2020). Python. Introduction au calcul numérique. Ressources informatiques. Éditions ENI, Paris.
- [Sanders, 1989] SANDERS, L. (1989). L'analyse des données appliquées à la géographie. Alidade. GIP-Reclus, Montpellier.
- [Saporta, 1975a] SAPORTA, G. (1975a). Dépendance et codages de deux variables aléatoires. Revue de statistique appliquée, 23(1):43–63.
- [Saporta, 1975b] SAPORTA, G. (1975b). Liaisons entre plusieurs ensembles de variables et codage de données qualitatives. Thèse de doctorat, Université Pierre et Marie Curie - Paris VI, Paris.
- [Saporta, 1976] SAPORTA, G. (1976). Quelques applications des opérateurs d'Escoufier au traitement des variables qualitatives. Quelques applications des opérateurs d'Escoufier au traitement des variables qualitatives, 1(1):38–46.
- [Saporta, 1979] SAPORTA, G. (1979). Ponderation optimale de variables qualitatives en analyse des données. pages 19–31.
- [Saporta, 1981a] SAPORTA, G. (1981a). Méthodes exploratoires d'analyse de données temporelles. Thèse de doctorat, Université Pierre et Marie Curie - Paris VI, Paris.
- [Saporta, 1981b] SAPORTA, G. (1981b). Méthodes exploratoires d'analyse de données temporelles. Cahiers du bureau universitaire de recherche opérationnelle, (37-38):7–194.
- [Saporta, 2001] SAPORTA, G. (2001). Discussions et commentaires. data mining et statistique. Journal de la société française de statistique, 142(1):81–84.
- [Scheid, 1986] SCHEID, F. (1986). Analyse numérique. Série Schaum. McGraw-Hill, Paris.
- [Schultz et Bussonnier, 2020] SCHULTZ, E. et BUSSONNIER, M. (2020). Python pour les SHS. Introduction à la programmation pour le traitement de données. Pratique de la statistique. Presses universitaires de Rennes, Rennes.

- [Spiegel, 2014] SPIEGEL, B. (2014). Digital Studies : organologie des services et technologies de la connaissance. FYP, Limoges.
- [Spiegel, 1974] SPIEGEL, M. R. (1974). Formules et tables de mathématiques. Série Schaum. McGraw-Hill, Paris.
- [Spiegel, 1984] SPIEGEL, M. R. (1984). Théorie et applications de la statistique. Série Schaum. McGraw-Hill, Paris. réédition de 1972.
- [Svensson, 2012] SVENSSON, P. (2012). Envisionning the digital humanities. Digital Humanities Quarterly, 6(1).
- [Tenenhaus, 1977] TENENHAUS, M. (1977). Analyse en composantes principales d'un ensemble de variables nominales ou numériques. Revue de statistique appliquée, 25(2):39–56.
- [Tenenhaus, 1979] TENENHAUS, M. (1979). La régression qualitative. Revue de statistique appliquée, 27(2):5–21.
- [Tenenhaus, 2007] TENENHAUS, M. (2007). Statistique. Méthodes pour décrire, expliquer et prévoir. Dunod, Paris.
- [Tenenhaus et Priouret, 1974] TENENHAUS, M. et PRIURET, B. (1974). Analyse des séries chronologiques multidimensionnelles. Revue française d'automatique, informatique, recherche opérationnelle. Recherche opérationnelle, 8(V2):5–16.
- [Terras, 2011] TERRAS, M. (2011). Peering inside the big tent : Digital humanities and the crisis of inclusion. Melissa Terras' Blog. <https://melissaterras.blogspot.com/2011/07/peering-inside-big-tent-digital.html>.
- [Thabut, 2018] THABUT, G. (2018). Les statistiques : pour quoi faire ?
- [Turner, 2012] TURNER, F. (2012). Aux sources de l'utopie numérique de la contre-culture à la cyberculture : Steward Brand, un homme d'influence. C. & F., Caen.
- [Van Hooland et al., 2016] VAN HOOLAND, S., GILLET, F., HENGCHEN, S. et DE WILDE, M. (2016). Introduction aux humanités numériques : méthodes et pratiques. Méthodes en sciences humaines. De Boeck, Louvain-la-Neuve.
- [Veyseyre, 2004] VEYSSEYRE, R. (2004). Aide-mémoire. Statistique et probabilités pour l'ingénieur. L'usine nouvelle. Dunod, Paris.
- [Vigneron, 1997] VIGNERON, E. (1997). Géographie et statistique. Que sais-je ? PUF, Paris.
- [Ward, 1963] WARD, J. H. (1963). Hierarchical grouping to optimize an objective function. Journal of American Statistical Association, 58:236–244.
- [Wonnacott et Wonnacott, 1995] WONNACOTT, T. H. et WONNACOTT, R. J. (1995). Statistique. Économie – Gestion – Sciences – Médecine. Économica, Paris.
- [Youness et Saporta, 2001] YOUNESS, G. et SAPORTA, G. (2001). Une méthodologie pour la comparaison de partitions. Revue de statistique appliquée, 52(1):97–120.