ADF PR
digital lifelong learning

Introduction à la Data Analytics

CERTIFICAT EXÉCUTIF - Analyse des données avec Excel avancé, SPSS, IA/Python & Power BI

Plan

- **Chapitre 1 : Comprendre l'analyse des données**
  - Qu'est-ce que l'analyse des données ?
  - Objectifs
  - Processus d'analyse des données
  - Les types d'analyse des données
- **Chapitre 2 : Concepts de base en analyse des données**
  - Types de variables
  - Mesures de tendance centrale
  - Mesures de dispersion
- **Chapitre 3 : Préparation des données**
  - Présentation générale
  - Nettoyage des données
  - Transformation des données
- **Chapitre 4 : Techniques d'exploration des données**
  - Analyse exploratoire des données (AED)
  - Analyse des relations entre les variables

Introduction à l'analyse des données

Chapitre 1

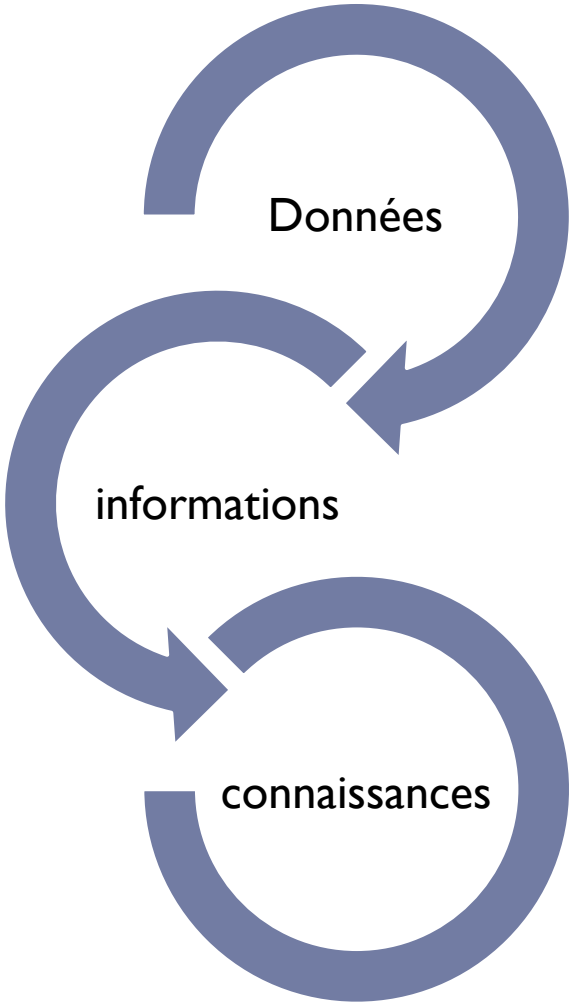
Problématique

Qu'est-ce qu'**une information** ?

Qu'est-ce qu'**une donnee** ?

Qu'est-ce qu'**une connaisance** ?

Définition de l'information



Définition de l'information

1er niveau : **la Donnée**

- Chaîne de caractère associé à des objets, des personnes ou des événements..
- Traitée par des personnes ou systèmes
- Représentée par un attribut et une valeur
- La donnée peut être brute ou calculée
- C'est la matière première de l'information

Exemples de données:

Nom : Joseph,

Date de naissance : 11/10/1950,

Taux de croissance de l'entreprise : 5%,

Référence machine : 1275DX, etc.

Définition de l'information

2ème niveau : l'Information

- De la donnée à l'information
- Une donnée est l'enregistrement d'une observation, objet, fait destiné à être interprété, traité par l'homme. La donnée est généralement objective Exemples :
  - température =35°
  - âge = 2 mois
- Une information est le signifiant attaché à la donnée ou à un ensemble de données par association. L'information est généralement subjective, définie selon un contexte Exemples :
  - (température=35°) : temps chaud
  - (âge=2 mois) : nourrisson
  - La donnée Age de l'employé est interprétée par la DRH devient une information qui sert à décider si une personne ouvre droit à la retraite ou pas

Définition de l'information

3ème niveau : **la Connaissance**

- De l'information à la connaissance
- Une connaissance est une information nouvelle, apprise par association d'informations de base, de règles, de raisonnement d'expérience, d'expertise, etc.

Exemple :

- temps chaud et enfant nourrisson alors risque de déshydratation

Qu'est-ce que l'analyse des données ?

- L’analyse des données est un ensemble de techniques pour découvrir la structure, éventuellement compliquée, d’un tableau de nombres à plusieurs dimensions et la traduire en une structure plus simple qui la résume au mieux. Cette structure peut, le plus souvent, être représentée graphiquement.

(J-P. Fénelon).

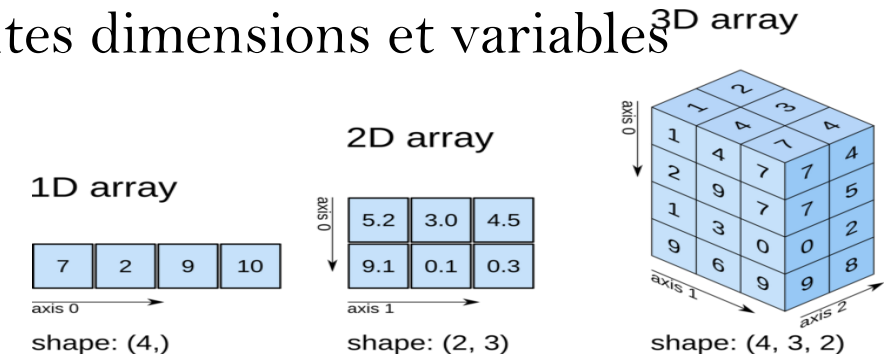
Jean-Paul Fénelon est un chercheur connu pour ses contributions à l'analyse des données, particulièrement dans les années 1980. Son ouvrage "Qu’est-ce que l’Analyse des Données?" publié en 1981 est une référence importante dans ce domaine.

Qu'est-ce que l'analyse des données ?

- L’analyse des données (ADD) est la science qui consiste à examiner les données pour en tirer des informations permettant de prendre des décisions ou d’approfondir les connaissances sur divers sujets.
- Elle consiste à soumettre les données à des opérations. Ce processus permet d’obtenir des conclusions précises qui nous aident à atteindre nos objectifs

Qu'est-ce que l'analyse des données ?

- L'analyse des données n'est pas une méthode unique mais un ensemble de méthodes et d'outils utilisés pour extraire des informations significatives à partir de données brutes.
- Ces techniques sont employées pour identifier les motifs, les relations et les tendances cachés dans les données.
- Les données peuvent être complexes, avec plusieurs dimensions et variables.
- Les données analysées sont souvent organisées sous forme de tableaux ou de matrices, avec des lignes et des colonnes représentant différentes dimensions et variables



Objectifs

- L'objectif de l'analyse est de simplifier cette complexité en une structure plus compréhensible.
  - Répondre aux problèmes posés par des tableaux de grandes dimensions
- Cette simplification permet de résumer les données de manière à ce qu'elles soient plus faciles à interpréter.
- Souvent, la structure simplifiée est représentée visuellement, par exemple à l'aide de graphiques, de diagrammes ou de tableaux de bord.
  - Organiser et visualiser les informations
- Les représentations graphiques permettent de visualiser les données de manière intuitive, facilitant ainsi la compréhension des informations clés.

Processus d'analyse des données

- La Data Analytics est un processus complexe qui se compose de plusieurs étapes clés, chacune étant essentielle pour transformer des données brutes en informations exploitables.
- La Data Analytics est le processus de **collecte**, **d'analyse** et de **visualisation** des données pour prendre des décisions éclairées.

Processus d'analyse des données

- Étape 1 : Définir les objectifs et les questions
- Étape 2 : Collecte de données
- Étape 3 : Nettoyage des données
- Étape 4 : Analyse des données
- Étape 5 : Interprétation et visualisation des données

Étape 1 : Définir les objectifs et les questions

- La première étape du processus d'analyse des données consiste à définir les objectifs et à formuler des questions claires et spécifiques auxquelles votre analyse vise à répondre.
- Il s'agit de comprendre le problème ou la situation en question, d'identifier les données nécessaires pour y répondre et de définir les mesures ou les indicateurs permettant de mesurer les résultats.
- **Objectif** : Identifier le besoin ou le problème à résoudre.
- **Exemple** : Une entreprise de e-commerce se demande : « Pourquoi le taux de conversion est-il plus faible en décembre que le reste de l'année ? »

Étape 2 : Collecte de données

- Une fois les objectifs et les questions définis, l'étape suivante consiste à collecter les données pertinentes. Cela peut se faire par le biais de différentes méthodes telles que les enquêtes, les entretiens, les observations ou l'extraction à partir de bases de données existantes.
- **Objectif** : Rassembler les données pertinentes pour répondre à la question.
- **Exemple** : Extraire les données de ventes, trafic web, provenance des visiteurs, paniers abandonnés depuis Google Analytics, CRM, ou base SQL.

Étape 3 : Nettoyage des données

- Le nettoyage des données est une étape essentielle du processus d'analyse des données. Il s'agit de vérifier si les données comportent des erreurs et des incohérences, et de les corriger ou de les supprimer. Cette étape garantit la qualité et la fiabilité des données, ce qui est essentiel pour obtenir des résultats précis et significatifs de l'analyse.
- **Objectif** : Nettoyer, transformer et structurer les données pour l'analyse.
- **Exemple** :
  - Supprimer les valaeurs aberrantes
  - Remplir les valeurs manquantes
  - Convertir les dates en format utilisable
  - Créer une variable "taux d'abandon de panier"

la qualité des analyses dépend avant tout de celle des données utilisées.

Étape 4 : Analyse des données

- Une fois les données nettoyées, il est temps de procéder à l'analyse proprement dite. Il s'agit d'appliquer des techniques statistiques ou mathématiques aux données afin de découvrir des modèles, des relations ou des tendances.
- **Objectif** : Appliquer des techniques statistiques ou technologique pour répondre à la question.
- **Exemple** :
  - Utiliser un test de corrélation entre le trafic et les conversions
  - Appliquer une régression logistique pour prédire la probabilité de conversion selon l'origine du visiteur
  - Identifier des segments de clients qui convertissent moins

Étape 5 : Interprétation et visualisation des données

- Après l'analyse des données, l'étape suivante consiste à interpréter les résultats et à les visualiser de manière à ce qu'ils soient faciles à comprendre.
- **Objectif** : Comprendre les relations entre les variables et présenter les résultats de façon visuelle.
- **Exemple** :
  - Tracer un histogramme des ventes par canal d'acquisition
  - Créer un graphique en boîte des montants dépensés par catégorie de clients
  - Utiliser un nuage de points pour corréler durée de visite et conversion

Les types d'analyse des données

- L'analyse des données peut être classée en quatre catégories principales, chacune ayant un objectif unique et fournissant des informations différentes. Il s'agit d'analyses descriptives, diagnostiques, prédictives et normatives.

L'analyse descriptive

- L'analyse descriptive, comme son nom l'indique, décrit ou résume des données brutes et les rend interprétables. Il s'agit d'analyser des données historiques pour comprendre ce qui s'est passé dans le passé.

- Par exemple, une entreprise peut utiliser l'analyse descriptive pour comprendre les ventes mensuelles moyennes de l'année écoulée.

analyse descriptive

Analyse diagnostique

- L'analyse diagnostique va plus loin que l'analyse descriptive en déterminant le pourquoi d'un événement. Elle implique une exploration plus détaillée des données et la comparaison de différents ensembles de données pour comprendre la cause d'un résultat particulier.

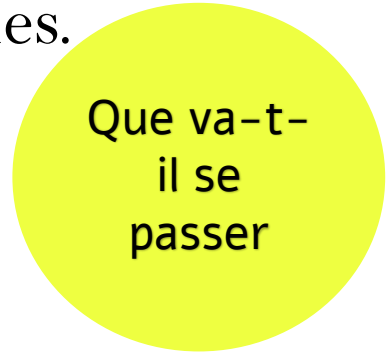
- Par exemple, si les ventes d'une entreprise ont chuté au cours d'un mois donné, l'analyse diagnostique peut être utilisée pour en **déterminer les raisons.**

Analyse diagnostique

Analyse prédictive

- L'analyse prédictive utilise des modèles statistiques et des techniques de prévision pour comprendre l'avenir. Il s'agit d'utiliser des données du passé pour prédire ce qui pourrait se produire dans le futur. Ce type d'analyse est souvent utilisé pour l'évaluation des risques, le marketing et les prévisions de ventes.

- Par exemple, une entreprise peut utiliser l'analyse prédictive pour prévoir les ventes du prochain trimestre sur la base de données historiques.



Analyse prédictive

Analyse normative

- L'analyse prescriptive est le type d'analyse de données le plus avancé. Il ne se contente pas de prédire les résultats futurs, il propose également des actions pour tirer parti de ces prédictions. Il utilise des outils et des technologies sophistiqués comme l'apprentissage automatique et l'intelligence artificielle pour recommander des décisions.

- Par exemple, une analyse prescriptive peut suggérer les meilleures stratégies de marketing pour augmenter les ventes futures.

Analyse prescriptive

Principaux Outils d'Analyse de Données

Microsoft Excel : est un tableur largement utilisé pour l'analyse de données, offrant des fonctionnalités puissantes pour les calculs, les graphes et la visualisation des données.

Caractéristiques Principales :

- **Fonctions et Formules**
- **Tableaux Croisés Dynamiques**
- **Graphiques et Visualisations**
- **Macros et VBA**
- **Outils d'Analyse de Données**

Environ 81% des entreprises utilisent Excel pour des tâches analytiques et des rapports financiers. (AlloExcel) (atelier-mosesu).

Principaux Outils d'Analyse de Données

Python *:* est un langage de programmation polyvalent et populaire pour l'analyse de données, offrant des bibliothèques puissantes pour le traitement des données, les statistiques, et le machine learning.

Caractéristiques Principales :

- **Pandas**
- **NumPy**
- **Matplotlib et Seaborn**
- **Scikit-learn**

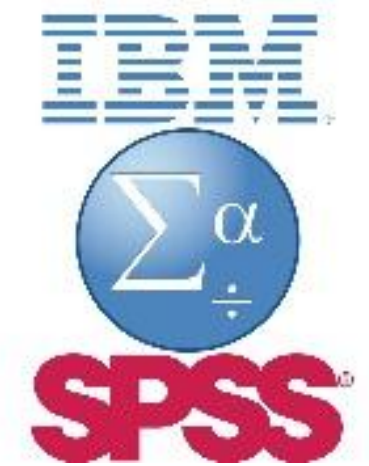
Environ 60% des data scientists utilisent Python pour leurs projets d'analyse de données.
(Geekflare) (Developpez Python).

Principaux Outils d'Analyse de Données

SPSS (Statistical Package for the Social Sciences) : est un logiciel d'analyse statistique utilisé principalement dans les sciences sociales et la recherche.

Caractéristiques Principales :

- **Interface Utilisateur Graphique**
- **Analyses Statistiques Avancées**
- **Manipulation des Données**
- **Rapports et Visualisations**
- **Extensions et Scripts**



Environ 84% des chercheurs en sciences sociales utilisent SPSS pour l'analyse de leurs données. (Techno-Science.net).

Principaux Outils d'Analyse de Données

Microsoft PowerBI : est un outil de business intelligence qui permet de créer des tableaux de bord interactifs et des rapports visuels à partir de diverses sources de données.

Caractéristiques Principales :

- **Tableaux de Bord Interactifs**
- **Intégration de Données**
- **Visualisations Riches**
- **Fonctionnalités de Drag-and-Drop**
- **Collaborations et Partage**

Les entreprises utilisant Power BI rapportent une amélioration de 25% de leur prise de décision basée sur les données. (Power BI) (Hub Collab).

Principaux Outils d'Analyse de Données

Chacun de ces outils - Excel, Python, SPSS, et Microsoft Power BI - offre des capacités uniques pour l'analyse des données.

- Excel est idéal pour les tâches analytiques quotidiennes et la visualisation de base,
- Python excelle dans les analyses avancées et le machine learning,
- SPSS est spécialisé dans les analyses statistiques complexes,
- Microsoft Power BI est parfait pour la création de tableaux de bord interactifs et la visualisation en temps réel.

➔ **Ensemble, ces outils permettent une analyse des données complète et efficace, adaptée à divers besoins professionnels et académiques.**

Concepts de base en analyse des données

Types de Variables

Les variables représentent les caractéristiques mesurées dans un ensemble de données et peuvent être classées en plusieurs types :

a) Variables Qualitatives (Qualitative Variables)

- **Nominales** : catégories sans ordre particulier (exemple1 : genre, couleur des yeux) - (exemple2 : secteurs d'activité, types de produits financiers).
- **Ordinales** : catégories avec un ordre logique (exemple1 : niveau d'éducation) (ex2: notations de crédit, niveaux de risque).

b) Variables Quantitatives (Quantitative Variables)

- **Discrètes** : valeurs numériques entières (exemple : nombre d'enfants) (ex: nombre de transactions, nombre d'actions).
- **Continues** : valeurs numériques pouvant prendre n'importe quelle valeur sur un intervalle (exemple : taille, poids) (ex: rendements, prix des actions, taux d'intérêt).

Introduction

Les variables représentent les caractéristiques mesurées dans un ensemble de données et peuvent être classées en plusieurs types :

Variables Qualitatives
(données catégorielles)

Nominale
- Variables sans ordre ni hiérarchie naturelle.
- Exemples : Sexe, Couleur, Nationalité, Race...

Ordinale
- Variables avec un ordre ou une séquence logique.
- Exemples : Niveau d'études, Groupe sanguin, Niveau de satisfaction, Performance...

Binaire
- Variables avec deux options seulement.
- Exemples : Succès/Échec, Oui/Non, Vrai/Faux...

Variables Quantitatives
(données numériques)

Discrète
- Données avec un nombre fini de valeurs possibles
- Exemple : Nombre de pièces endommagées lors d'un envoi.

Continue
- Données mesurables sur une échelle continue
- Peuvent prendre presque n'importe quelle valeur numérique.
- Exemples : Longueur, Taille, Largeur, Température, Temps...

Mesures de Tendance Centrale

- Les mesures de tendance centrale indiquent la valeur autour de laquelle les données se regroupent.
- Les 3 mesures les plus utilisées sont :

  1 – le **mode :** La valeur la plus fréquente dans un ensemble de données.

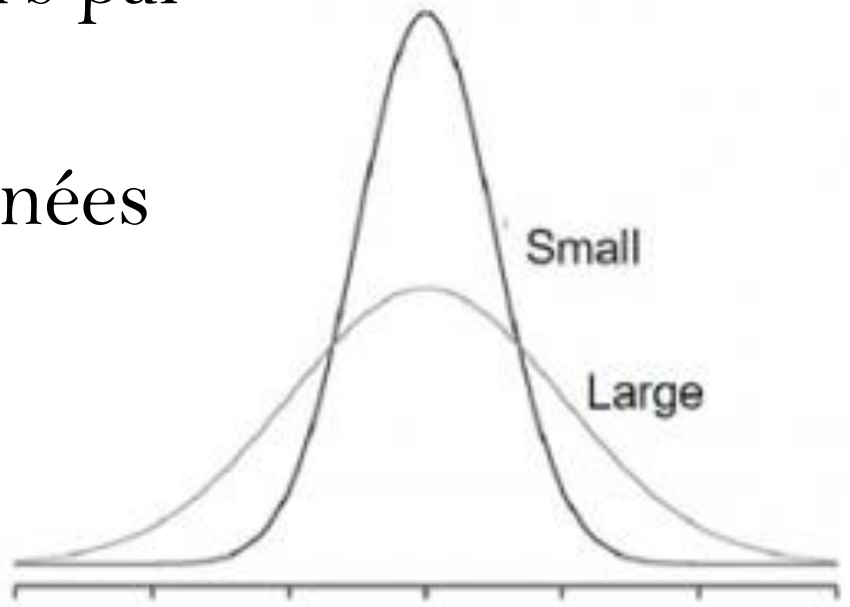
  2 – la **médiane :** la valeur qui divise en 2 parties égales un ensemble ordonné de scores.

  3 – la **moyenne :** la valeur "**typique**" ou **centrale** d'un ensemble de données. C'est **la somme des valeurs** divisée par le **nombre total d'observations**

- Sur la base des données recueillies dans un échantillon, les indicateurs de position (Mesures de Tendance Central) fournissent des informations sur l'emplacement du "centre" de la distribution.
- Les mesures de localisation peuvent être utilisées pour résumer ou décrire une liste de données avec un seul paramètre.

Mesures de dispersion

- Les mesures de dispersion permettent d'évaluer l'homogénéité ou l'hétérogénéité d'un ensemble de données en mesurant l'écart des valeurs par rapport à la **moyenne** ou à la **médiane**.
- Une faible dispersion indique que les données sont concentrées autour de la moyenne.
- Une forte dispersion indique une grande variabilité des données.



Mesures de dispersion

1. Étendue (Amplitude)

- L'étendue est la différence entre la valeur maximale et la valeur minimale.
- Etendue de $\mathbf{X = X_{max} - X_{min}}$
- **Interprétation :**
  - Mesure simple mais sensible aux valeurs extrêmes.
  - Utile pour une première évaluation de la variabilité des données
- L'inconvénient de l'étendue est qu'elle dépend uniquement des deux valeurs les plus extrêmes de la distribution. Elle indique donc la différence maximum entre deux valeurs mais pas la différence typique.
- Exemple : Notation des professeurs X et Y :
  - L'étendue des notes données par le professeur **X** est de **(13-7)=6**, ce qui signifie que l'écart maximum entre deux notes du professeur **X** est de 6.
  - L'étendue des notes données par le professuer **Y** est de **(20-0)=20** ce qui signifie que l'écart maximum entre deux notes du professeur **Y** est de **20.**

    → La dispersion des notes du professeur Y est donc beaucoup plus forte que celle des notes du professeur X.

Mesures de dispersion

2. Variance

- La variance mesure la moyenne des carrés des écarts (σ) par rapport à la moyenne.
- **Interprétation :**
  - Plus la variance est élevée, plus les valeurs sont dispersées.
  - La variance est exprimée en unités carrées, ce qui complique son interprétation directe.

La **variance** d'une série de n observation $x_1, \ldots, x_n$ de moyenne $\overline{x}$ est :

$$\sigma^2 = \frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n}.$$

C'est la *moyenne du carré des écarts à la moyenne* $\overline{x}$.

Mesures de dispersion

2. Variance

Ex :

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n}$$

- Exemple. Voici les âges d'un groupe de 6 enfants :

  $x_1 = 6, x_2 = 6, x_3 = 7, x_4 = 7, x_5 = 7, x_6 = 8$
- L'âge moyen est $\bar{x} = 6{,}8$ ans.
- La variance est
- $\sigma^2 = \frac{(6-6{,}8)^2 + (6-6{,}8)^2 + (7-6{,}8)^2 + (7-6{,}8)^2 + (7-6{,}8)2 + (8-6{,}8)^2}{6}$

  $= 0{,}47$

Mesures de dispersion

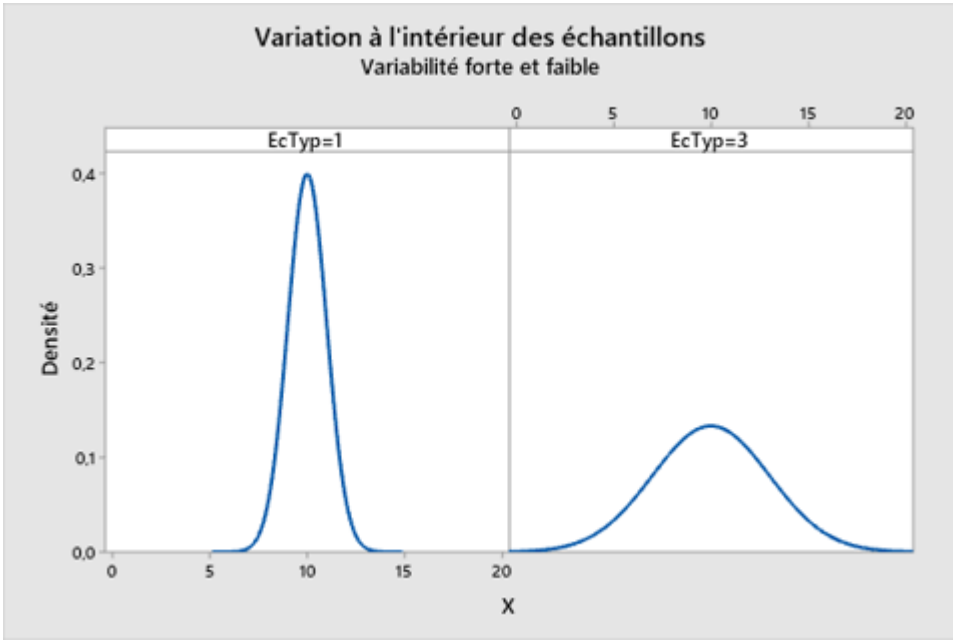
3. Écart-Type :

- est une mesure de dispersion des données autour de la moyenne.
- L'écart-type est la racine carrée de la variance. Il est exprimé dans la même unité que les données.

$$écart - type = \sqrt{variance}$$

Interprétation :

- Plus l'écart-type est élevé, plus les données sont dispersées.
- Une faible valeur indique une concentration des données autour de la moyenne.

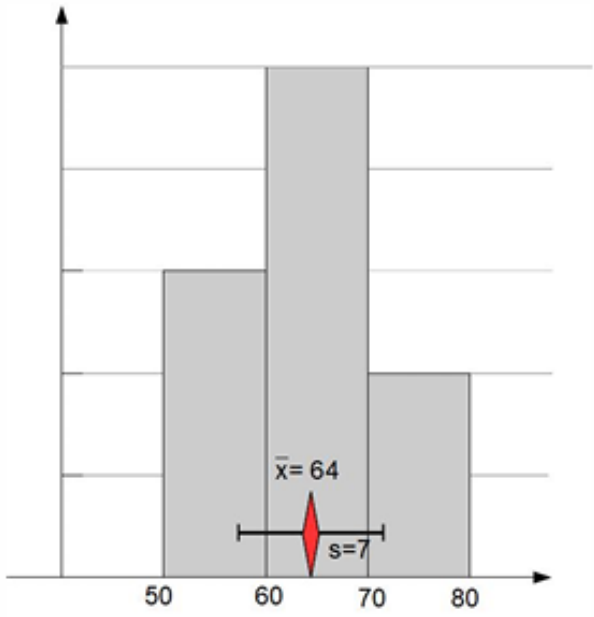


Mesures de dispersion

3. Écart-Type :

- La **variance** et **l'écart-type** mesurent l'écart des données par rapport à la **moyenne**. Ils sont les paramètres de dispersion les plus importants.

Exemple. Poids en grammes des oeufs.
Variance : $\sigma^2 \simeq 57$ grammes^2
Écart-type : $\sigma \simeq \sqrt{57} \simeq 7,6$ grammes
Exemple. On peut penser approximativement que le poids d'un oeuf s'éloigne typiquement de 7,6g du poids moyen de 64g.



Chapitre 3 : Préparation des données

Préparation des données – Présentation générale

- Le terme « préparation des données » désigne les opérations de nettoyage et transformation qui doivent être appliqués aux données brutes avant leur traitement et analyse.
- Il s'agit d'une étape importante avant le traitement proprement
- Il implique souvent de reformater et corriger les données et de combiner des datasets pour enrichir certaines données.

- Par exemple, le processus de préparation des données comprend généralement les opérations suivantes :
  - standardisation des formats (types) de données,
  - enrichissement des données source
  - suppression des valeurs aberrante
  - …

Préparation des données – Présentation générale

- La préparation des données est la partie la plus fastidieuse de leur travail, mais aussi que les décisions efficaces et précises ne peuvent être prises qu'avec des données « propres ».
- GIGO (garbage in, garbage out) est l'idée selon laquelle des données d'entrée défectueuses ou absurdes produisent des sorties absurdes ou « déchets ».

Préparation des données – Présentation générale

- La préparation des données permet d'obtenir les résultats suivants :
  - Corriger les erreurs rapidement
  - Obtenir des données de grande qualité
  - Prendre des décisions plus avisées

Préparation des données – Workflow général

Chapitre 4 : Techniques d’exploration des données

Techniques d'exploration des données

- Analyse exploratoire des données (AED) : visualisation des données, histogrammes, diagrammes de dispersion

- Analyse des relations entre les variables : corrélation, matrices de covariance

Analyse exploratoire des données (AED)

- **Qu'est-ce que l'Analyse exploratoire des données ?**
- L'analyse exploratoire des données (AED) est une approche initiale de l'analyse des données qui vise à résumer leurs principales caractéristiques, souvent à l'aide de méthodes visuelles et statistiques. Elle permet de :
  - Comprendre la structure des données disponibles.
  - Détecter les valeurs manquantes, aberrantes ou extrêmes.
  - Identifier les tendances, motifs et relations entre les variables.
- L'AED constitue une étape cruciale avant l'application de modèles d'apprentissage automatique ou prédictif, car elle permet de poser les bonnes hypothèses et de mieux préparer les données.

Analyse exploratoire des données (AED)

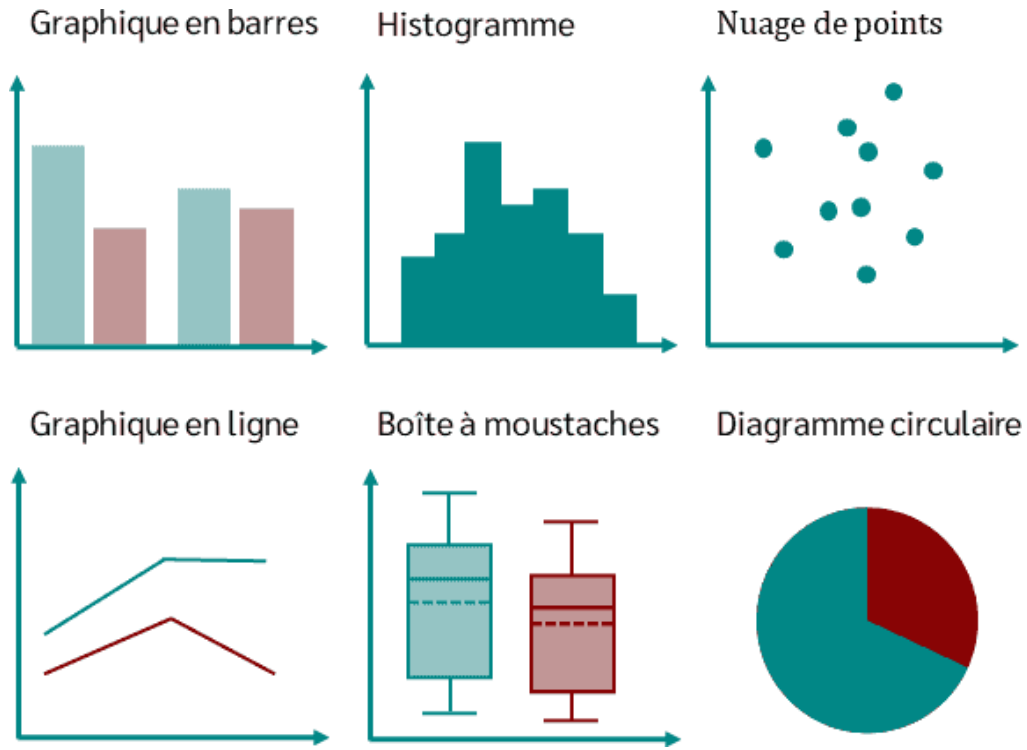
Visualisation des données

La visualisation des données est essentielle pour :

- Identifier des tendances globales.
- Détecter des valeurs aberrantes.
- Communiquer efficacement les résultats.

Types de visualisations utiles :

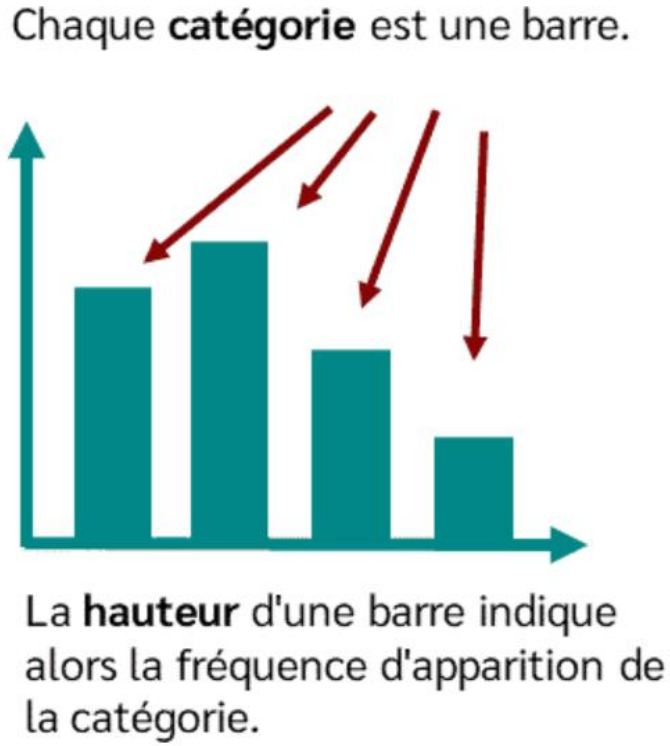
- **Diagrammes en barres**
- **Histogrammes**
- **Diagrammes de dispersion**
- **Graphique linéaire**
- **Diagrammes en boîte - Boxplots (Boîtes à moustaches)**



Analyse exploratoire des données (AED)

Diagrammes en barres

- Les diagrammes en barres sont probablement les diagrammes les plus couramment utilisés en statistiques.
- Ils sont généralement utilisés pour montrer la fréquence de différentes catégories, mais aussi pour visualiser des données numériques, telles que des chiffres de vente ou des statistiques démographiques.
- Dans un diagramme en barres, la longueur de chaque barre est proportionnelle à la valeur qu'elle représente. Les barres sont généralement disposées horizontalement ou verticalement.



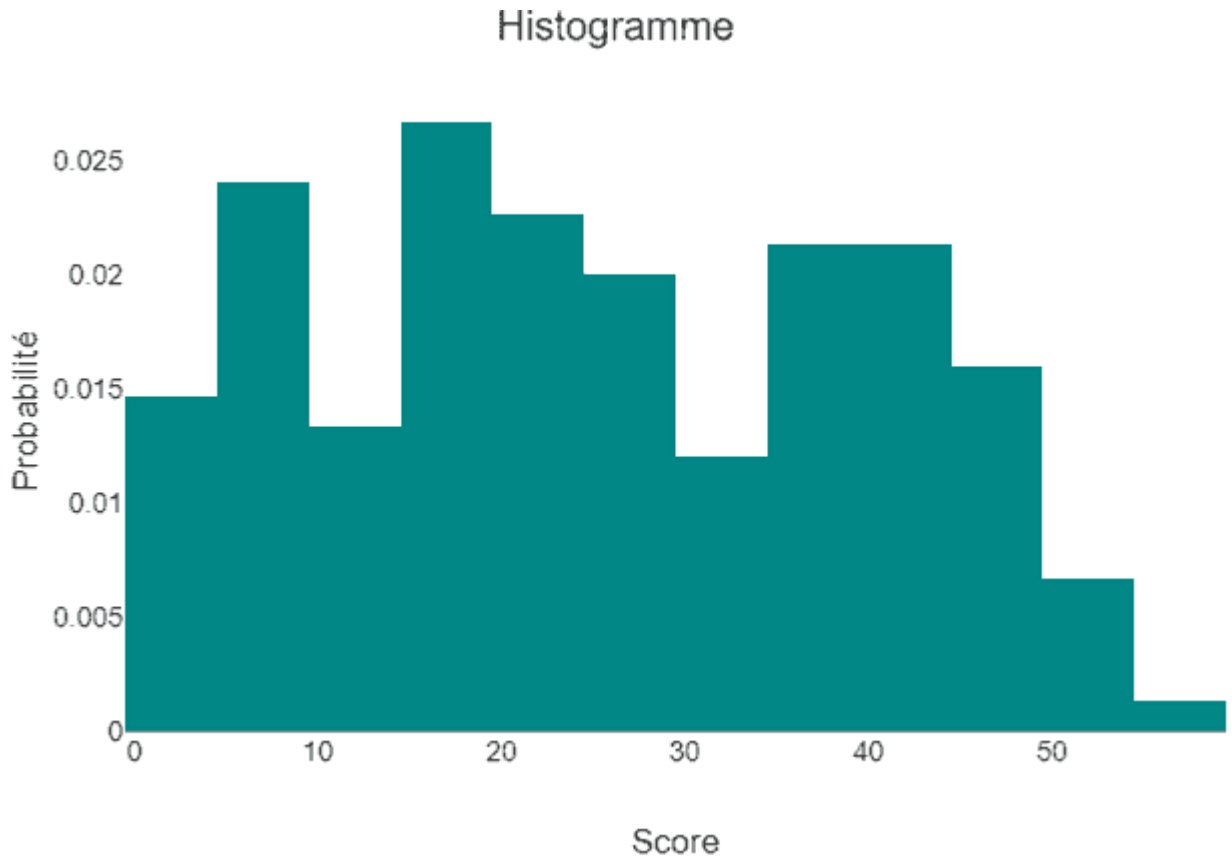
Analyse exploratoire des données (AED)

Histogramme

- Un histogramme est une représentation graphique de la distribution de fréquence d'une variable métrique.
- Pour afficher une distribution de données dans un histogramme, les données doivent d'abord être divisées en classes, également appelées bins.
- Ces classes ou cellules sont ensuite représentées par des rectangles situés directement l'un à côté de l'autre.

Analyse exploratoire des données (AED)

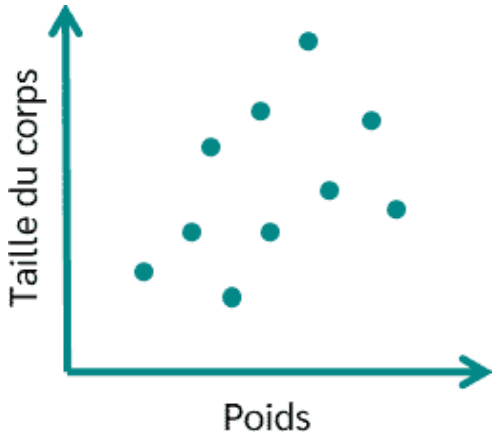
Histogramme

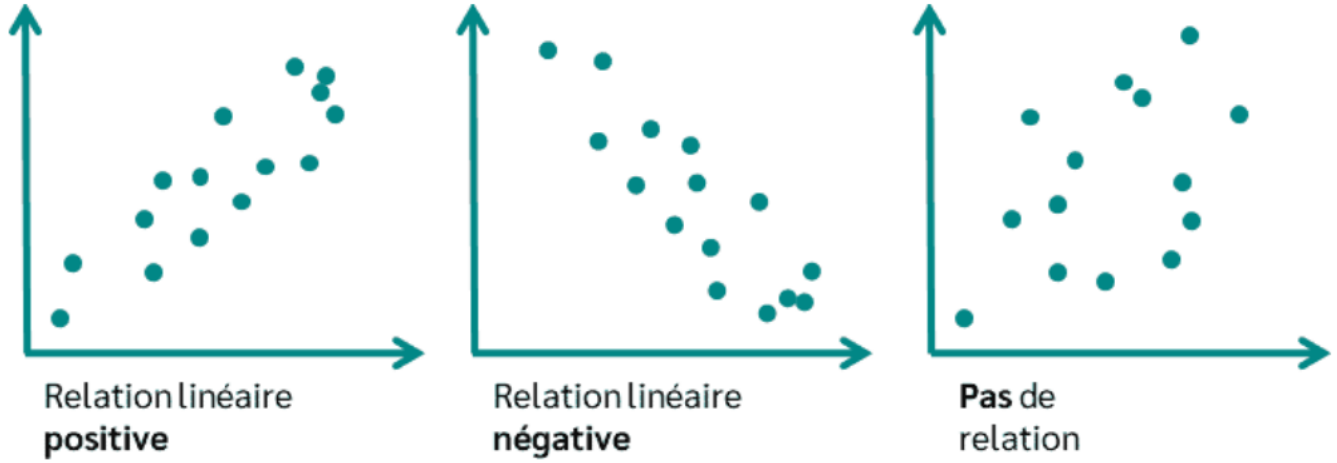


Analyse exploratoire des données (AED)

Diagrammes de dispersion

- Les diagrammes de dispersion sont utilisés en statistique pour visualiser les corrélations entre les données.
- Dans un diagramme de dispersion, deux variables peuvent toujours être représentées, ce qui se fait en représentant chaque paire de valeurs d'un cas comme un point dans un système de coordonnées. Si, par exemple, on demande à 10 personnes de donner leur poids et leur taille, le nuage de points montre 10 points.
- Le nuage de points vous donne une première indication de la corrélation entre les deux variables visualisées.

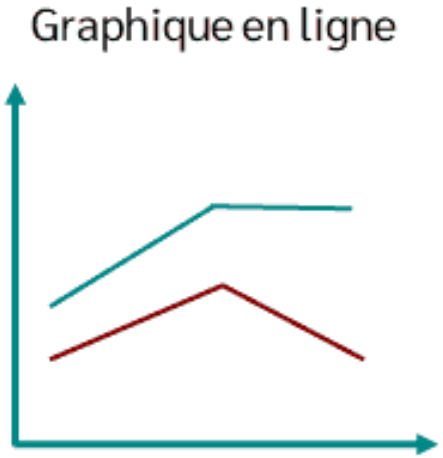




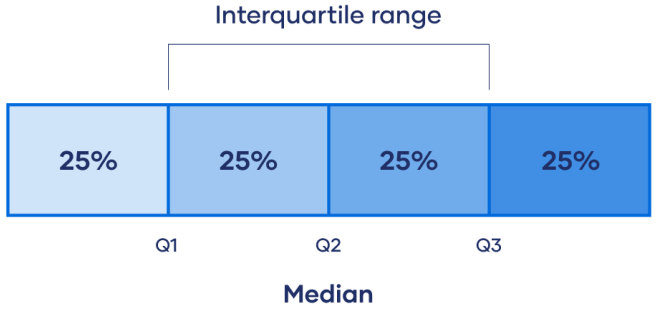
Analyse exploratoire des données (AED)

Graphique linéaire

- Un graphique linéaire est un graphique composé d'une série de points de données reliés par une ligne.
- Il est utilisé, par exemple, pour montrer une évolution continue des données dans le temps. Dans un graphique linéaire, le temps ou l'autre variable continue est représenté sur l'axe horizontal, tandis que les valeurs des données à illustrer sont représentées sur l'axe vertical.
- Les graphiques linéaires sont particulièrement utiles pour visualiser les tendances et les changements dans le temps.

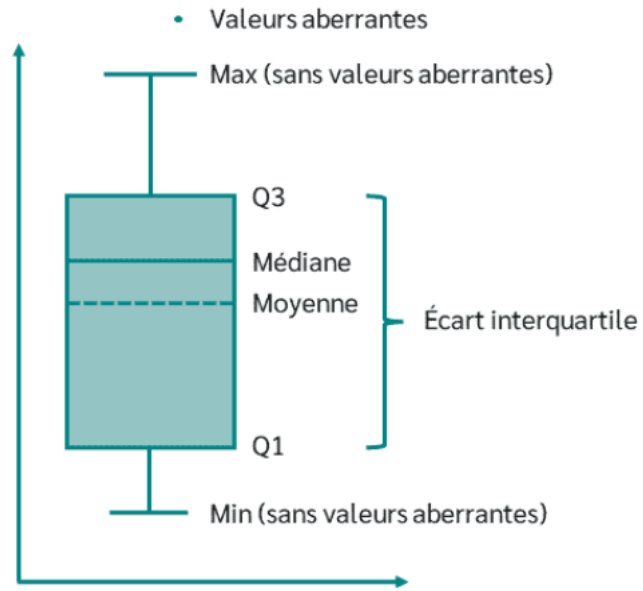


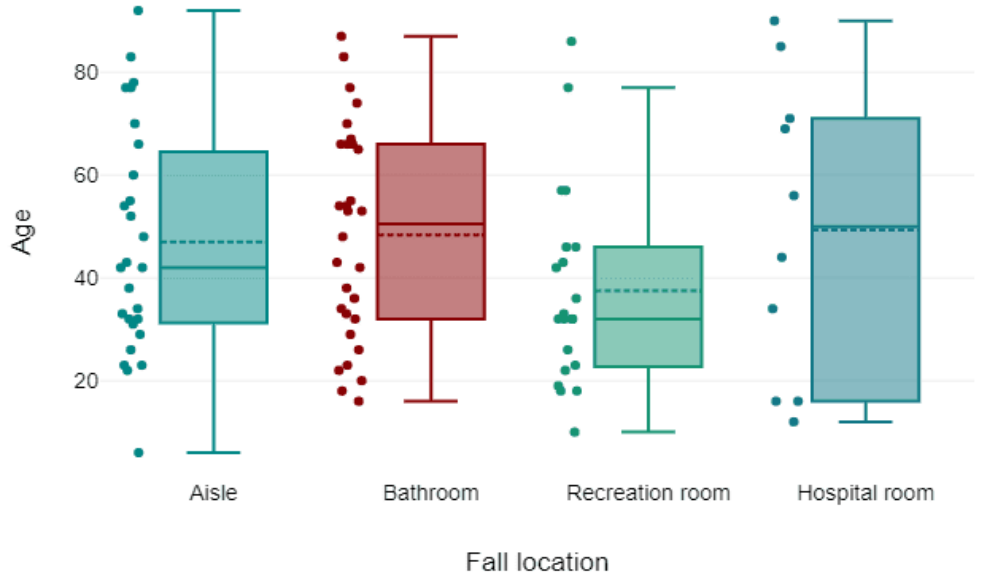
Analyse exploratoire des données (AED)



Diagrammes en boîte

- Les diagrammes en boîte sont des graphiques utilisés pour représenter des distributions de données. Ils fournissent un résumé visuel des données en présentant des mesures statistiques importantes telles que la médiane, les quartiles et les valeurs aberrantes dans un seul graphique.





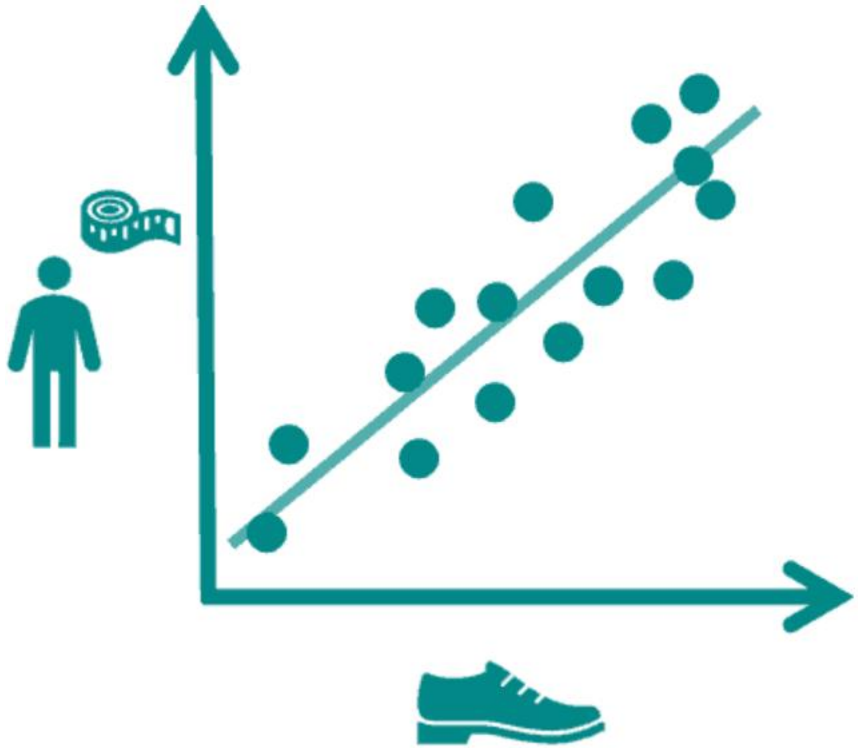
Analyse des relations entre les variables : corrélation, matrices de covariance

- L'analyse des relations entre les variables est une étape essentielle dans de nombreuses disciplines, notamment en sciences sociales, en économie, en finance, en biologie et en data science. Elle permet de comprendre comment différentes variables interagissent et influencent les phénomènes étudiés.
- Deux concepts fondamentaux pour analyser les relations entre les variables : la **corrélation** et les **matrices de covariance**.
  - **La corrélation**
  - **La matrice de covariance**

Analyse des relations entre les variables

Corrélation

- La corrélation mesure la force et la direction de la relation linéaire entre deux variables. Elle est exprimée par le **coefficient de corrélation**
- Une analyse de corrélation est une technique statistique qui fournit des informations sur le lien entre deux variables, par exemple s'il existe un lien entre la taille du corps et la taille des chaussures.

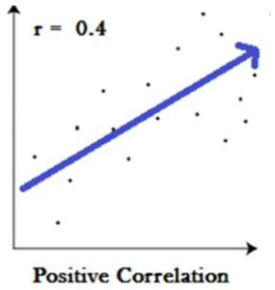


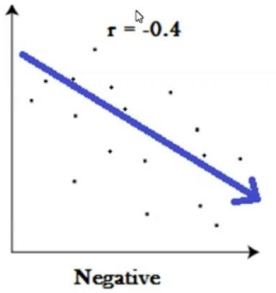
Analyse des relations entre les variables

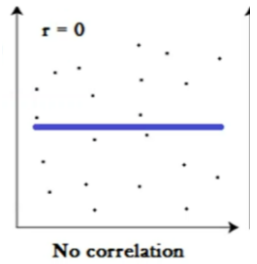
Corrélation

Types de corrélation :

- **Corrélation positive** :
  - Si une variable augmente, l'autre augmente également.
  - Exemple : Le prix des actions d'une entreprise et le marché global (dans certains cas).
  - Coefficient de corrélation proche de **+1**.
- **Corrélation négative** :
  - Si une variable augmente, l'autre diminue.
  - Exemple : La demande d'obligations peut être corrélée négativement aux taux d'intérêt.
  - Coefficient de corrélation proche de **-1**.
- **Absence de corrélation** :
  - Les deux variables n'ont aucune relation linéaire.
  - Exemple : Le rendement d'un fonds en Chine et celui d'une entreprise dans un secteur très différent.
  - Coefficient de corrélation proche de **0**.







Analyse des relations entre les variables

Corrélation

Comment la mesurer ?

- La corrélation de Pearson est une méthode fréquemment utilisée pour mesurer cette relation. Sa formule est :

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \cdot \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

  - r est le coefficient de corrélation.
  - X et Y sont les deux variables.
  - $\bar{X}$ et $\bar{Y}$ sont les moyennes des variables X et Y.

Analyse des relations entre les variables

Matrices de covariance

- La covariance mesure comment deux variables varient ensemble. Contrairement à la corrélation, elle ne mesure pas la force de la relation mais plutôt la direction (positive ou négative).
- Une covariance positive indique que les deux variables ont tendance à augmenter ensemble, tandis qu'une covariance négative indique qu'elles varient en sens opposé.

Analyse des relations entre les variables

Matrices de covariance

Formule de la covariance :

- Pour deux variables X *et* Y avec n observations, la covariance est calculée comme suit :

$$\mathrm{Cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

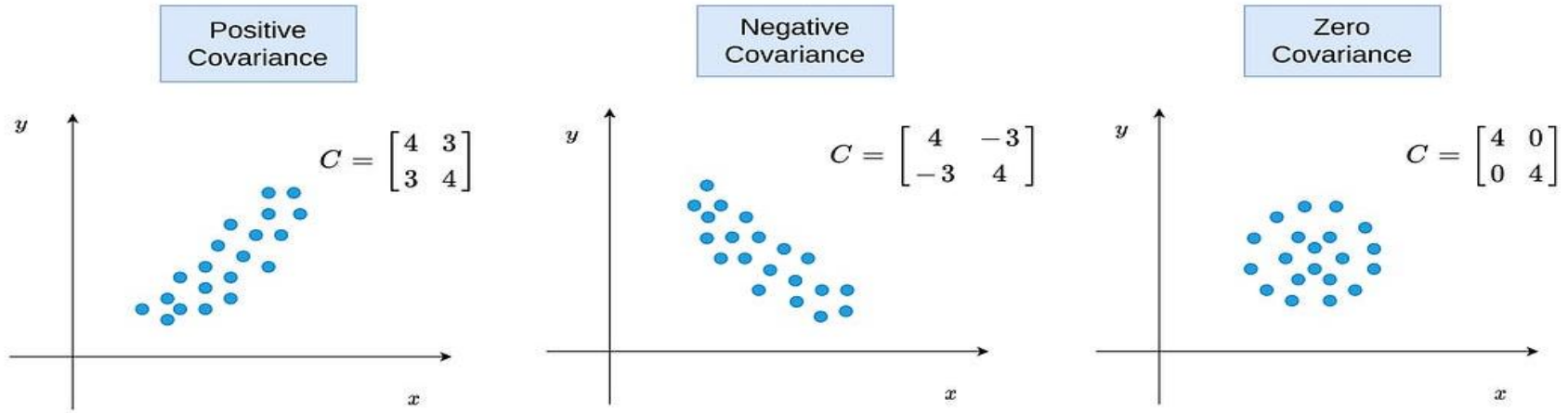
- X_i et Y_i sont les valeurs individuelles des variables.
- $\bar{X}$ et $\bar{Y}$ sont les moyennes des variables X et Y.

Analyse des relations entre les variables

Matrices de covariance

Interprétation :

1. **Covariance positive** :
    1. Si la covariance est **positive**, cela signifie que les deux variables ont tendance à augmenter ou diminuer ensemble.
    2. Exemple : Les rendements d'actions liées à un même secteur (comme les banques).
2. **Covariance négative** :
    1. Une covariance **négative** indique qu'une augmentation de l'une est associée à une diminution de l'autre.
    2. Exemple : Les obligations et les actions ont souvent une covariance négative.
3. **Covariance nulle** :
    1. Une covariance proche de **zéro** signifie qu'il n'y a pas de relation linéaire claire entre les deux variables.



Conclusion

La Data Analytics est un outil puissant qui aide à transformer des données brutes en informations exploitables. Elle joue un rôle crucial dans la prise de décisions, l'amélioration des processus et la compréhension des tendances.

Fin