

Chapitre 2:

Collecte et préparation des données

Pourquoi cette étape est-elle cruciale ?

- 80 % du temps d'un projet en analyse de données est consacré à la préparation.
- Une donnée mal collectée ou mal nettoyée = des résultats biaisés ou inutilisables.
- Étape indispensable pour garantir la qualité, la fiabilité et la pertinence de l'analyse.

- Principe fondamental:



Si les données en entrée sont de mauvaise qualité, les résultats le seront aussi.

2.1. Méthodes de collecte de données

Panorama des méthodes de collecte

- **Enquêtes et sondages** (questionnaires, formulaires en ligne: Google Forms, SurveyMonkey).
- **Expériences** (tests en laboratoire, essais cliniques).
- **Observations** (caméras, capteurs, suivi comportemental).
- **Sources numériques** : Web scraping, API, logs informatiques.

Enquêtes et échantillonnage

- **Enquêtes :**
 - Rapides, économiques, mais dépend de la sincérité des réponses.
- **Échantillonnage :**
 - Aléatoire simple.
 - Stratifié (par sous-groupes).
 - En grappes (cluster sampling).

Échantillonnage aléatoire simple

- **Définition:** chaque individu de la population a la **même probabilité** d'être sélectionné. La sélection est totalement aléatoire (tirage au sort).
- **Exemple:** une université veut interroger 100 étudiants sur leurs habitudes de lecture. Elle tire 100 noms au hasard dans la liste complète des étudiants inscrits.
- **Avantage:** méthode simple, réduit les biais.
- **Limite:** nécessite une liste exhaustive de la population (pas toujours disponible).

Échantillonnage stratifié

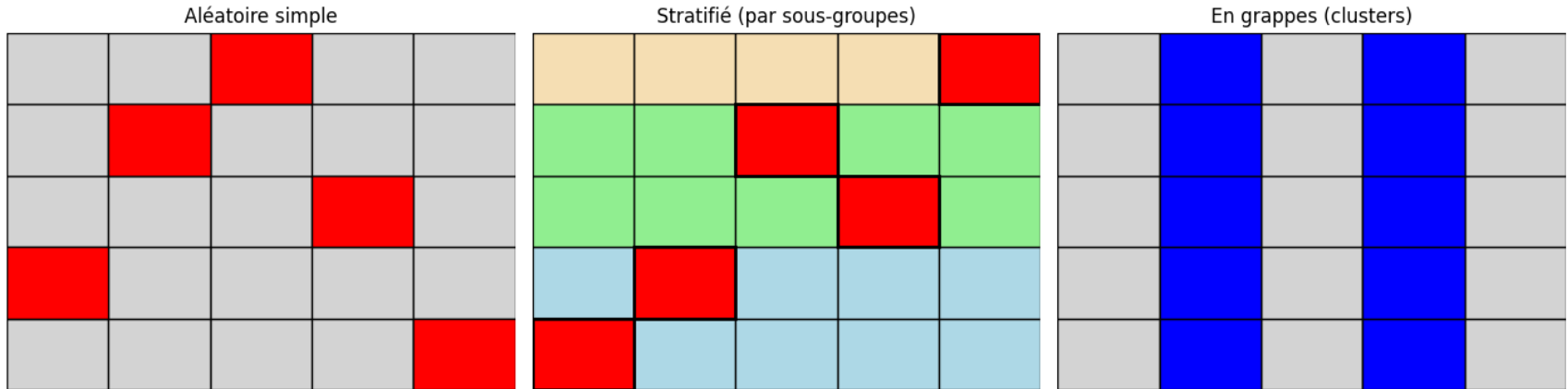
- **Définition:** la population est divisée en sous-groupes homogènes (strates) selon une caractéristique (sexe, âge, revenu, etc.), puis on tire un échantillon dans chaque strate proportionnellement à sa taille.
- **Exemple:** pour un sondage politique, on divise la population en strates selon l'âge (18–34, 35–54, 55+). Si 30% de la population a entre 18–34 ans, on veille à ce que 30% de l'échantillon appartienne à cette tranche.
- **Avantage:** garantit une bonne représentativité des sous-groupes.
- **Limite:** nécessite de connaître la structure de la population à l'avance.

Échantillonnage en grappes (Cluster sampling)

- **Définition** : la population est divisée en **grappes naturelles** (souvent géographiques ou organisationnelles). On tire au hasard quelques grappes et on interroge **tous les individus** de ces grappes (ou un sous-échantillon à l'intérieur).
- **Exemple** : une enquête sur la santé dans une ville. Plutôt que de tirer des individus au hasard dans toute la ville, on tire au hasard 5 quartiers (grappes), puis on interroge les habitants de ces quartiers.
- **Avantage** : moins coûteux et plus pratique (surtout quand la population est dispersée).
- **Limite** : moins représentatif que le stratifié, risque de biais si les grappes sont très différentes entre elles.

Méthodes d'échantillonnage

Méthodes d'échantillonnage



- Aléatoire simple : chaque individu de la population a la même probabilité d'être choisi (les cases rouges sont tirées au hasard).
- Stratifié : la population est divisée en sous-groupes (couleurs différentes) et l'échantillon est prélevé proportionnellement dans chaque groupe.
- En grappes (clusters) : la population est divisée en grappes naturelles (ex. écoles, quartiers) et on sélectionne certaines grappes entières (colonnes en bleu).

Données numériques (Web & API)

- **Web scraping** : extraction automatisée depuis un site web.
 - Exemple : récupérer les prix des produits sur un site e-commerce.
- **API** (Application Programming Interface) : accès direct et structuré aux données.
 - Exemple : *Twitter API* pour collecter des tweets, *OpenWeather* pour les données météo.

Enjeux et limites de la collecte

- **Avantages** : richesse des sources, rapidité, accessibilité.
- **Limites** :
 - Biais d'échantillonnage.
 - Données incomplètes ou bruitées.
 - Questions légales et éthiques (RGPD, confidentialité).



2. 2. Nettoyage des données

Gestion des données manquantes

- Bien que le volume de données disponibles ne cesse de croître avec l'essor du Big Data, la question des données manquantes demeure fréquente dans les analyses de données et requiert une attention spécifique.
- Les ignorer peut non seulement réduire la précision des résultats, mais aussi introduire d'importants biais dans les modèles d'analyse.

Types de données manquantes

- Afin d'aborder correctement l'imputation des données manquantes il faut en distinguer les causes, surtout si elles ne sont pas au hasard.
- Une typologie a été développée par Little & Rubin (1987), les répartissant en trois catégories :
 - MCAR (missing completely at random)
 - MAR (Missing at random)
 - MNAR (Missing not at random)

MCAR (missing completely at random)

- L'emplacement des valeurs manquantes dans l'ensemble de données est purement aléatoire et ne dépend d'aucune autre donnée.
- **Exemple:** Un capteur météorologique mesure la température et envoie les données à une base de données. Certaines entrées de la base de données sont manquantes lorsque le capteur est tombé en panne.

MAR (Missing at random)

- L'emplacement des valeurs manquantes dans l'ensemble de données dépend d'autres données observées.
- **Exemple:** Il manque des valeurs de température dans la base de données lorsque le capteur a été éteint pour maintenance. L'équipe de maintenance ne travaillant jamais le week-end, l'emplacement des valeurs manquantes dépend du jour de la semaine.

MNAR (Missing not at random)

- L'emplacement des valeurs manquantes dans l'ensemble de données dépend des valeurs manquantes elles-mêmes.
- **Exemple:** Par temps extrêmement froid, le capteur météo gèle et cesse de fonctionner. Il n'enregistre donc pas les températures très basses. Par conséquent, l'emplacement des valeurs manquantes dans la variable de température dépend des valeurs de cette variable.

Gestion des données manquantes

- Que se passe-t-il si l'on supprime simplement les observations incomplètes ?
 - Si les données sont de type MCAR, leur suppression entraîne une perte d'information.
 - Si les données sont de type MAR ou MNAR, leur suppression introduit un biais dans les modèles construits sur ces données. Dans ce cas, les valeurs manquantes doivent être imputées.
- De nombreuses méthodes d'imputation supposent l'existence d'un MAR ; il est donc important de le détecter.

Détection et traitement des valeurs aberrantes (outliers)

- **Définition**

- Une **valeur aberrante (outlier)** est une donnée qui s'écarte fortement du reste des observations.
- Elle peut être due à une **erreur de saisie**, un **problème de mesure**, ou refléter un **cas particulier réel** (mais rare).

- **Méthodes de détection**

- Règle des **écarts-types** : valeur située à plus de 2σ de la moyenne.
- **IQR (Interquartile Range)** : valeurs en dehors de $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$.

Règle des écarts-types

- Principe :
 - On calcule la moyenne (μ) et l'écart-type (σ) de l'échantillon.
 - Une valeur est considérée comme aberrante si :
 - Elle est située à plus de 3σ au-dessus ou en dessous de la moyenne.
- Formule :
 - Valeur aberrante si $x < \mu - 3\sigma$ ou $x > \mu + 3\sigma$

Règle des écarts-types: Exemple

- Le jeu de données comporte 11 valeurs. Il contient quelques valeurs extrêmes; vous utiliserez donc la méthode règle des écarts-types pour vérifier s'il s'agit de valeurs aberrantes.

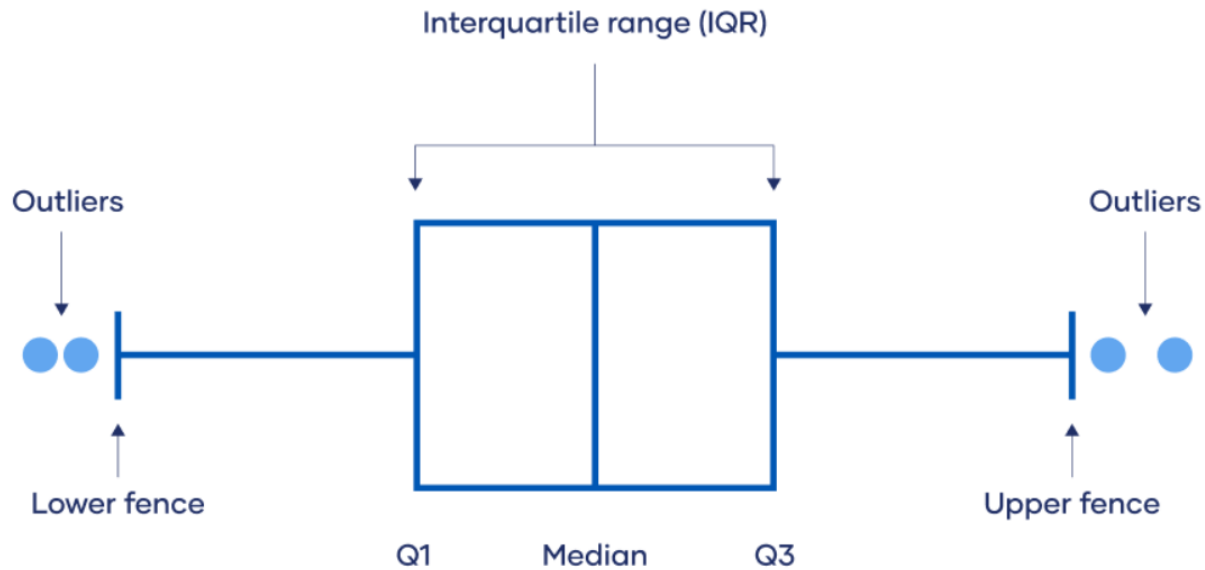
26, 37, 24, 28, 35, 22, 31, 53, 41, 64, 29

- Moyenne (μ) = 35.45455
- Écart-type (σ) = 12.94113

Limite supérieure = $\mu + 3\sigma = 74.27792$

Limite inférieure = $\mu - 3\sigma = -3.368831$

IQR (Interquartile Range)



IQR (Interquartile Range)

1. Triez les données du plus bas au plus haut.
2. Identifiez le premier quartile (Q1), la médiane et le troisième quartile (Q3).
3. Calculez l'**IQR = Q3 - Q1**.
4. Calculez la limite supérieure = $Q3 + (1.5 * IQR)$.
5. Calculez la limite inférieure = $Q1 - (1.5 * IQR)$.
6. Utilisez les limites pour mettre en évidence les valeurs aberrantes, c'est-à-dire celles qui se situent en dehors de ces limites.

IQR (Interquartile Range): Exemple

- Le jeu de données comporte 11 valeurs. Il contient quelques valeurs extrêmes; vous utiliserez donc la méthode IQR pour vérifier s'il s'agit de valeurs aberrantes.

26, 37, 24, 28, 35, 22, 31, 53, 41, 64, 29

IQR (Interquartile Range): Exemple

- 1. Triez les données du plus bas au plus haut.**
2. Identifiez le premier quartile (Q1), la médiane et le troisième quartile (Q3).
3. Calculez l'IQR = $Q3 - Q1$.
4. Calculez la limite supérieure = $Q3 + (1.5 * IQR)$.
5. Calculez la limite inférieure = $Q1 - (1.5 * IQR)$.
6. Utilisez les limites pour mettre en évidence les valeurs aberrantes, c'est-à-dire celles qui se situent en dehors de ces limites.

22, 24, 26, 28, 29, 31, 35, 37, 41, 53, 64

IQR (Interquartile Range): Exemple

1. Triez les données du plus bas au plus haut.
2. **Identifiez le premier quartile (Q1), la médiane et le troisième quartile (Q3).**
3. Calculez l'IQR = $Q3 - Q1$.
4. Calculez la limite supérieure = $Q3 + (1.5 * IQR)$.
5. Calculez la limite inférieure = $Q1 - (1.5 * IQR)$.
6. Utilisez les limites pour mettre en évidence les valeurs aberrantes, c'est-à-dire celles qui se situent en dehors de ces limites.

22, 24, **26**, 28, 29, **31**, 35, 37, **41**, 53, 64

↑ ↑ ↑
First quartile **Second quartile** **Third quartile**
(Q1 or lower quartile) (Q2 or median) (Q3 or upper quartile)

IQR (Interquartile Range): Exemple

1. Triez les données du plus bas au plus haut.
2. Identifiez le premier quartile (Q1), la médiane et le troisième quartile (Q3).
3. **Calculez l'IQR = Q3 - Q1.**
4. Calculez la limite supérieure = $Q3 + (1.5 * IQR)$.
5. Calculez la limite inférieure = $Q1 - (1.5 * IQR)$.
6. Utilisez les limites pour mettre en évidence les valeurs aberrantes, c'est-à-dire celles qui se situent en dehors de ces limites.

$$IQR = Q3 - Q1$$

$$Q1 = 26$$

$$Q3 = 41$$

$$IQR = 41 - 26 = 15$$

IQR (Interquartile Range): Exemple

1. Triez les données du plus bas au plus haut.
2. Identifiez le premier quartile (Q1), la médiane et le troisième quartile (Q3).
3. Calculez l'IQR = $Q3 - Q1$.
4. **Calculez la limite supérieure = $Q3 + (1.5 * IQR)$.**
5. **Calculez la limite inférieure = $Q1 - (1.5 * IQR)$.**
6. Utilisez les limites pour mettre en évidence les valeurs aberrantes, c'est-à-dire celles qui se situent en dehors de ces limites.

$$\begin{aligned}\text{Limite supérieure} &= Q3 + (1.5 * IQR) \\ &= 41 + (1.5 * 15) = 41 + 22.5 = 63.5\end{aligned}$$


$$\begin{aligned}\text{Limite inférieure} &= Q1 - (1.5 * IQR) \\ &= 26 - (1.5 * 15) = 26 - 22.5 = 3.5\end{aligned}$$

IQR (Interquartile Range): Exemple

1. Triez les données du plus bas au plus haut.
2. Identifiez le premier quartile (Q1), la médiane et le troisième quartile (Q3).
3. Calculez l'IQR = $Q3 - Q1$.
4. Calculez la limite supérieure = $Q3 + (1.5 * IQR)$.
5. Calculez la limite inférieure = $Q1 - (1.5 * IQR)$.
6. **Utilisez les limites pour mettre en évidence les valeurs aberrantes, c'est-à-dire celles qui se situent en dehors de ces limites.**

Limite supérieure = 63.5

Limite inférieure = 3.5

22, 24, 26, 28, 29, 31, 35, 37, 41, 53, **64**  outlier

Traitement des outliers

- **Vérifier leur origine**
 - Erreur de saisie?
 - Mesure exceptionnelle mais valide ?
- **Options de traitement :**
 - Corriger (si erreur évidente).
 - Supprimer (si aberration non représentative).
 - Garder (si l'outlier est pertinent pour l'analyse).

Exemple concret : Températures quotidiennes

- Relevés sur une ville en été :
29°C, 31°C, 30°C, 32°C, 28°C, 33°C, 45°C
- La majorité des valeurs se situe entre **28°C** et **33°C**, sauf un jour à **45°C**.
- Cette valeur peut être :
 - Une **erreur de capteur** (et donc corrigée ou supprimée).
 - Un **événement climatique exceptionnel** (Vague de chaleur réelle, à conserver).

Normalisation et standardisation

- Normalisation : transformation des données pour les ramener dans une échelle fixe (souvent $[0,1]$).
- Standardisation : transformation des données pour qu'elles aient une moyenne **nulle** et un écart-type de **1**.

Pourquoi les utiliser ?

- Éviter que des variables à grandes valeurs (ex. revenus en MD) dominant celles à petites valeurs (ex. âge).
- Améliorer la performance des algorithmes sensibles aux échelles (k-means, SVM, régression logistique, réseaux de neurones).
- Faciliter la comparaison entre variables hétérogènes.

Méthodes

- **Normalisation Min-Max**

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

- Valeurs transformées entre 0 et 1.

- **Standardisation (Z-score)**

$$Z = \frac{x - \mu}{\sigma}$$

- Moyenne = 0, écart-type = 1.

Exemple

Client	Âge	Revenu (TND)
1	20	1 000
2	25	2 000
3	30	5 000
4	35	8 000
5	40	20 000
6	60	50 000

- Sans transformation: les revenus dominant l'âge dans l'analyse.
- Après **normalisation**: chaque variable est ramenée sur la même échelle (0 à 1).
- Après **standardisation**: les deux variables deviennent comparables avec une moyenne et un écart-type normalisés.

Exemple: Normalisation (min-max)

- Âge : min = 20, max = 60
- Revenu : min = 1 000, max = 50 000
 - **Âges normalisés :**
 - $20 \rightarrow (20-20)/(60-20) = 0$
 - $25 \rightarrow (25-20)/40 = 0,125$
 - $30 \rightarrow 0,25$
 - $35 \rightarrow 0,375$
 - $40 \rightarrow 0,5$
 - $60 \rightarrow 1$
 - **Revenus normalisés :**
 - $1\ 000 \rightarrow (1\ 000-1\ 000)/49\ 000 = 0$
 - $2\ 000 \rightarrow (2\ 000-1\ 000)/49\ 000 \approx 0,0204$
 - $5\ 000 \rightarrow (5\ 000-1\ 000)/49\ 000 \approx 0,0816$
 - $8\ 000 \rightarrow 0,1429$
 - $20\ 000 \rightarrow 0,3878$
 - $50\ 000 \rightarrow 1$

Exemple: Standardisation (avg = 0, sd = 1)

- Calcul de la moyenne et écart-type :
- Âge :
 - $\mu = (20+25+30+35+40+60)/6 = 35$
 - $\sigma = \sqrt{[(20-35)^2 + \dots + (60-35)^2]/6}$
 $= \sqrt{[(225+100+25+0+25+625)/6]} = \sqrt{(1000/6)} \approx 12,91$
- Revenu :
 - $\mu = (1\,000+2\,000+5\,000+8\,000+20\,000+50\,000)/6 = 14\,333,33$
 - $\sigma \approx \sqrt{[(1\,000-14\,333,33)^2 + \dots + (50\,000-14\,333,33)^2]/6]} \approx 18\,621,9$

Exemple: Standardisation (avg = 0, sd = 1)

- **Âges standardisés :**

- $20 \rightarrow (20-35)/12,91 \approx -1,16$
- $25 \rightarrow -0,78$
- $30 \rightarrow -0,39$
- $35 \rightarrow 0$
- $40 \rightarrow 0,39$
- $60 \rightarrow 1,94$

- **Revenus standardisés :**

- $1\ 000 \rightarrow (1\ 000-14\ 333,33)/18\ 621,9 \approx -0,71$
- $2\ 000 \rightarrow -0,66$
- $5\ 000 \rightarrow -0,49$
- $8\ 000 \rightarrow -0,34$
- $20\ 000 \rightarrow 0,31$
- $50\ 000 \rightarrow 1,95$

Exemple

Client	Âge (brut)	Âge (norm)	Âge (std)	Revenu (TND) (brut)	Revenu (norm)	Revenu (std)
1	20	0	-1,16	1 000	0	-0,71
2	25	0,125	-0,78	2 000	0,0204	-0,66
3	30	0,25	-0,39	5 000	0,0816	-0,49
4	35	0,375	0	8 000	0,1429	-0,34
5	40	0,5	0,39	20 000	0,3878	0,31
6	60	1	1,94	50 000	1	1,95

Vérification de la cohérence et des unités

- **Pourquoi vérifier ?**
 - Des données **syntactiquement valides** peuvent être **logiquement fausses**.
- Les incohérences ou erreurs d'unités → biais et mauvaise interprétation.
- Exemple :
 - Âge = 200 ans
 - Prix en € et en \$ mélangés

Vérification de la cohérence

- Détection des valeurs impossibles :
 - Âge < 0 ou > 120 ans
 - Pourcentage $> 100\%$
 - Dates inversées (fin $<$ début)
- Contrôles de dépendances :
 - Cohérence entre plusieurs variables
- Outils : règles métiers, scripts de validation

Vérification des unités

- Problèmes fréquents :
 - Poids en **kg** vs **g**
 - Distance en **miles** vs **km**
 - Devises différentes (€, \$, £)
- Solutions :
 - Conversion vers une **unité standardisée**
 - Documentation des unités (métadonnées)

Vérification de la cohérence et des unités

- Établir des **règles de validation** (plages, dépendances logiques).
- Standardiser les **unités de mesure**.
- Vérifier régulièrement la cohérence pour éviter des biais.
- Documenter les unités et transformations appliquées.

Vérification de la cohérence et des unités

Exemple

Client	Âge (brut)	Âge corrigé	Poids (brut)	Poids corrigé	Montant (brut)	Montant corrigé
1	150 ans ✗	50 ans ✓	70 kg	70 kg	200 €	200 €
2	-5 ans ✗	5 ans ✓	70 000 g ✗	70 kg ✓	300 \$ ✗	280 € ✓ (converti)
3	35 ans	35 ans	80 kg	80 kg	500 €	500 €

2. 3. Transformation des données

Transformation des données

- Les données brutes ne sont pas toujours prêtes à être utilisées.
- La transformation permet :
 - De rendre les variables **compatibles avec les algorithmes**.
 - De créer de **nouvelles informations utiles**.
 - De **synthétiser** les données pour l'analyse.
- Principales étapes :
 - Encodage des variables catégorielles
 - Ingénierie des variables (Feature Engineering)
 - Agrégation et filtrage

Encodage des variables catégorielles

- Les algorithmes nécessitent des données **numériques**.
- Les variables qualitatives doivent être **encodées**.
- Deux approches principales :
 - **Label Encoding**
 - **One-Hot Encoding**

Label Encoding

- Associer un entier à chaque modalité.
- Exemple :
 - Couleur : Rouge = 0, Vert = 1, Bleu = 3 etc...
 - Réponse : Oui = 1, Non = 0
- Avantages : simple, efficace pour binaires.
- Inconvénients : introduit parfois un **ordre artificiel**.

Exemple de Label Encoding

Client	Couleur (brute)	Couleur (encodée)
1	Rouge	0
2	Vert	1
3	Bleu	2
4	Vert	1
5	Bleu	2
6	Noir	3

One-Hot Encoding

- Crée une colonne binaire pour chaque modalité.
- Exemple : Couleur = {Rouge, Bleu, Vert}
 - Rouge = [1,0,0]
 - Bleu = [0,1,0]
 - Vert = [0,0,1]
- Avantages : pas d'ordre artificiel.
- Inconvénients : explosion du nombre de colonnes.

Exemple de One-Hot Encoding

Produit	Couleur (brute)	Rouge	Bleu	Vert
A	Rouge	1	0	0
B	Vert	0	0	1
C	Bleu	0	1	0

Ingénierie des variables (Feature Engineering)

- Objectif: enrichir le dataset avec de nouvelles informations.
- Deux approches :
 - **Création de nouvelles variables.**
 - **Combinaison de variables existantes.**

Création de nouvelles variables

- Exemple 1 :

$\hat{\text{Âge}} = \text{Date_actuelle} - \text{Date_naissance}.$

- Exemple 2 :

$\text{Durée_abonnement} = \text{Date_fin} - \text{Date_début}.$

- Exemple 3 :

Année, Mois, Jour extraits d'une date complète.

Combinaison de variables

- Exemple 1 :

Ratio \rightarrow Revenu / Nombre de personnes dans le foyer.

- Exemple 2 :

Différence \rightarrow Score_initial – Score_final.

- Exemple 3 :

Interaction \rightarrow Taille \times Poids \rightarrow IMC.

Agrégation et filtrage

- **Agrégation** : regrouper les données pour les résumer.
 - Ventes totales par mois.
 - Âge moyen par département.
- **Filtrage** : sélectionner les données pertinentes.
 - Conserver uniquement les clients actifs.
 - Filtrer les ventes > 100 TND.

Exemple d'agrégation

Région	Ventes	Total
Nord	$200 + 150 + 300$	650
Sud	$400 + 250$	650
Est	$500 + 100 + 200$	800
Ouest	$300 + 200$	500

Résumé des ventes par **région** au lieu de ventes par transaction.

Transformation des données

- L'encodage transforme le qualitatif en numérique.
- Le feature engineering permet d'ajouter de la valeur.
- L'agrégation et le filtrage facilitent la **synthèse** et la **compréhension** des données.