



Analyse de Données



Akil ELKAMEL

akil.elkamel@isimm.rnu.tn

Plan du cours

- Fondements de l'analyse de données
- Collecte et préparation des données
- Analyse exploratoire des données
- Analyse statistique et inférence
- Méthodes analytiques avancées

Chapitre 1:

Fondements de l'analyse de données

1.1. Introduction à l'analyse de données

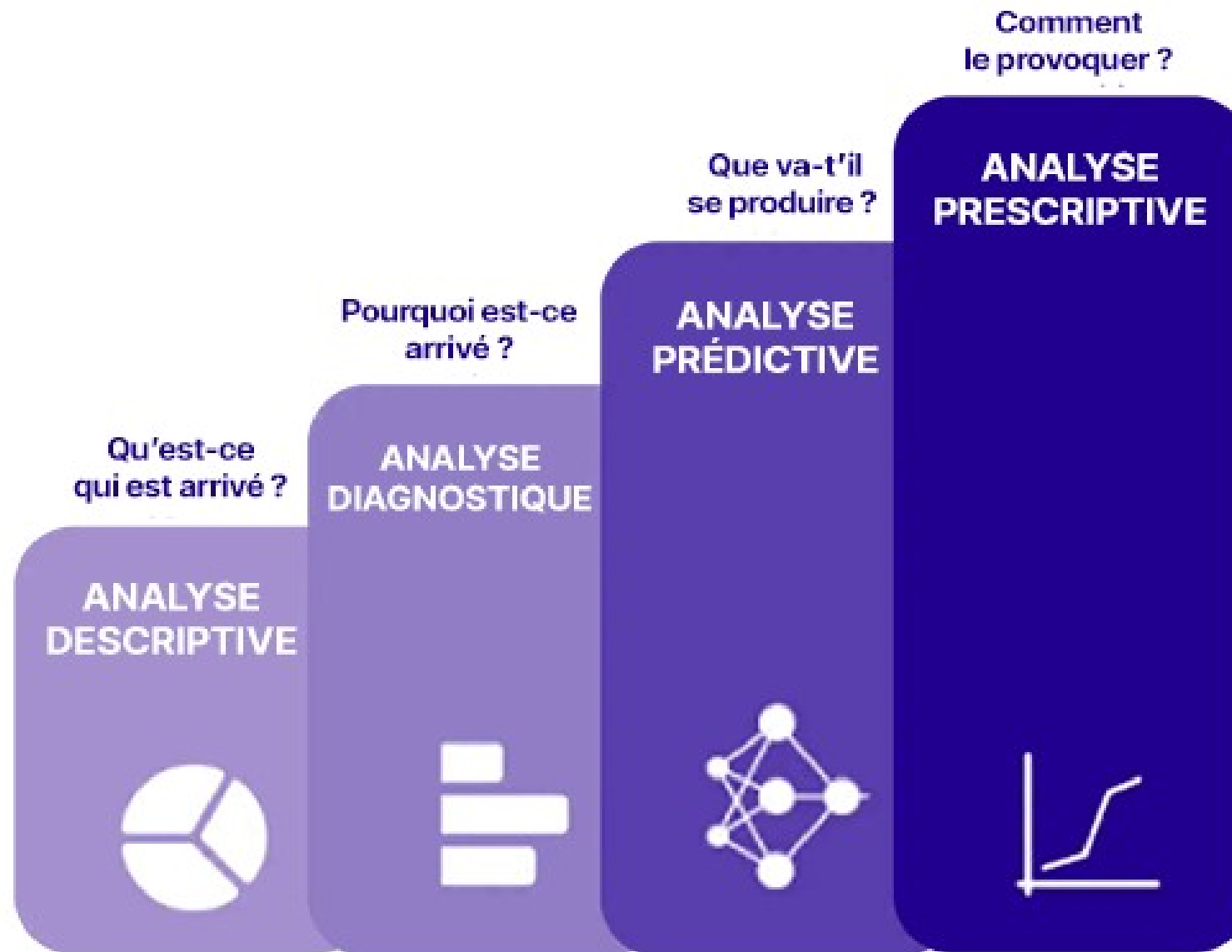
C'est quoi l'analyse de données?

- L'analyse de données est un processus **méthodique** et **structuré** qui consiste à **collecter**, **nettoyer**, **transformer** et **examiner** des ensembles de données dans le but d'identifier des tendances, d'évaluer des hypothèses, de répondre à des questions spécifiques et de soutenir la prise de décision.





Les différents types d'analyse de données

- Permettent de tirer des connaissances ou informations des données de différentes manières.
- Chacun de ces types d'analyse a une approche unique pour explorer les données et comprendre les tendances et les modèles cachés

Les différents types d'analyse de données



Les différents types d'analyse de données

-  Descriptive (que se passe-t-il ?)
 - Décrit les caractéristiques des données telles qu'elles sont.
-  Diagnostique (pourquoi ?)
 - Tente de comprendre les raisons sous-jacentes à un phénomène observé dans les données.
-  Prédictive (que va-t-il se passer ?)
 - Utilise les données historiques pour faire des prévisions sur les événements futurs.
-  Prescriptive (que faut-il faire ?)
 - Utilise les données et les algorithmes pour recommander des actions pour atteindre un objectif spécifique.
- Ces différents types d'analyse de données peuvent être utilisés seuls ou en combinaison pour aider à comprendre les données et à prendre des décisions basées sur les données (Data Driven Decisions)

L'analyse descriptive



-
- L'analyse descriptive est le niveau le plus élémentaire de l'analyse de données. Elle consiste à décrire et à résumer les données de manière factuelle, claire et concise.
 - L'objectif principal de cette analyse est de donner un aperçu des événements passés et des tendances actuelles :
 - Statistiques de base : calculs de moyenne, médiane, mode, écart-type, etc.
 - Graphiques et tableaux : représentations visuelles des données, telles que les histogrammes, les diagrammes en barres, les camemberts, etc.
 - Segmentations : divisions des données en groupes homogènes pour mieux comprendre les caractéristiques de chaque groupe.

L'analyse diagnostique



-
- L'analyse diagnostique vise à comprendre pourquoi certains événements se sont produits en identifiant les facteurs ou les causes qui les ont influencés. Elle s'appuie sur les résultats de l'analyse descriptive pour approfondir la compréhension des données :
 - Analyse de corrélation : identification des relations entre les différentes variables pour comprendre comment elles interagissent les unes avec les autres.
 - Analyse de régression : évaluation de la relation entre une variable dépendante et une ou plusieurs variables indépendantes pour expliquer les variations dans les données.
 - Analyse de causalité : détermination des facteurs qui ont conduit à un événement spécifique.

L'analyse prédictive



- L'analyse prédictive consiste à utiliser des modèles statistiques et des algorithmes pour anticiper les comportements futurs des clients et les tendances du marché, en se basant sur les données passées et actuelles. Cela permet aux entreprises de prendre des décisions anticipées, d'optimiser les campagnes et de cibler efficacement les clients potentiels :
 - Modélisation prédictive : utilisation de modèles mathématiques pour prédire des résultats futurs, tels que la prédiction des ventes, des tendances du marché, etc.
 - Analyse de séries temporelles : prévision des valeurs futures en se basant sur les variations historiques dans le temps.
 - Analyse de clustering : regroupement de données similaires pour identifier les schémas et les tendances émergentes.



-
- L'analyse prescriptive va au-delà de la prédiction en proposant des solutions et des recommandations pour atteindre des objectifs spécifiques. Elle aide les entreprises à choisir la meilleure approche à suivre en fonction des résultats des analyses précédentes :
 - Optimisation : identification de la meilleure combinaison de variables pour atteindre un objectif défini, par exemple, l'optimisation des dépenses publicitaires pour maximiser les conversions.
 - Scénarios et simulations : évaluation des résultats possibles en fonction de différentes stratégies pour prendre des décisions éclairées.
 - Systèmes experts : utilisation de règles et d'algorithmes pour fournir des recommandations en temps réel.

Les différents types d'analyse de données

- Avec l'essor récent de l'intelligence artificielle, les analyses prédictives et prescriptives sont de plus en plus performantes et constituent un réel atout pour les entreprises qui s'appuient dessus.
- En combinant ces quatre catégories d'analyse de données, les entreprises peuvent tirer le meilleur parti de leurs données.

Les différents types d'analyse de données

- Quelques exemples concrets de décisions basées sur les données (Data Driven Decisions) guidées par ces différents types d'analyse de données :
 - **Pour une campagne marketing** : l'analyse descriptive peut informer que les emails du mercredi performant mieux. Vous ajusterez donc la planification en fonction.
 - **Pour le taux de conversion** : une analyse diagnostique va permettre de révéler que les abandons de panier surviennent surtout de l'application mobile. Une refonte UX ciblée pourra être initiée pour résoudre ce problème et optimiser cet indicateur.
 - **Pour la prévision des ventes** : l'analyse prédictive permettra d'anticiper la demande et d'avoir une meilleure gestion des stocks.
 - **Pour l'optimisation du pricing** : les modèles d'analyse prescriptive vont identifier le bon prix selon la saison, la demande et le profil client.

Intérêt de l'analyse de données

- Aide à la prise de décision
 - Fournit des informations objectives pour orienter les choix stratégiques, scientifiques ou opérationnels.
- Identification de tendances et de modèles
 - Permet de découvrir des relations cachées, des corrélations et des comportements récurrents.
- Amélioration de la performance
 - Optimise les processus dans les entreprises, la recherche, la santé, l'éducation, etc.
- Prédiction et anticipation
 - Utilisée pour prévoir des évolutions futures (ex. : ventes, risques financiers, propagation d'une maladie).

Intérêt de l'analyse de données

- Résolution de problèmes complexes
 - Aide à comprendre les causes profondes de phénomènes et à proposer des solutions efficaces.
- Innovation et compétitivité
 - Alimente le développement de nouveaux produits, services et modèles économiques.
- Validation scientifique
 - Soutient la recherche académique en confirmant ou en infirmant des hypothèses à partir de données empiriques.
- Amélioration de la qualité des données
 - Le processus de nettoyage et de vérification permet d'assurer la fiabilité des informations utilisées.

Applications de l'analyse de données

- Santé
- Finance et économie
- Commerce et marketing
- Industrie et ingénierie
- Sciences sociales et politiques publiques
- Environnement et énergie
- Sports et divertissement

Santé

- Diagnostic médical assisté
- Médecine personnalisée selon le profil patient
- Suivi et détection des épidémies



Finance et économie

- Prévisions économiques et boursières
- Détection de fraudes et anomalies de transactions
- Analyse de risques et scoring crédit



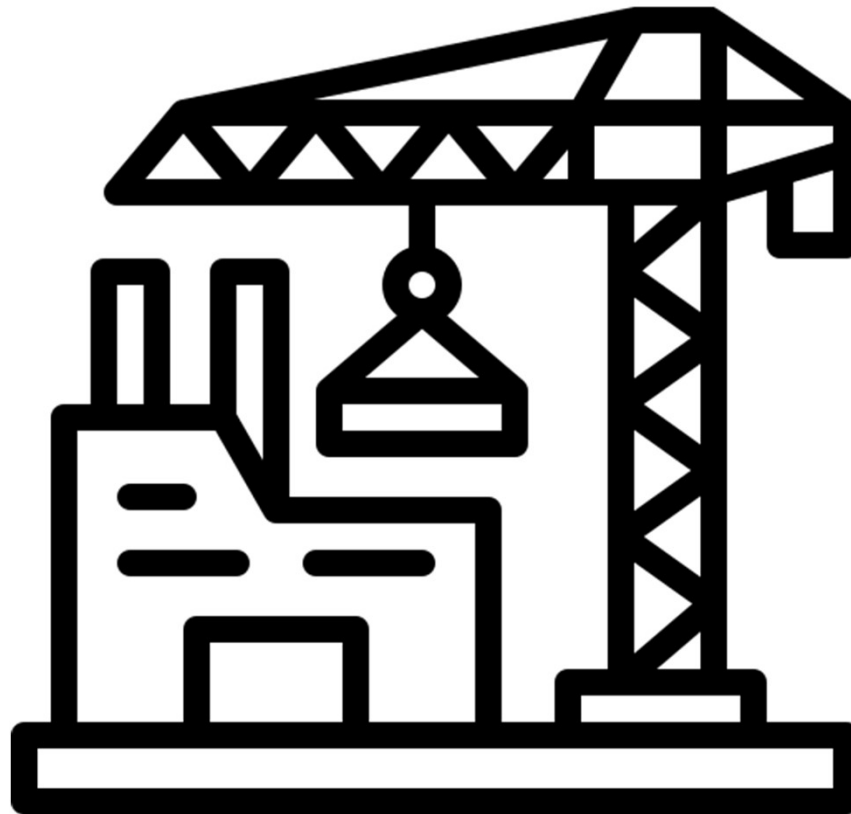
Commerce et marketing

- Segmentation des clients
- Personnalisation des recommandations
- Optimisation des campagnes publicitaires



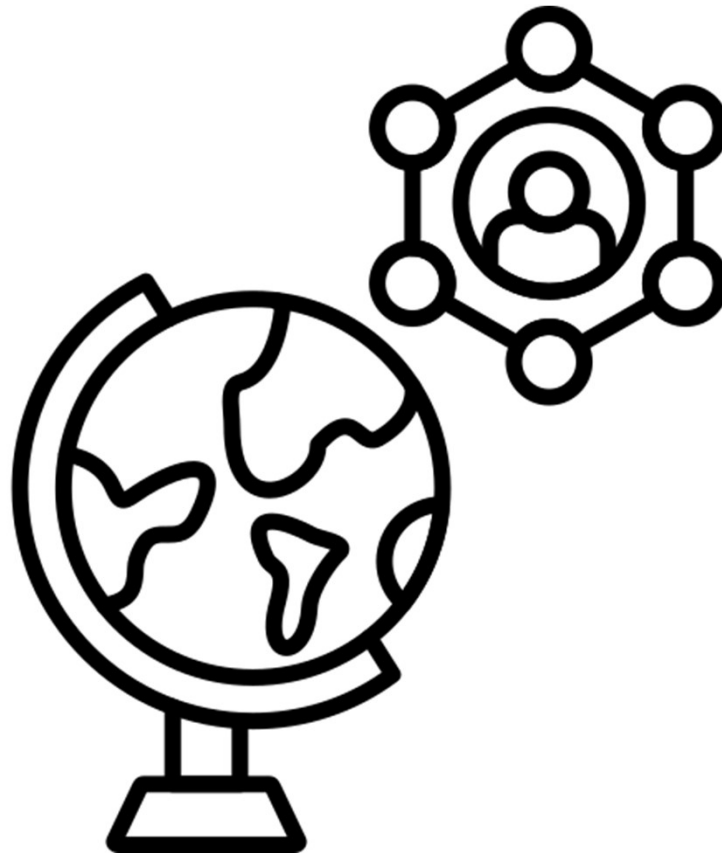
Industrie et ingénierie

- Maintenance prédictive (anticiper les pannes)
- Optimisation logistique et chaîne d'approvisionnement
- Contrôle qualité automatisé



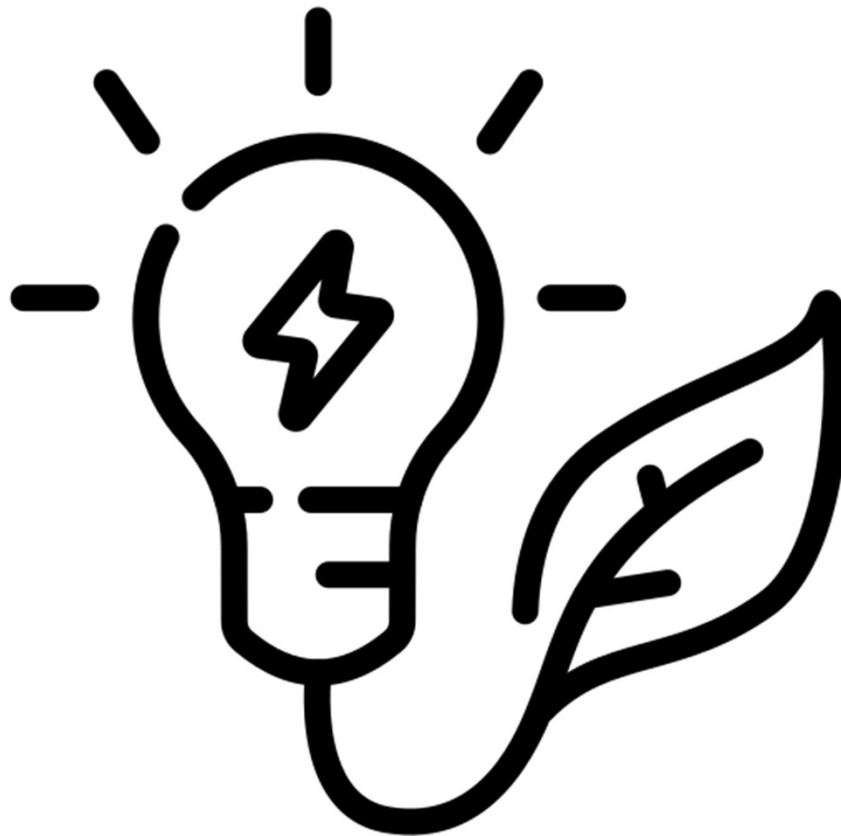
Sciences sociales et politiques publiques

- Analyse des enquêtes et sondages
- Planification de politiques publiques
- Études démographiques



Environnement et énergie

- Prévisions météorologiques & modélisation climat
- Smart cities : optimisation énergétique
- Suivi biodiversité & déforestation



Sports et divertissement

- Analyse des performances sportives
- Optimisation des stratégies d'équipe
- Personnalisation des contenus de divertissement



Les étapes de l'analyse de données



Spécifier le
problème



Collecte de
données



Nettoyage et trie
des données
collectées



Analyse des
données
collectées



Présentation
des résultats



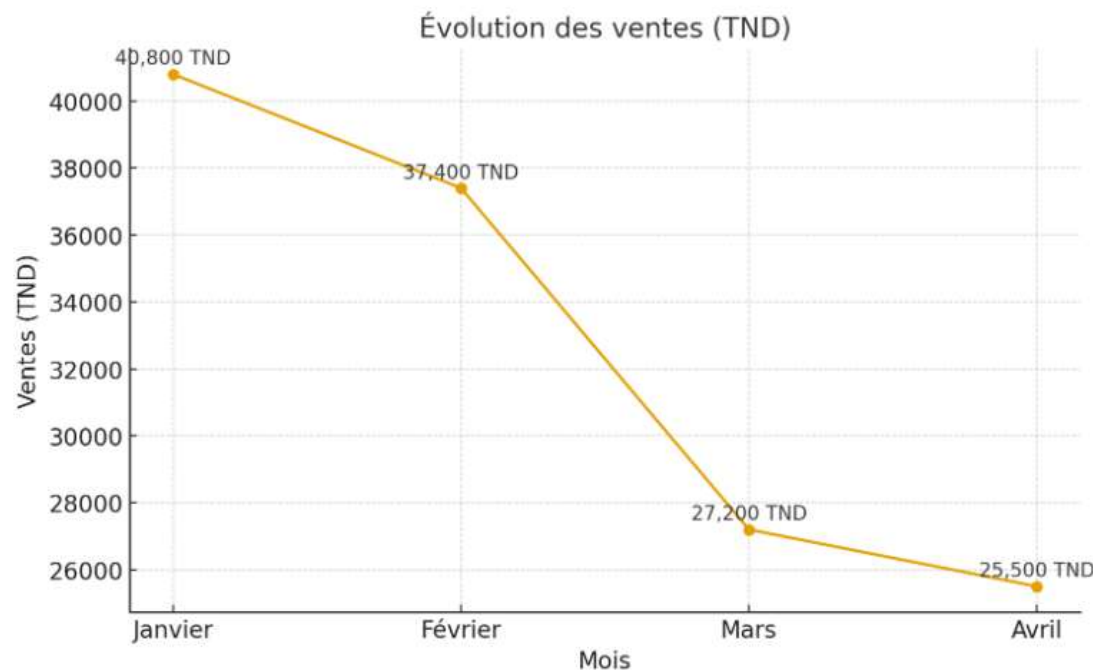
La prise de
décision

Exemple : Améliorer les ventes d'une boutique en ligne de vêtements

Spécifier le problème



- Définir l'objectif et formuler une question claire.
- Identifier les indicateurs clés à analyser.



- Dans notre cas:
 - La boutique constate une baisse des ventes depuis 3 mois.
 - La question devient : "*Quels sont les facteurs qui expliquent la baisse des ventes et comment y remédier ?*"



Collecte de données

- Identifier les sources de données.
- Rassembler les données nécessaires.
- Dans notre cas, on collecte :
 - Données de ventes (produits, prix, date, quantités).
 - Données clients (âge, sexe, localisation).
 - Données marketing (campagnes publicitaires, promotions).
 - Données externes (concurrence, tendances saisonnières).

Date	Produit	Catégorie	Prix (TND)	Qté vendue	Client âge	Sexe	Campagne pub
2025-02-01	T-shirt A	Homme	60	2	25	M	Facebook Ads
2025-02-03	Robe B	Femme	135	1	32	F	Google Ads
2025-02-05	Jean C	Homme	100	3	28	M	Instagram

Nettoyage et trie des données collectées



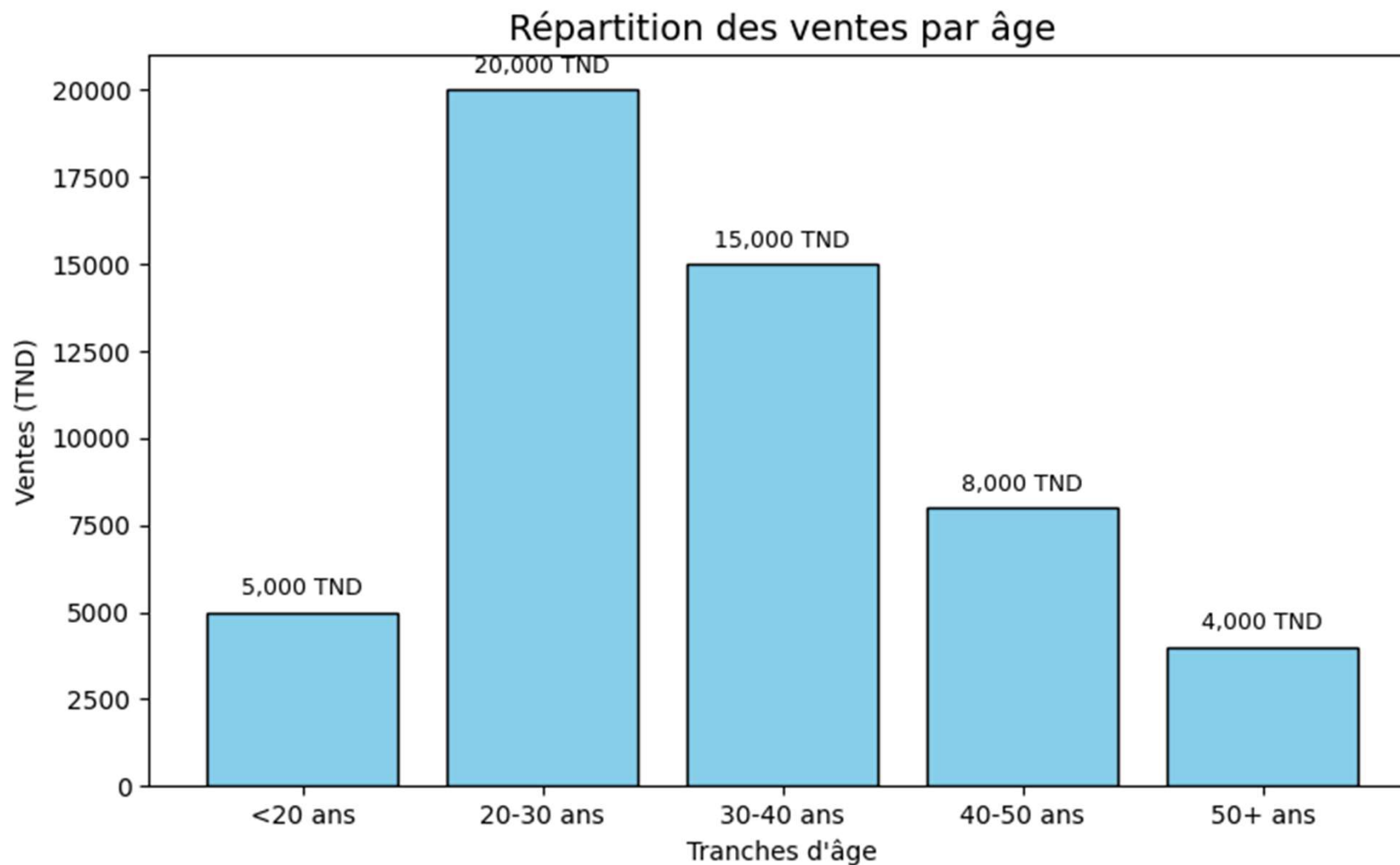
- Supprimer les doublons.
- Remplacer ou gérer les valeurs manquantes.
- Uniformiser les formats (dates, devises, catégories).
- Détecter les valeurs aberrantes (ex. prix = -50 TND).
- Dans notre cas :
 - Suppression des transactions dupliquées.
 - Correction de dates mal saisies ("2025-13-01").
 - Remplissage des âges manquants avec une estimation (moyenne par segment).
- Avant nettoyage :
`2025-13-01, Jean C, Homme, 100, 2, ?, M, Facebook Ads`
- Après nettoyage :
`2025-01-13, Jean C, Homme, 100, 2, 29, M, Facebook Ads`

Analyse des données collectées

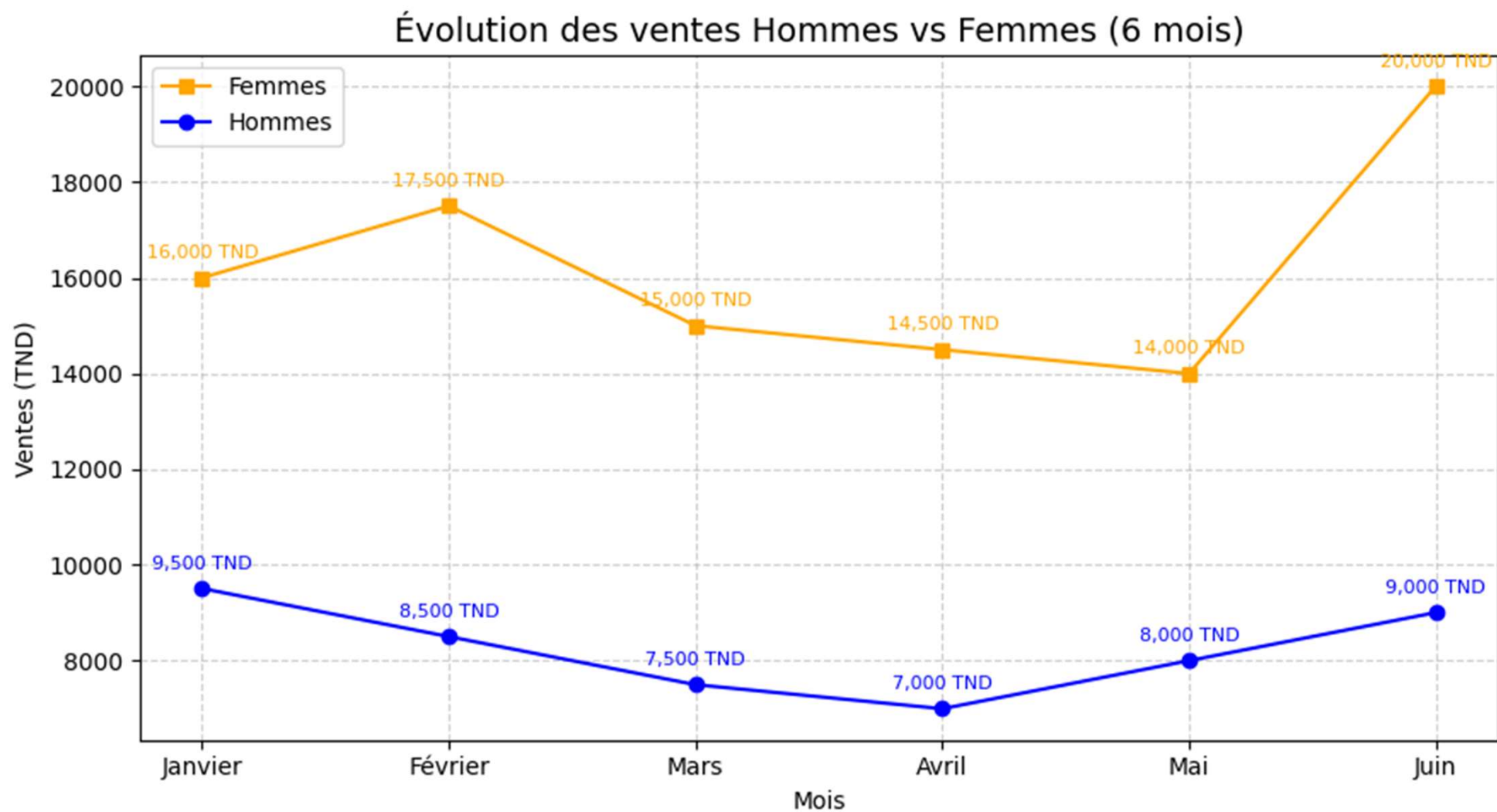


- Décrire les données (statistiques descriptives).
- Identifier des relations (corrélations, régressions, clustering).
- Répondre à la question posée.
- Dans notre cas :
 - 65% des ventes concernent les femmes.
 - Les clients 20-30 ans sont majoritaires.
 - Forte baisse sur les articles masculins.
 - Les campagnes Google Ads génèrent +40% de ventes.
 - Instagram = presque 0 impact.

Analyse des données collectées



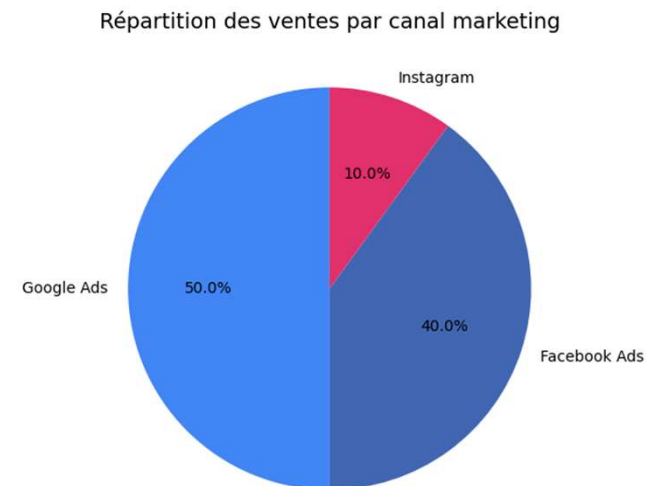
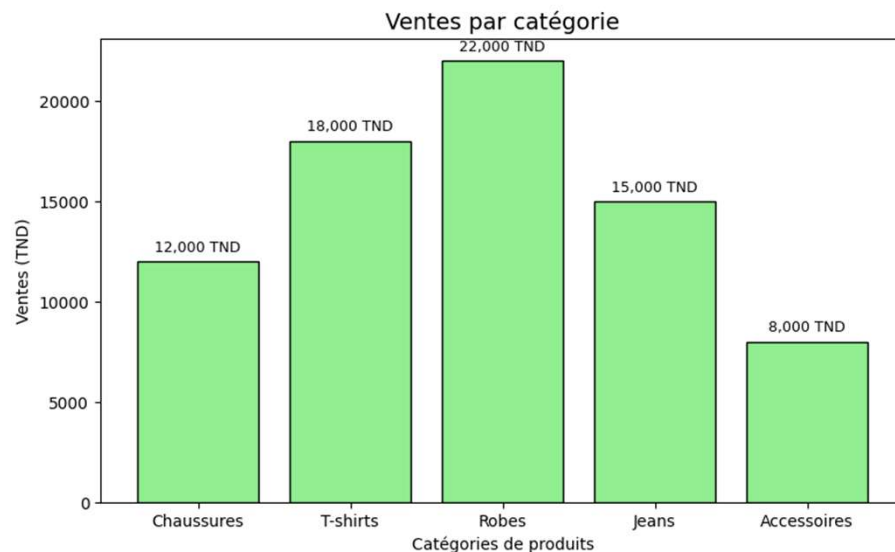
Analyse des données collectées



Présentation des résultats



- Visualiser les données (graphiques, dashboards).
- Rédiger un rapport clair, compréhensible par les décideurs.
- Dans notre cas :
 - Diagramme en barres : ventes par catégorie.
 - Pie chart : répartition des ventes par canal marketing.
 - Tableau récapitulatif avec recommandations.



La prise de décision



- Transformer l'analyse en actions concrètes.
- Définir un plan d'action mesurable (KPIs).
- Dans notre cas :
 - Réduire le budget marketing sur Instagram.
 - Renforcer les promotions sur les articles masculins.
 - Développer de nouvelles campagnes Google Ads ciblant les 20-30 ans.
- Décision finale :
 - **+20%** budget Google Ads
 - Nouvelle collection "homme printemps"
 - Objectif : **+15% ventes en 3 mois**

Analyse de données vs. statistique vs. science des données

Analyse de données

- **Définition** : Processus d'examen, de transformation et d'interprétation des données pour répondre à des questions spécifiques et appuyer la prise de décision.
- **Objectif** : Extraire des informations utiles d'un jeu de données particulier.
- **Exemple** : Analyser les ventes mensuelles d'un magasin pour comprendre les variations saisonnières.

Analyse de données vs. statistique vs. science des données

Statistique

- **Définition** : *«Le mot statistique désigne à la fois un ensemble de données d'observations et l'activité qui consiste dans leur recueil, leur traitement et leur interprétation»* (Encyclopedia Universalis). C'est une discipline mathématique qui développe des méthodes pour collecter, organiser, analyser et interpréter des données.
- **Objectif** : Fournir une base théorique et des outils rigoureux pour décrire des populations, tester des hypothèses et modéliser des phénomènes.
- **Exemple** : Utiliser un test d'hypothèse pour déterminer si une nouvelle méthode d'enseignement améliore significativement les résultats des étudiants.

Analyse de données vs. statistique vs. science des données

Science des données

- **Définition** : Domaine interdisciplinaire qui combine statistique, informatique, apprentissage automatique et expertise métier pour extraire de la connaissance et construire des modèles prédictifs ou prescriptifs à partir de données massives (big data).
- **Objectif** : Automatiser et optimiser l'exploitation des données pour générer des solutions innovantes et intelligentes.
- **Exemple** : Développer un algorithme de recommandation (Netflix, Amazon) en utilisant des millions de données utilisateurs.

Analyse de données vs. statistique vs. science des données

Aspect	Analyse de données	Statistique	Science des données
Nature	Application pratique	Discipline théorique et appliquée	Domaine interdisciplinaire
Objectif	Répondre à une question précise	Décrire, tester, modéliser	Découvrir, prédire, optimiser
Outils	Tableurs, visualisations, tests simples	Méthodes mathématiques, inférence	Statistique + ML + Big Data
Exemple	Résumer les ventes d'une année	Vérifier si une différence est significative	Construire un modèle de prévision de ventes

1.2. Types de données

Données (Data)

- Est une collection d'objets avec leurs attributs ou caractéristiques.

Attributs

		variable 1	...	variable j	...	variable p
Objets	individu 1	x_{11}		x_{1j}		x_{1p}
	individu i	x_{i1}		x_{ij}		x_{ip}
				.		.
	individu n	x_{n1}		x_{nj}		x_{np}

$\mathbf{X} = \{\mathbf{x}_{ij}\}$ est une matrice avec n lignes and p colonnes

-

-

-
- A complex network graph visualization. It features a central node colored red, which is highly connected to many other nodes. These peripheral nodes are represented by small grey and blue spheres. The connections between nodes are shown as a dense web of thin yellow lines. The overall structure is roughly circular, with the highest density of connections and nodes in the center, tapering off towards the edges.

-
- A photograph of a busy pedestrian street in Beijing, China. The street is wide and paved with light-colored tiles. On the left, a person is riding a bicycle. In the center, a group of people are walking, including a man in a white shirt and dark pants, and a woman in a red jacket. On the right, there is a row of white, curved structures that look like modern benches or barriers. The background shows trees and buildings, suggesting an urban environment.

-

-

Types de données

- Les données quantitatives ou numériques
- Les données qualitatives ou catégoriques

Les données quantitatives ou numériques

- **Les données quantitatives (ou numériques)** sont des données mesurables qui expriment des quantités. Elles représentent des valeurs numériques sur lesquelles on peut effectuer des opérations mathématiques (addition, soustraction, moyenne, variance, etc.).
- Elles se divisent généralement en deux sous-catégories :
 - **Données discrètes** : prennent uniquement des valeurs entières (ex. nombre d'enfants, nombre de ventes).
 - **Données continues** : peuvent prendre une infinité de valeurs dans un intervalle (ex. taille, poids, température).

Les données quantitatives continues

- Une donnée quantitative est dite continue lorsqu'elle prend un nombre infini de valeurs réelles à l'intérieur d'un intervalle donné.
- La taille d'une personne est un exemple de données quantitatives continues. Même si elle ne peut pas prendre toutes les valeurs réelles possibles, elles peuvent prendre une infinité de valeurs dans un intervalle défini selon l'objet mesuré. Le poids d'une personne, la hauteur d'un immeuble sont également des exemples de données quantitatives continues.
- Entre deux valeurs de poids par exemple, il y a des millions de poids possibles.
- En général, les données qui proviennent d'une mesure sont quantitatives.

Les données quantitatives discrètes

- On désigne par données quantitatives discrètes des données qui ne peuvent prendre qu'un nombre fini de valeurs réelles possibles au sein d'un intervalle donné.
- Le nombre d'employés d'une entreprise est également une donnée quantitative discrète. En prenant l'exemple des entreprises qui ont au plus 100 employés, le nombre de valeurs possibles prises par une telle variable ne peut bien évidemment pas excéder 100. On sait en effet qu'il est impossible pour une entreprise de disposer d'un nombre d'employés qui serait une fraction d'un nombre entier comme 72.3 par exemple.
- En général, les données discrètes sont des décomptes.

Les données quantitatives ou numériques

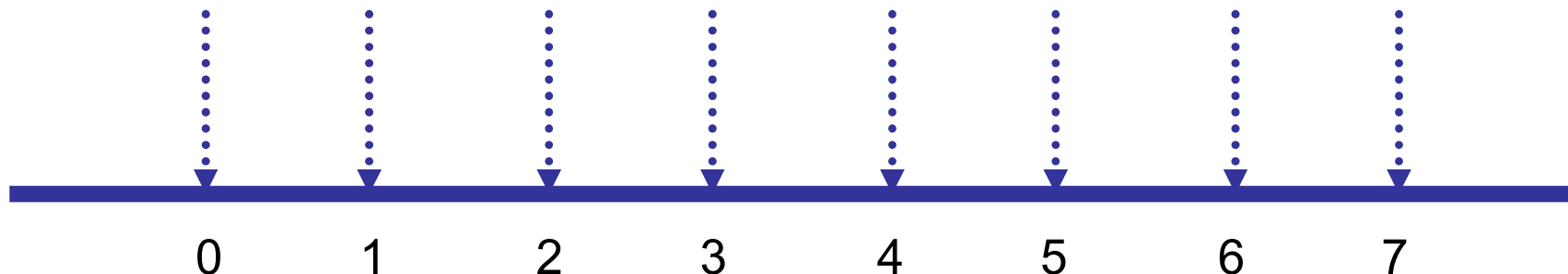
Données continues – Théoriquement,
aucun écart entre les valeurs possibles



0

1000

Données discrètes – Écarts entre les valeurs possibles



Les données qualitatives ou catégoriques

- **Les données qualitatives (ou catégoriques)** sont des données non numériques qui décrivent des caractéristiques, des qualités ou des catégories. Elles ne se prêtent pas directement aux opérations mathématiques (on ne peut pas faire une moyenne de couleurs ou de professions, par exemple), mais elles permettent de classer, de regrouper et de comparer des individus ou des objets selon leurs attributs.
- Elles se divisent en deux types principaux :
 - **Données nominales** : simples étiquettes sans ordre logique (ex. couleur des yeux, nationalité).
 - **Données ordinales** : catégories avec un ordre ou un classement (ex. niveaux de satisfaction : «faible», «moyen», «élevé» ; niveaux d'éducation).

Les données qualitatives nominales

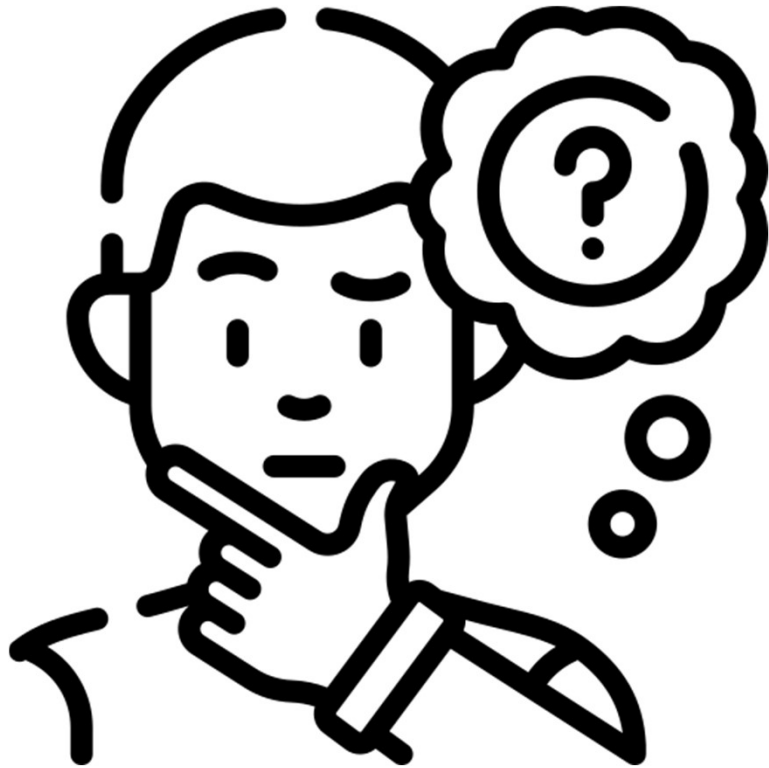
- Une donnée qualitative nominale décrit un nom ou une catégorie sans ordre particulier.
- Les données qualitatives nominales servent essentiellement pour étiqueter des variables.
- **Exemples:** Le mode de transport utilisé par les employés d'une entreprise, le sexe, la couleur des cheveux...

Les données qualitatives ordinales

- Une donnée qualitative ordinale est une donnée qui présente des valeurs définies par une relation d'ordre entre les différentes catégories possibles.
- L'appréciation des clients de la qualité des services d'une entreprise est un exemple de données qualitatives ordinales.
- Elle présente en effet des catégories comme «Mauvais», «Bon», «Très bon», «Excellent» entre lesquelles une relation évidente d'ordre peut être établie.
- La catégorie «Très bien» est meilleure que la catégorie «Bon», mais moins intéressante que la catégorie «Excellente».

Types de données

Type de données	Définition	Sous-types	Exemples
Quantitatives (numériques)	Données mesurables exprimées par des nombres. Permettent des calculs mathématiques (somme, moyenne, variance).	<ul style="list-style-type: none">- Discrètes : valeurs entières (ex. nombre d'enfants)- Continues : valeurs dans un intervalle infini (ex. taille, poids)	Âge, revenu, température, durée, nombre de ventes
Qualitatives (catégorielles)	Données descriptives qui expriment des catégories ou des attributs. Ne permettent pas directement de calculs numériques.	<ul style="list-style-type: none">- Nominales : catégories sans ordre (ex. couleur des yeux)- Ordinales : catégories avec ordre (ex. niveau d'éducation)	Profession, couleur, niveau de satisfaction, classe sociale



Les types de données collectées dans une étude déterminent le type d'analyse statistique utilisé.

Exemples

- Les données numériques sont généralement résumées à l'aide de « moyennes ».
 - Le nombre moyen d'élèves par classe est de 25.
 - Le poids moyen des hommes est de 85 kg.
 - Le poids moyen des femmes est de 67 kg.

Exemples

- Les données catégoriques sont généralement résumées sous forme de « pourcentages » (ou « proportions »).
 - 31 % des étudiants ont de très bonnes notes.
 - 28%, 33% et 39% des étudiants sont respectivement en première, deuxième et troisième année.