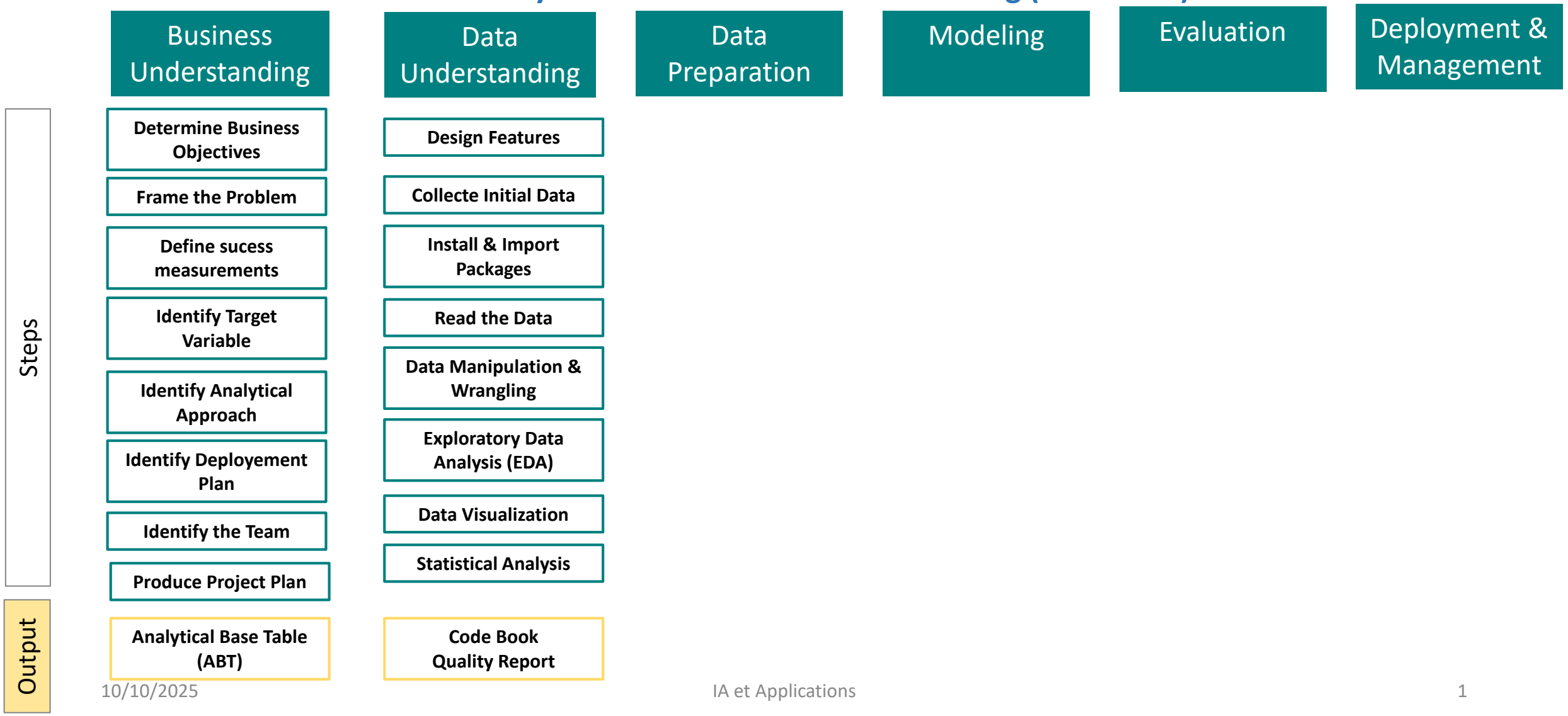
Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)



Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Objectifs du Data Understanding:

1- Se familiariser avec les données:

- Comprendre la **nature et la structure** du dataset.
- Identifier si les données sont **structurées, semi-structurées, ou non-structurées**.
- Reconnaître les types de variables : catégorielles (nominales, ordinales, binaires), numériques (discrètes, continues), temporelles.

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Objectifs du Data Understanding:

Données Structurées:

- Organisées dans des **formats tabulaires** (lignes, colonnes).
- Faciles à stocker dans des **bases relationnelles** (SQL).
- Chaque variable a un type défini (**numérique, texte, date…**).

Exemples:

- Table des ventes : ID client, produit, prix, date.
- Données bancaires : numéro de compte, solde, transactions.
- Données scientifiques : mesures de capteurs, résultats de laboratoire.

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Objectifs du Data Understanding:

Données semi Structurées

- Ne suivent pas strictement le modèle tableau, mais contiennent des **balises ou une organisation flexible**.
- Stockées dans des formats comme **JSON, XML, logs**.
- Plus difficiles à manipuler directement qu'un tableau, mais plus riches en informations.

Exemples:

- Données d'une API météo (JSON).
- Fichiers XML pour l'échange de données entre systèmes.
- Logs de serveurs (avec timestamp, message, code d'erreur).

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Objectifs du Data Understanding:

Données non Structurées:

- Pas de structure prédéfinie, souvent **brutes**.
- Nécessitent un **traitement avancé (NLP, Computer Vision, Speech processing)** pour devenir exploitables.

Exemples:

- Texte libre (emails, articles, tweets).
- Images, vidéos, audio (photos médicales, caméras de surveillance, enregistrements vocaux).
- Documents PDF scannés.

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Objectifs du Data Understanding:

Variables **qualitatives**, décrivent des catégories ou des classes. Pas de valeur numérique intrinsèque (mais on peut les coder).

- **Nominales** : pas d'ordre (ex. couleurs : rouge, bleu, vert).
- **Ordinales** : ordre mais sans échelle précise (ex. satisfaction : faible, moyen, élevé).
- **Binaires** : seulement 2 valeurs possibles (oui/non, 0/1).

Variables **quantitatives**, mesurées par des nombres.

- **Discrètes** : valeurs entières (comptage). Ex : nombre d'enfants = 0,1,2.
- **Continues** : valeurs réelles, avec infinité de possibilités. Ex : poids = 65.3 kg.

Variables Temporelles:

- Données liées au **temps**.
- Ont un ordre naturel (passé → présent → futur).
- Souvent utilisées dans les **séries temporelles** (prévision).

Exemples:

- Date d'un achat : 03/10/2025.
- Heures de fonctionnement d'un capteur : 14h23min.
- Périodes : semaine, mois, trimestre.

Spécificités:

- Peuvent être traitées comme **catégorielles** (jour de la semaine).
- Ou comme **numériques** (timestamp en secondes depuis 1970) ou encodage cyclique:

$$x_{sin} = \sin\left(\frac{2\pi \cdot x}{T}\right), \quad x_{cos} = \cos\left(\frac{2\pi \cdot x}{T}\right)$$

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Objectifs du Data Understanding:

2- Évaluer la qualité des données

- Détecter les valeurs manquantes, doublons, incohérences, Outliers.
- Repérer les problèmes de format (ex : majuscules/minuscules, caractères spéciaux).
- Vérifier la cohérence hiérarchique (ex : *ville* → *état* → *pays*).

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Objectifs du Data Understanding:

3- Identifier les enjeux légaux et éthiques

- Vérifier la présence de données personnelles ou sensibles (noms, adresses, identifiants).
- Respecter les règles de confidentialité, RGPD, anonymisation, consentement.
- Décider si certaines variables doivent être exclues (PII – personally identifiable information).

4- Comprendre le sens métier des variables

Savoir ce que représente chaque attribut et comment il se rapporte à la problématique étudiée.
Exemple : *Dans un dataset médical, la variable "Age" peut être en années ou en mois, il faut clarifier.*

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Objectifs du Data Understanding:

5- Explorer les patterns et relations initiales

- Réaliser une **analyse** pour trouver des tendances.
- Déterminer la distribution des variables.
- Détecter des corrélations potentielles entre features et cible.

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

	Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment & Management
Steps	Determine Business Objectives	Design Features				
	Frame the Problem	Collecte Initial Data				
	Define sucess measurements	Install & Import Packages				
	Identify Target Variable	Read the Data				
	Identify Analytical Approach	Data Manipulation & Wrangling				
	Identify Deployement Plan	Exploratory Data Analysis (EDA)				
	Identify the Team	Data Visualization				
	Produce Project Plan	Statistical Analysis				
Output	Analytical Base Table (ABT)					

Packages : pandas, numpy, seaborn, matplotlib, plotly, scipy, statsmodels.

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

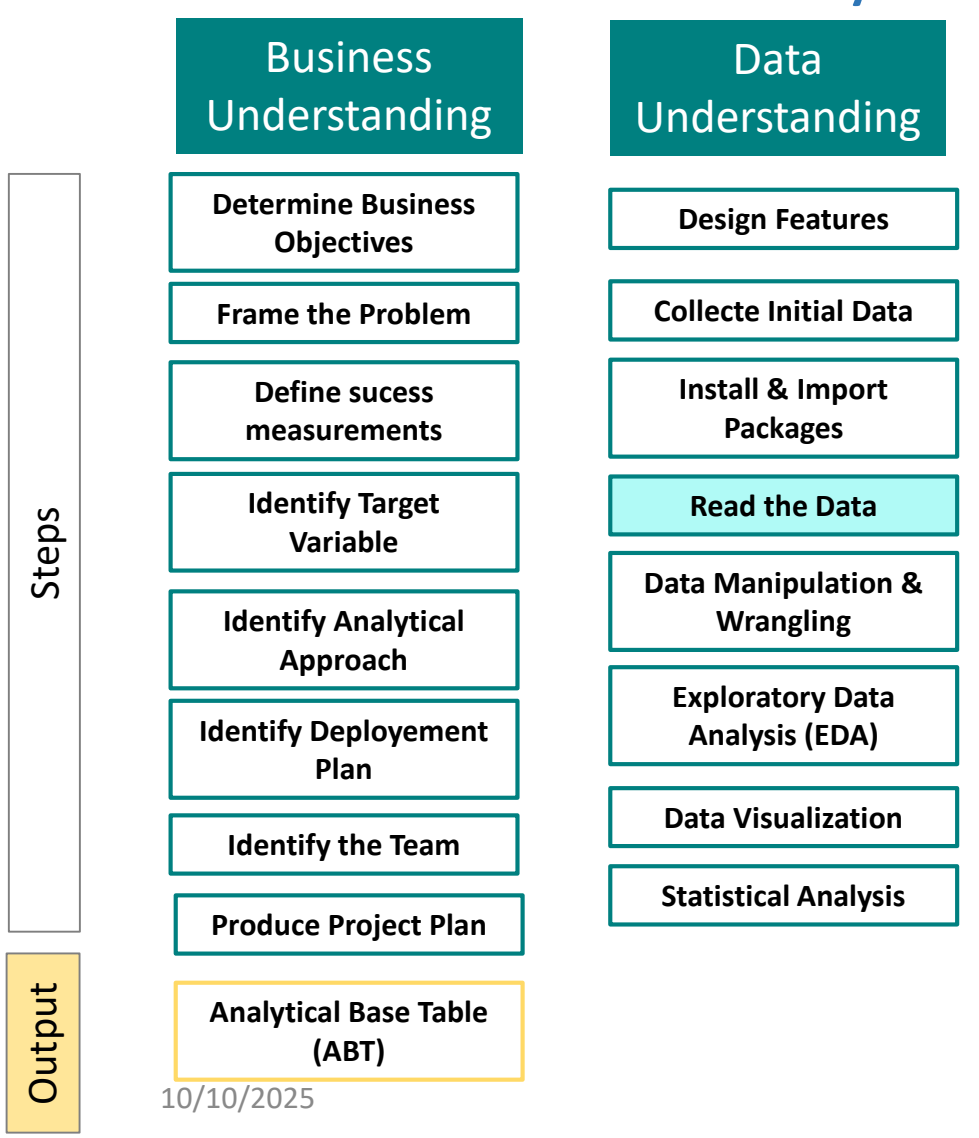
Méthodologie de développement de projets d'IA

Pandas et **NumPy** sont deux bibliothèques populaires en Python utilisées pour l'exploration et la préparation de données.

Critère	Pandas	NumPy
Objectif principal	Manipulation et analyse de données structurées (tableaux de données étiquetés)	Calculs numériques et traitement de données multidimensionnelles
Structures de données	DataFrame (table à deux dimensions) Series (tableau à une dimension)	ndarray (tableau multidimensionnel homogène)
Types de données	Peut contenir différents types de données dans un même DataFrame (entiers, chaînes, dates, flottants, etc.)	Homogène : tous les éléments doivent être du même type (par exemple, tous les entiers ou tous les flottants)
Manipulation de données	Manipulation de données étiquetées, filtrage, tri, fusion, jointure, gestion des données manquantes	Manipulation d'objets numériques sous forme de matrices et de vecteurs, opérations arithmétiques rapides sur les tableaux

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)



- Lire différents formats (CSV, Excel, SQL, JSON).
- Vérifier structure (head, info, shape).

Bonnes pratiques :

- Harmoniser noms de colonnes (lowercase, suppression des espaces et caractères spéciaux).
- Identifier clés uniques (IDs).
- Données hierarchiques
- Données constantes

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment & Management

Steps

- Determine Business Objectives
- Frame the Problem
- Define sucess measurements
- Identify Target Variable
- Identify Analytical Approach
- Identify Deployement Plan
- Identify the Team
- Produce Project Plan

Output

- Analytical Base Table (ABT)

- Design Features
- Collecte Initial Data
- Install & Import Packages
- Read the Data
- **Data Manipulation & Wrangling**
- Exploratory Data Analysis (EDA)
- Data Visualization
- Statistical Analysis

Première interaction avec les données.
Ici, on explore la structure et on fait des manipulations simples (sans nettoyer en profondeur → ça sera le preprocessing).

Actions typiques :

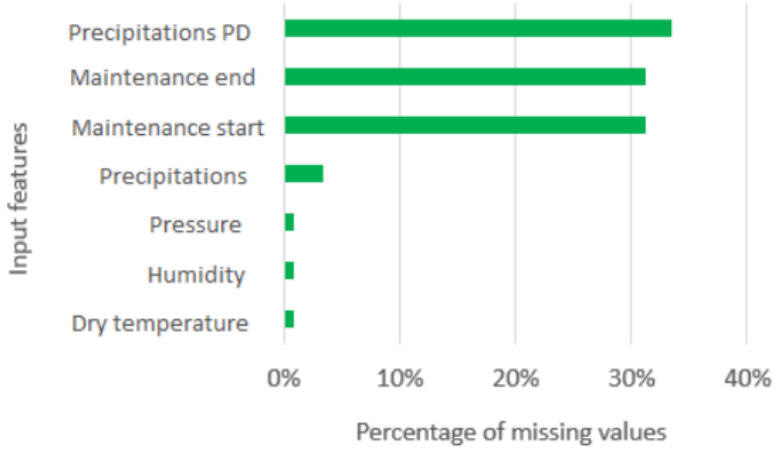
- Vérifier colonnes, types (df.info()).
- Sélectionner/filtrer des sous-ensembles.
- Créer de nouvelles variables (ratios, combinaisons).
- Détecter les valeurs manquantes , outliers, ou doublons (mais pas les corriger).

Objectif : **mieux comprendre comment les données sont organisées**

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Données Manquantes:



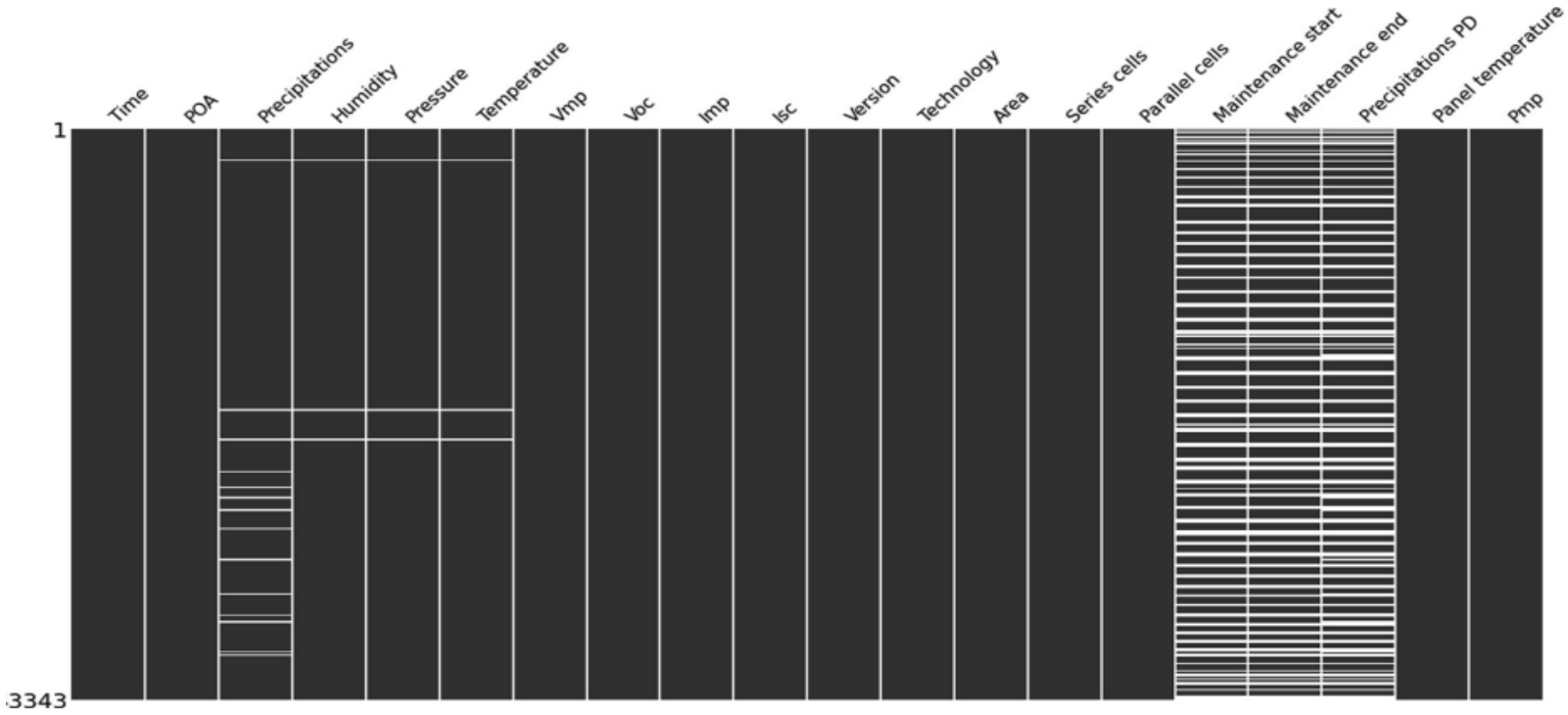
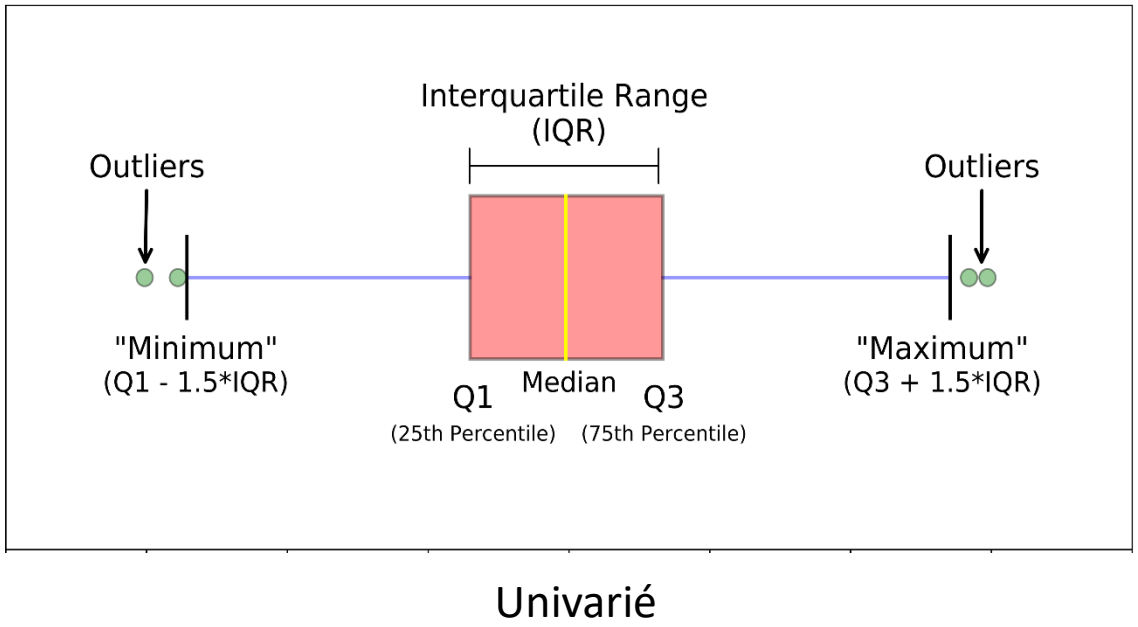


Fig. 2. Distribution of null values in data samples of one panel (white lines represent null values).

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Outliers:



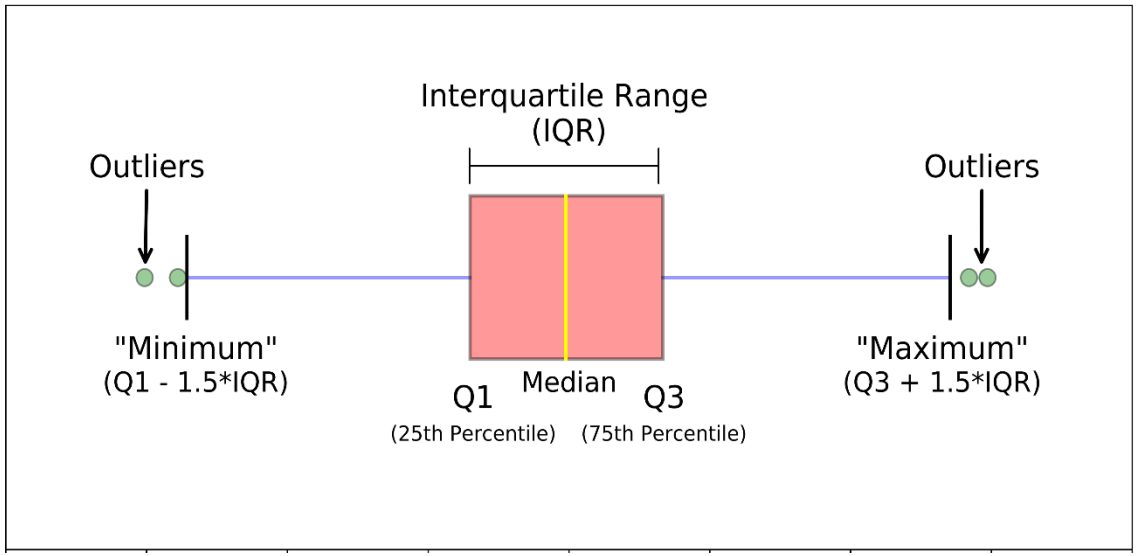
Univarié

Radiation (W/m²)	PV Power (W/m²)
0.5	500
500	0.9
550	330
10	8
18	4
1000	800

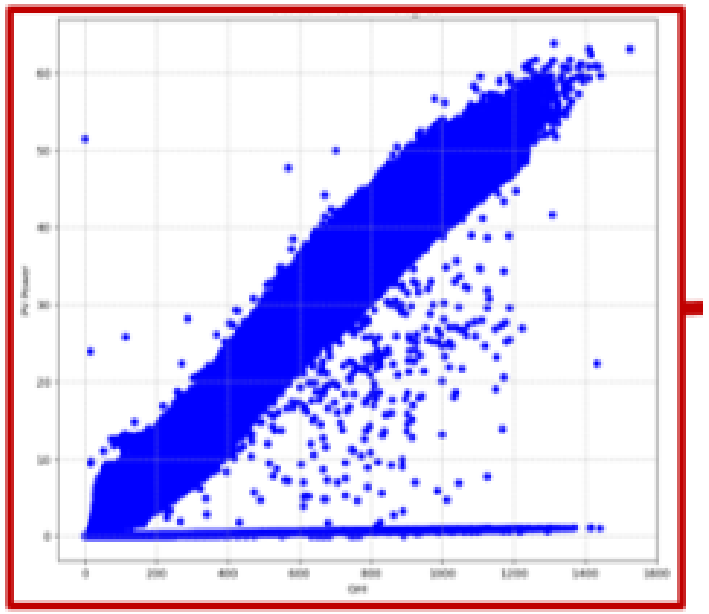
Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Outliers:



Univariées



Multivariées

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment & Management

Steps

- Determine Business Objectives
- Frame the Problem
- Define sucess measurements
- Identify Target Variable
- Identify Analytical Approach
- Identify Deployement Plan
- Identify the Team
- Produce Project Plan

Output

- Analytical Base Table (ABT)

- Design Features
- Collecte Initial Data
- Install & Import Packages
- Read the Data
- Data Manipulation & Wrangling
- **Exploratory Data Analysis (EDA)**
- Data Visualization
- Statistical Analysis

Étape où on essaie de comprendre le contenu du dataset.
Différence avec Wrangling : ici, on cherche à extraire des patterns et insights.
Actions typiques :

- Statistiques descriptives (mean, median, std).
- Distribution des variables.
- Relations entre features et target.
- Premières hypothèses (ex : "Les passagers femmes ont plus survécu").

Objectif : **générer des intuitions sur le dataset**.

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

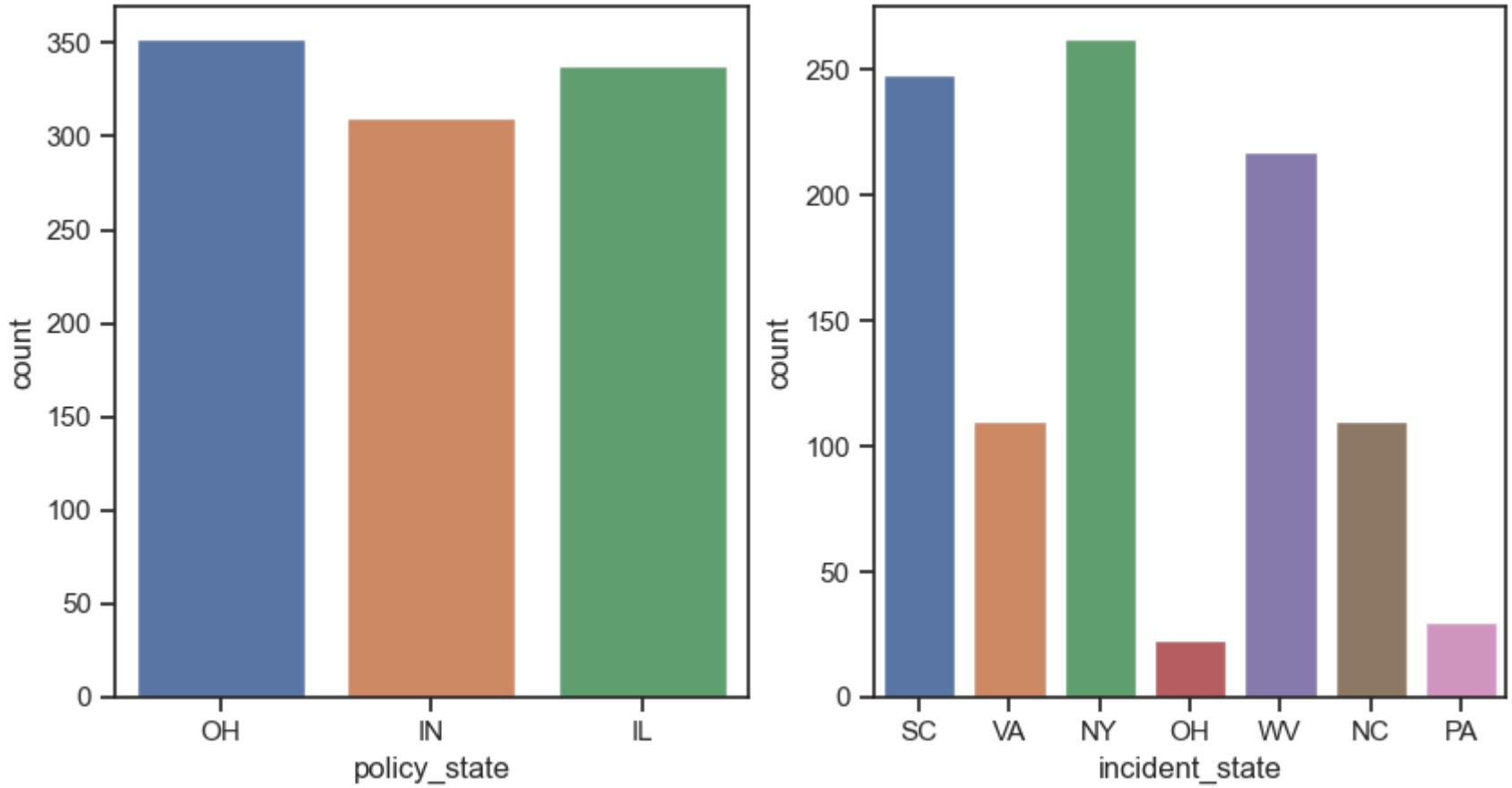
Statistiques sur les données:

	months_as_customer	age	fraud_reported
count	1000.000000	1000.000000	1000.000000
mean	203.954000	38.948000	0.247000
std	115.113174	9.140287	0.431483
min	0.000000	19.000000	0.000000
25%	115.750000	32.000000	0.000000
50%	199.500000	38.000000	0.000000
75%	276.250000	44.000000	0.000000
max	479.000000	64.000000	1.000000

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Statistiques sur les données:



Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

La distribution décrit la manière dont les valeurs d'une variable se répartissent dans un ensemble de données.

Pourquoi c'est important en Data Understanding :

- Identifier la forme des données.
- Adapter les méthodes statistiques ou modèles.
- Détecter les anomalies ou biais.

Normale

Asymétrique

Uniforme

Multimodale

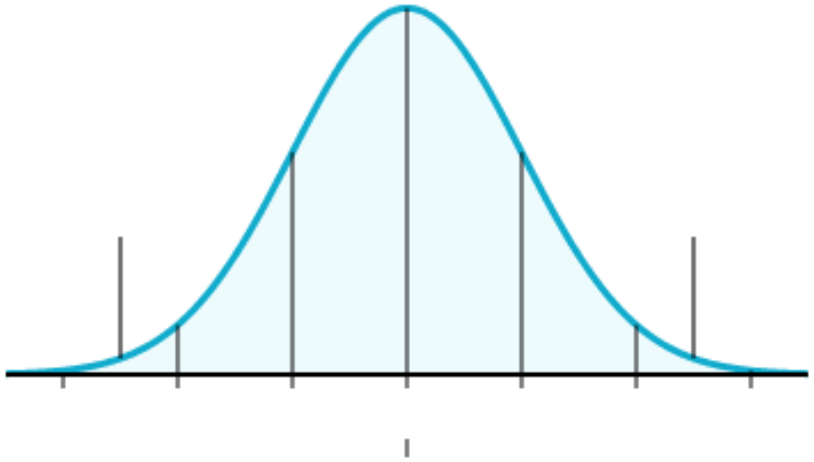
Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Distribution Normale

Caractéristiques :

- Symétrique, en forme de cloche.
- Moyenne = médiane = mode.

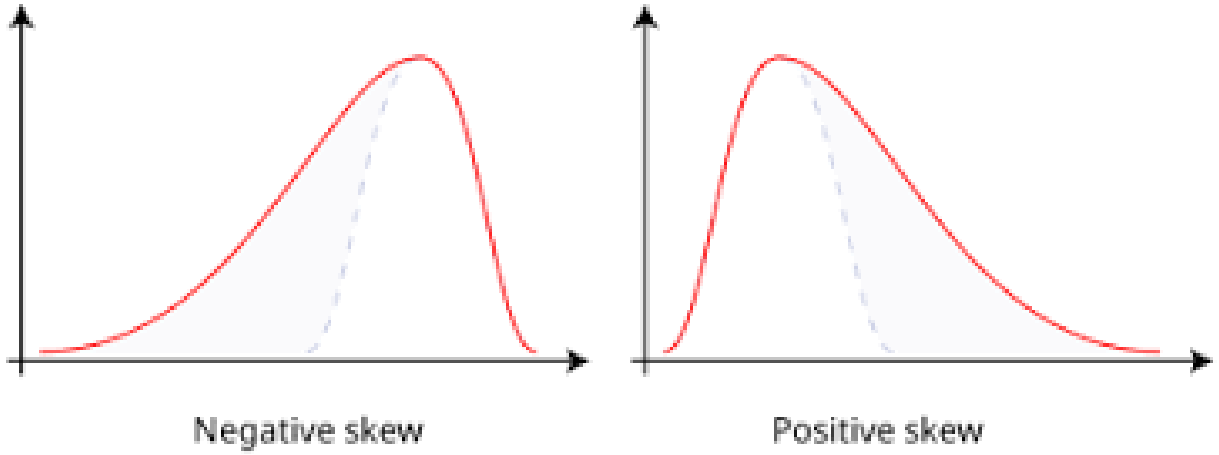


Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Distribution Asymétrique

- **Skewness positive (queue à droite)** :
  - Moyenne > médiane.
  - Exemple : revenus d'une population (quelques très riches).
- **Skewness négative (queue à gauche)** :
  - Moyenne < médiane.
  - Exemple : âge de décès (beaucoup vivent longtemps, peu meurent jeunes).

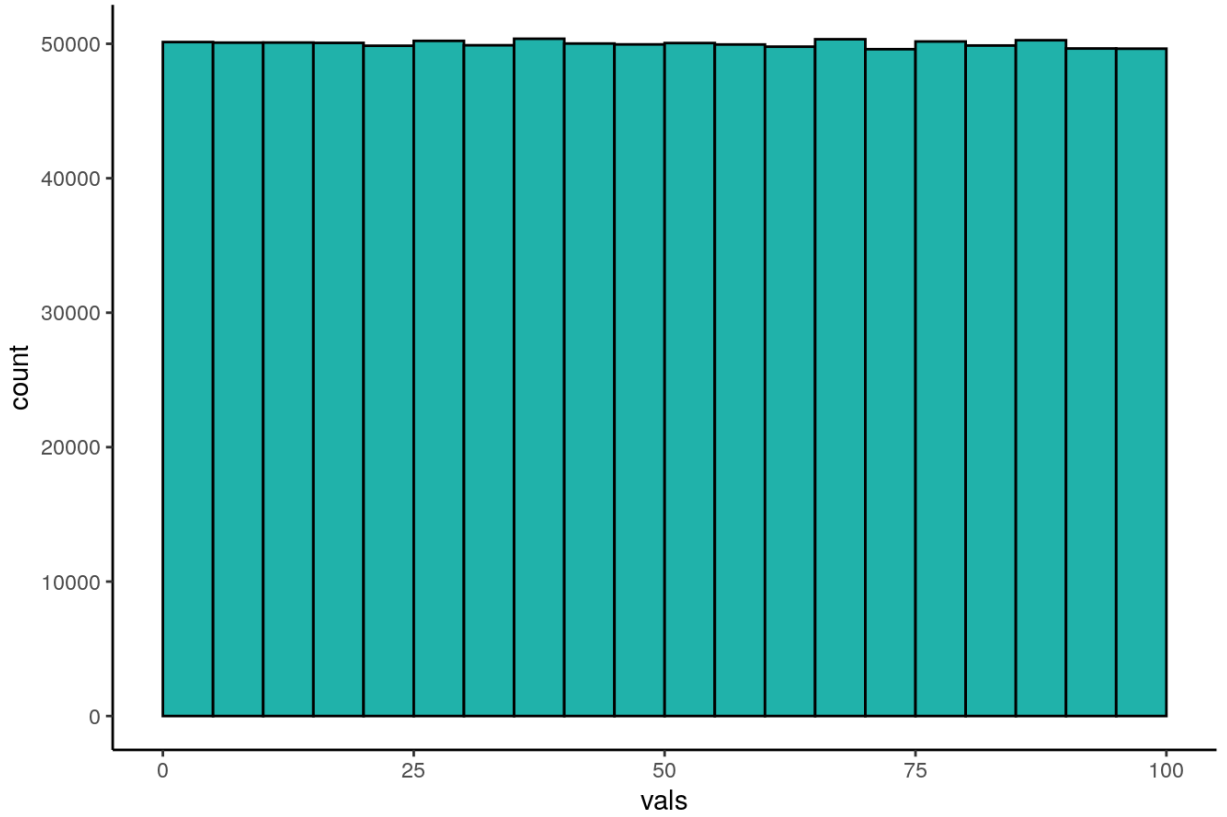


Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Distribution Uniforme

Toutes les valeurs ont la même probabilité

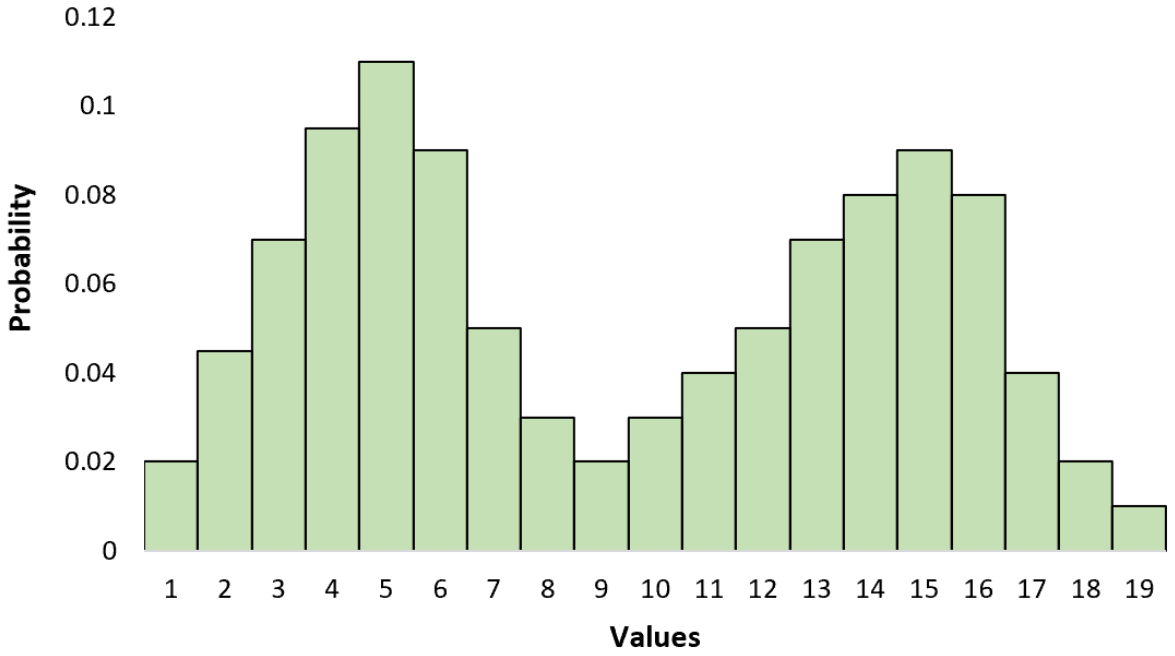


Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

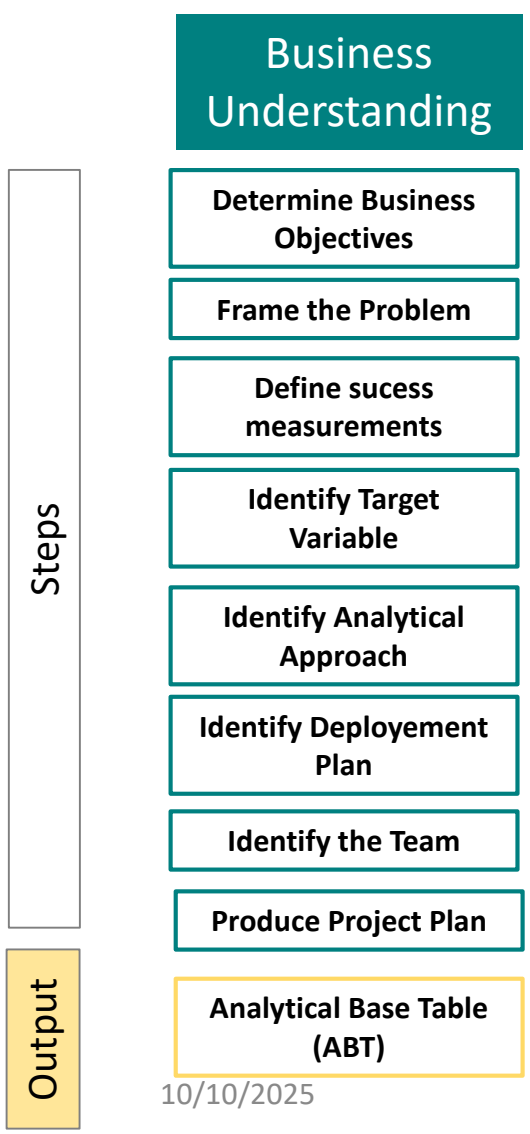
Distribution Binomial

Plusieurs pics



Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)



Support visuel de l'EDA et SA.
On utilise les graphes pour rendre lisible et interprétable ce qu'on observe.

Actions typiques :

- Histogrammes, boxplots (distributions).
- Scatterplots, barplots (relations).
- Heatmaps (corrélations).
- Graphiques interactifs (plotly).

Objectif : **voir les patterns cachés que les chiffres seuls ne montrent pas**.

La Bibliothèque Pandas

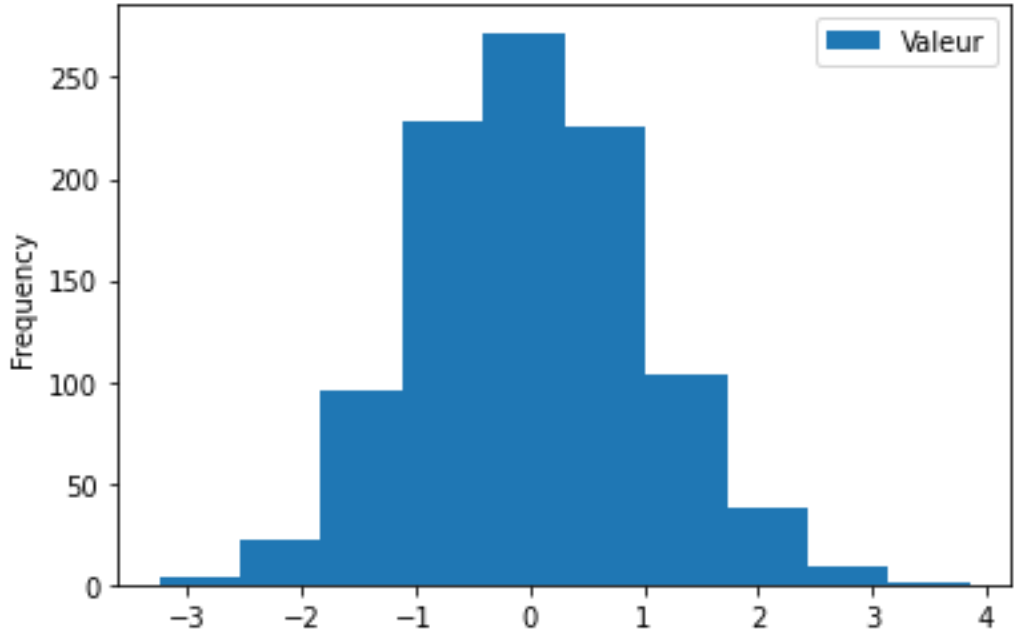
La Visualisation des Données

df.plot.hist()

```
1  np.random.seed(42)
2  data_continu = np.random.randn(1000)
3  df = pd.DataFrame(data_continu, columns=['Valeur'])
4  df
```

	Valeur
0	0.496714
1	-0.138264
2	0.647689
3	1.523030
4	-0.234153
...	...
995	-0.281100
996	1.797687
997	0.640843
998	-0.571179
999	0.572583

1000 rows × 1 columns



La Bibliothèque Pandas

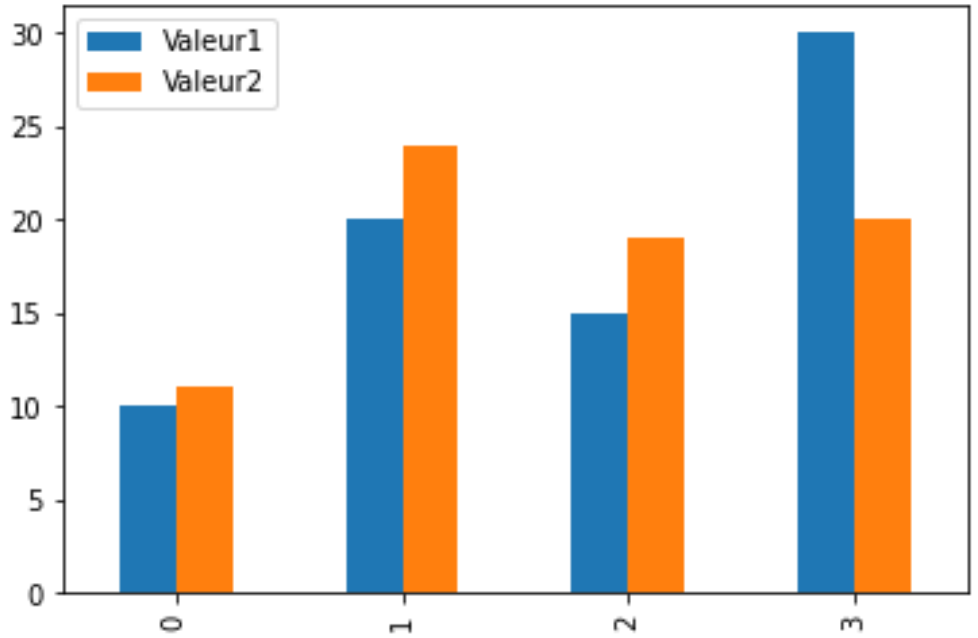
La Visualisation des Données

df.plot.bar()

```python
df = pd.DataFrame({
    'Categorie': ['A', 'B', 'C', 'D'],
    'Valeur1': [10, 20, 15, 30],
    'Valeur2': [11, 24, 19, 20]
})

df
```

	Categorie	Valeur1	Valeur2
0	A	10	11
1	B	20	24
2	C	15	19
3	D	30	20



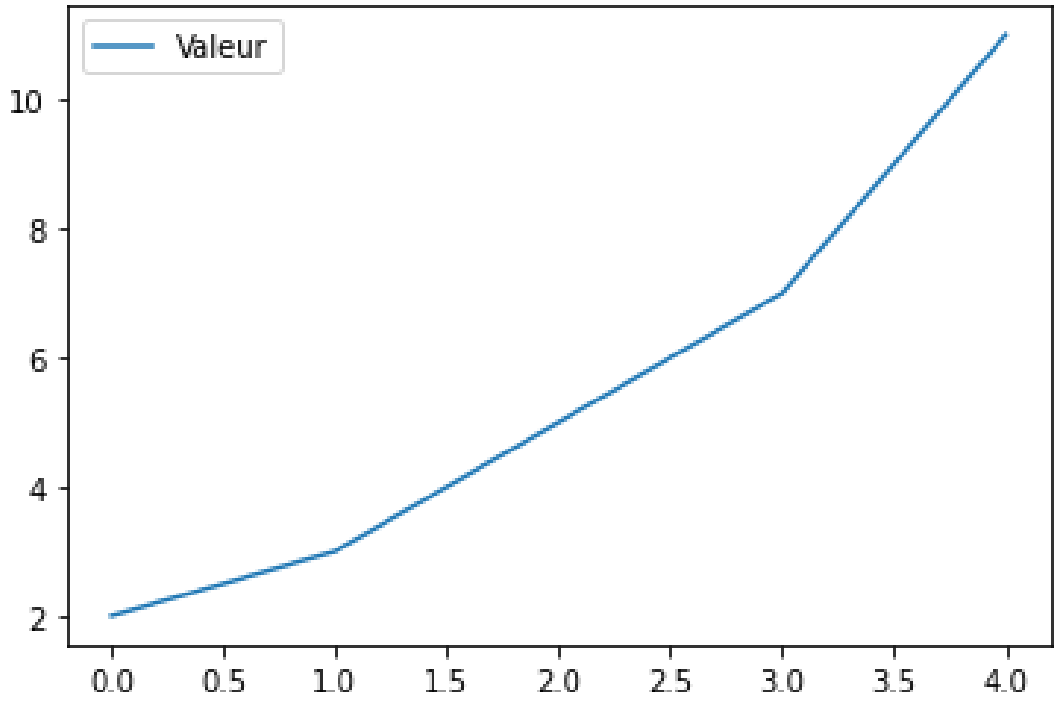
La Bibliothèque Pandas

La Visualisation des Données

df.plot()

```
1  df = pd.DataFrame({
2      'Valeur': [2, 3, 5, 7, 11]
3  })
4  df
```

	Valeur
0	2
1	3
2	5
3	7
4	11



La Bibliothèque Pandas

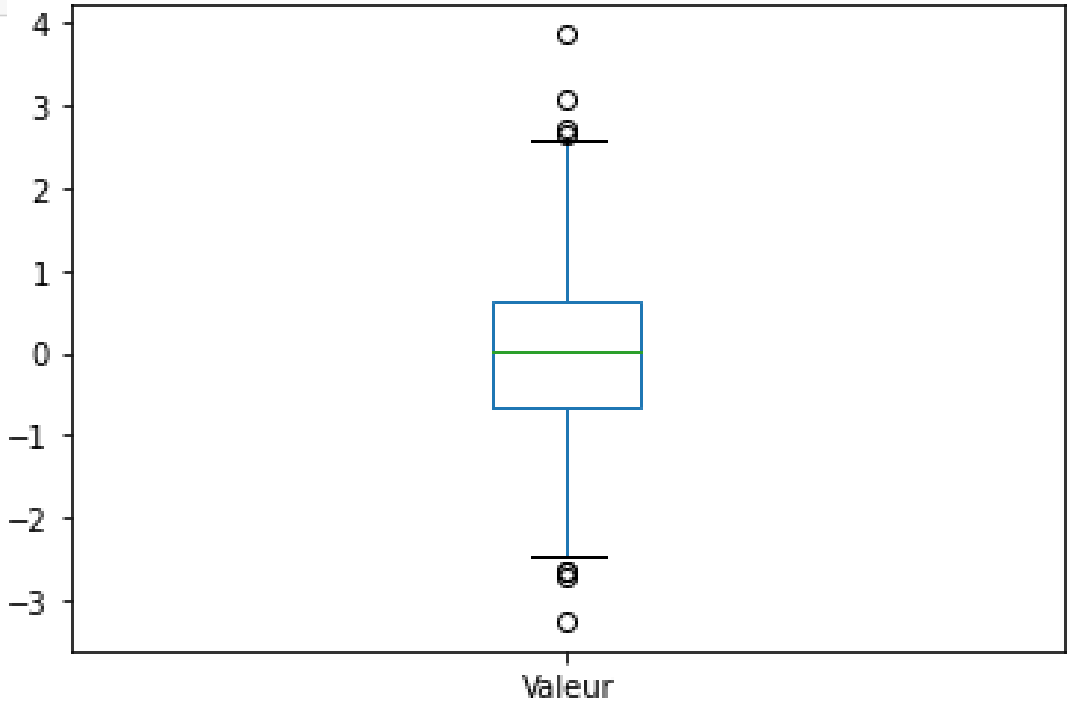
La Visualisation des Données

df.plot.box()

```
1  np.random.seed(42)
2  data_continu = np.random.randn(1000)
3  df = pd.DataFrame(data_continu, columns=['Valeur'])
4  df
```

	Valeur
0	0.496714
1	-0.138264
2	0.647689
3	1.523030
4	-0.234153
...	...
995	-0.281100
996	1.797687
997	0.640843
998	-0.571179
999	0.572583

1000 rows × 1 columns



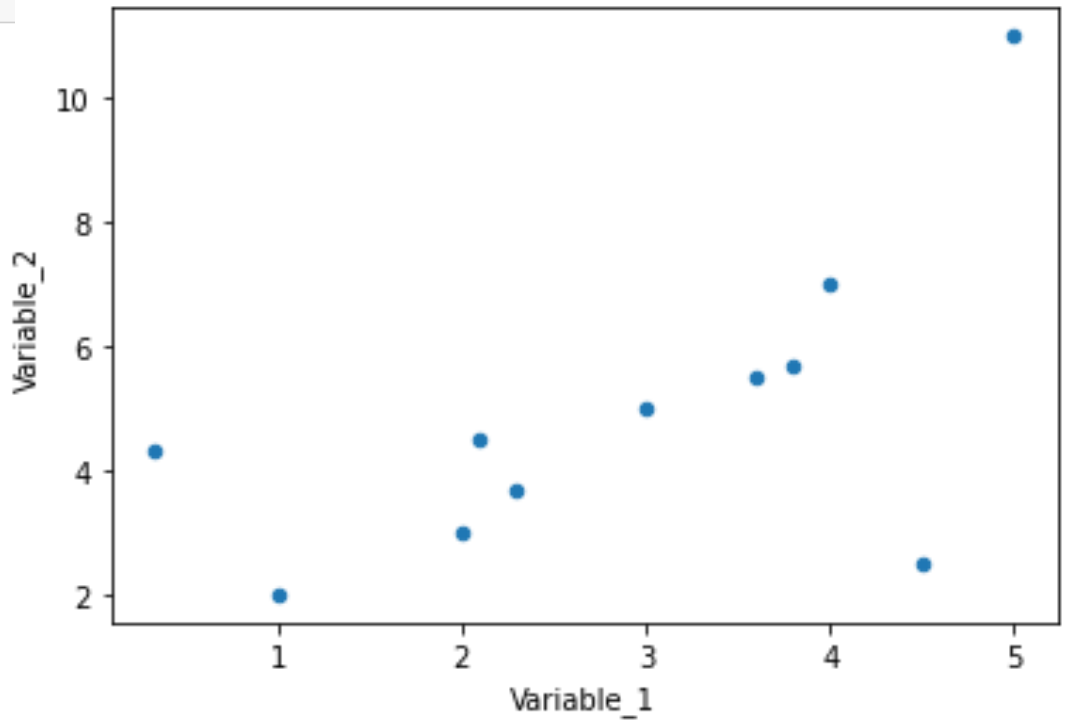
La Bibliothèque Pandas

La Visualisation des Données

df.plot.scatter(x=‘ ‘,y=‘ ‘)

```
df = pd.DataFrame({
    'Variable_1': [1, 2, 3, 4, 5,4.5,3.6,3.8,2.3,0.33,2.1],
    'Variable_2': [2, 3, 5, 7, 11,2.5,5.5,5.7,3.7,4.3,4.5]
})
df
```

	Variable_1	Variable_2
0	1.00	2.0
1	2.00	3.0
2	3.00	5.0
3	4.00	7.0
4	5.00	11.0
5	4.50	2.5
6	3.60	5.5
7	3.80	5.7
8	2.30	3.7
9	0.33	4.3
10	2.10	4.5



Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

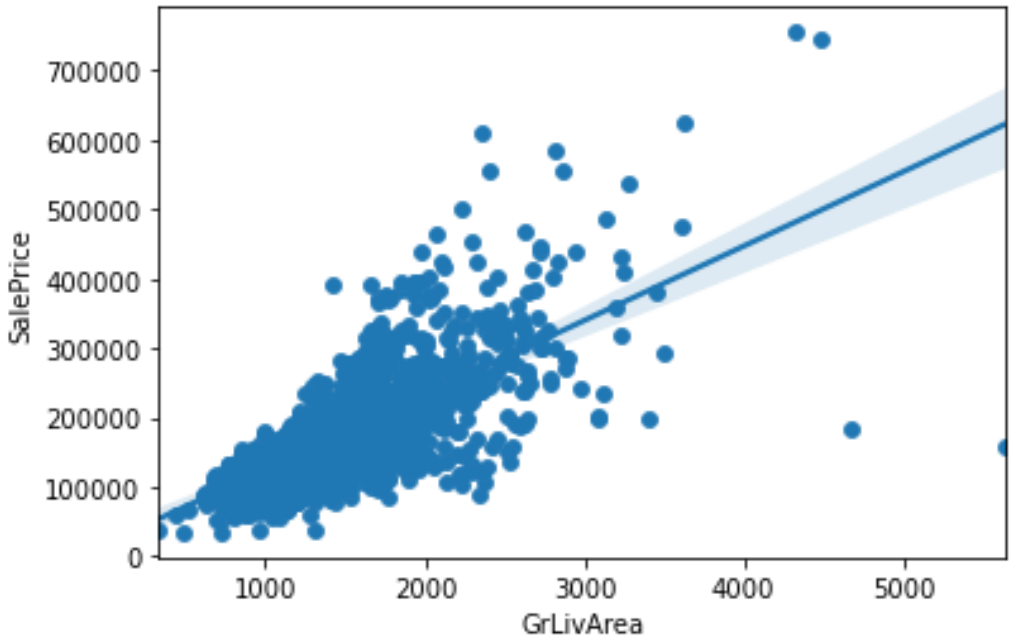
Visualization de features Vs target:

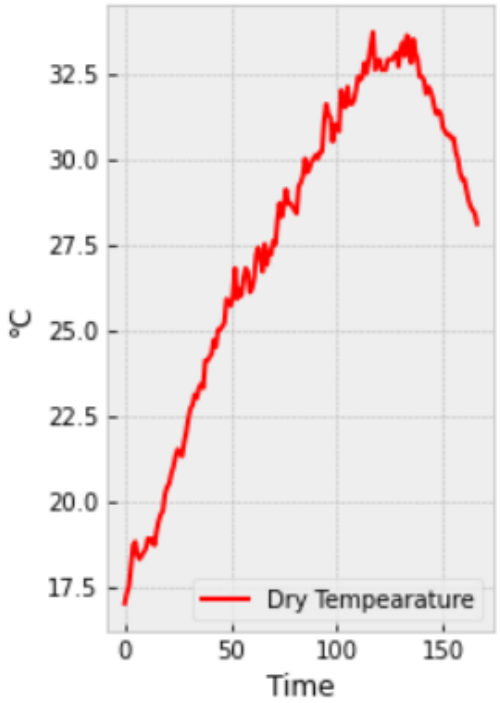
- **Feature numérique Vs Target numérique**
- **Feature numérique Vs Target catégoriel**
- **Feature catégoriel Vs Target numérique**
- **Feature catégoriel Vs Target catégoriel**

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

- **Feature numérique Vs Target numérique**

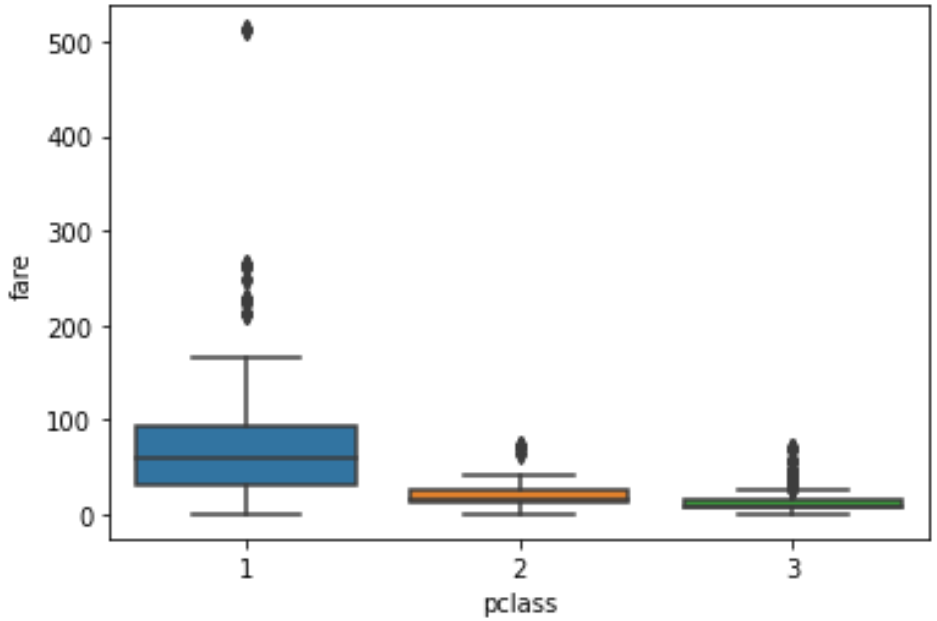




Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

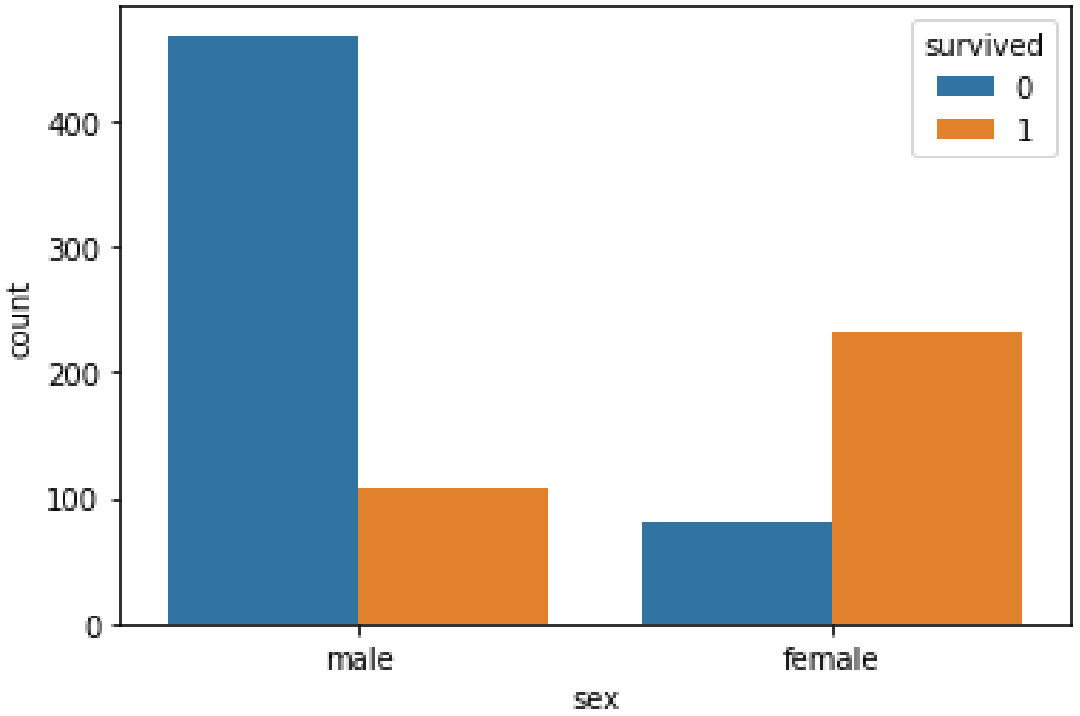
- **Feature numérique Vs Target categoriel**



Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

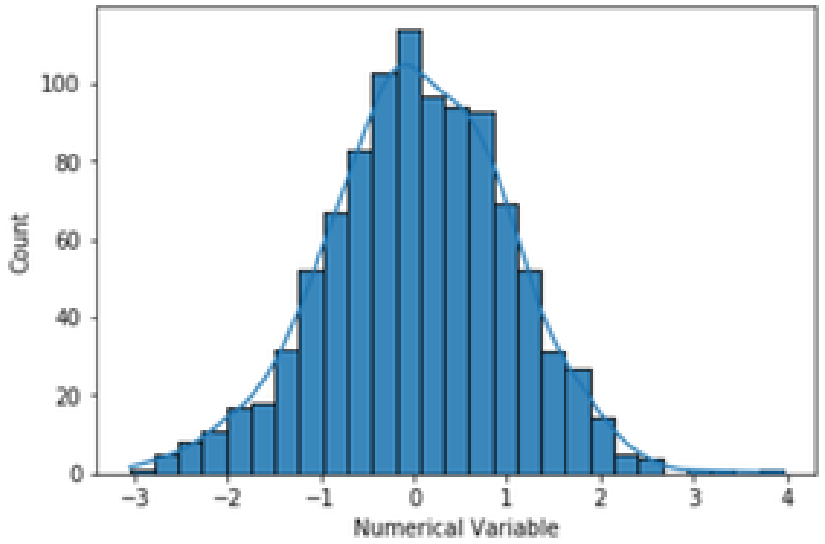
- **Feature categorical Vs Target categoriel**



Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

- **Feature categorical Vs Target numerique**



Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment & Management

Steps

- Determine Business Objectives
- Frame the Problem
- Define sucess measurements
- Identify Target Variable
- Identify Analytical Approach
- Identify Deployement Plan
- Identify the Team
- Produce Project Plan

Output

- Analytical Base Table (ABT)

- Design Features
- Collecte Initial Data
- Install & Import Packages
- Read the Data
- Data Manipulation & Wrangling
- Exploratory Data Analysis (EDA)
- Data Visualization
- Statistical Analysis

Output

- Quality Report

Étape plus formelle que l'EDA.
Ici, on fait des tests statistiques pour valider ou infirmer des hypothèses.

Actions typiques :

- Moyenne, écart-type, variance.
- Test t (comparer deux moyennes).
- Test Khi² (indépendance entre variables catégorielles).
- Corrélations (Pearson, Spearman).

Exemple : L'âge moyen des survivants est-il significativement plus bas que celui des non-survivants ?

Objectif : **passer d'observations descriptives à des conclusions validées.**

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Notion d'hypothèse statistique:

Une hypothèse statistique est une idée que l'on veut tester avec les données.

- H_0 (hypothèse nulle) : il n'y a pas de différence ou pas de relation ou pas d'effet.
- H_1 (hypothèse alternative) : il y a une différence ou une relation ou un effet.

Exemple:

H_0 : l'âge moyen est le même pour les survivants et les non-survivants.
H_1 : l'âge moyen est différent.

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

La p-value (valeur p)

C’est la probabilité que la différence observée soit due au hasard.

Exemple:
Imagine que tu compares deux groupes.

- Si p est grande > 0.05 (ex. 0.6) → la différence peut venir du hasard → on garde H_o.
- Si p est petite <=0.05 (ex. 0.01) → la différence est très improbable par hasard → on rejette H_o.

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Tests paramétriques vs non paramétriques

Paramétriques → données normales (courbe en cloche).
Non paramétriques → pas besoin de normalité, souvent basés sur les **rangs**.

Types de relations possibles:

- **Numérique Vs Numérique:** Température ↔ Production
- **Numérique Vs Catégorique (2):** Revenu ↔ Sexe
- **Numérique Vs Catégorique (>2):** Production ↔ Type de panneau
- **Catégorique Vs Catégorique**: Sexe ↔ Survie

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

1- TESTS DE MOYENNES

Test t de Student

Comparer la **moyenne** de deux groupes.

Principe :

On calcule la **p-value** :

- Si $p < 0.05$ → la différence est **significative** → on rejette H_0.
- Sinon → pas de différence prouvée.

Âge moyen : survivants = 28.4 ans, non-survivants = 31.7 ans

$p = 0.01$ → Rejet de H_0 → différence significative.

Si non normal → Mann–Whitney

Mann–Whitney :

- Compare les **rangs** au lieu des valeurs.
- Moins sensible aux valeurs extrêmes.
- Même objectif : vérifier si une différence existe.

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Test d'ANOVA

Comparer les moyennes de **plus de 2 groupes**.

- $p < 0.05$ → au moins un groupe est différent.
- $p > 0.05$ → toutes les moyennes sont proches.

Exemple :
Moyenne production (Type A=10.1, B=13.2, C=11.0)
$p = 0.03$ → au moins un type est différent.

Si non normal → Kruskal–Wallis
Même principe que ANOVA, mais sur les **rangs**.

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Test Khi² (Chi-square)

Tester si deux variables **catégorielles** sont liées.
Principe :
On compare les **fréquences observées** avec celles qu'on aurait eues **si les variables étaient indépendantes**.

Sexe	Survivant	Non-survivant
Homme	109	468
Femme	233	81

Si effectifs faibles (<5) → Fisher Exact Test
Même logique mais plus précis pour petits échantillons.

Ne mesure pas la **force** du lien.
→ Utiliser le **V de Cramer** (0 = pas de lien, 1 = lien fort).

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

2- TESTS DE CORRÉLATION

Corrélation de Pearson

Mesurer la relation linéaire entre deux variables numériques.

r	Interprétation
r = 1	lien parfait positif
r = -1	lien parfait négatif
r ≈ 0	pas de lien linéaire

Exemple :

Radiation ↑ → Production ↑ → r = 0.95 → corrélation forte positive.

Limite :

- Suppose une relation **linéaire** et **normale**.
- Sensible aux valeurs extrêmes.

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Corrélation de Spearman

Mesurer le lien **monotone** (croissant ou décroissant).
Principe :
On remplace les valeurs par leur **rang** (1er, 2e, 3e…).
→ On calcule ensuite un r basé sur ces rangs.
Moins sensible aux valeurs extrêmes et ne suppose pas la normalité.

Corrélation de Kendall

→ Plus robuste sur petits échantillons.

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

Relation	Exemple	Tests possibles
Numérique ↔ Numérique	Température ↔ Production	Corrélation (Pearson, Spearman, Kendall)
Numérique ↔ Catégorique (2 groupes)	Revenu ↔ Sexe	Test t ou Mann–Whitney
Numérique ↔ Catégorique (>2 groupes)	Production ↔ Type de panneau	ANOVA ou Kruskal–Wallis
Catégorique ↔ Catégorique	Sexe ↔ Survie	Khi² ou Fisher

Méthodologie de développement de projets d'IA

Cross-Industry Standard Process for Data Mining (CRISP-DM)

TP2: OSTI Data - USA

Eugene

Golden

Cocoa

- **1 year**
- **High Quality**
- **11** PV Panels
- **6** PV Technologies
- **5 min** of data resolution