INTRODUCTION AUX DATA POUR L'IA

Partie 1 : Les statistiques descriptives

Les statistiques descriptives

- L’analyse de données est primordial pour pouvoir développer des IA de qualité
- Il y a un enjeu majeur à explorer les données à notre disposition :
  - *Les comprendre*
  - *Vérifier leur cohérence et leur justesse*
  - *Les nettoyer*
  - *Les transformer*
- Dans ce cadre, les stats descriptives sont un outil indispensable

Les statistiques descriptives pour comprendre les données

- Les statistiques descriptives fournissent une vue d'ensemble rapide et claire des données à travers des mesures de tendance centrale (comme la moyenne, la médiane, et le mode) et des mesures de dispersion (comme la variance et l'écart-type).

- Ces informations aident les data scientists à comprendre les caractéristiques fondamentales des données, telles que la distribution, la centralité, et la variabilité, ce qui est crucial pour toute analyse préliminaire.

Les statistiques descriptives pour vérifier la qualité des données

- Les statistiques descriptives aident à identifier les anomalies ou les valeurs aberrantes (outliers) qui pourraient biaiser ou dégrader la performance du modèle d'IA.

- La détection précoce des valeurs aberrantes permet de réaliser des corrections nécessaires, soit en les supprimant, soit en les traitant de manière appropriée.

Les statistiques descriptives pour sélectionner les caractéristiques

- Les analyses descriptives permettent d'identifier les caractéristiques les plus pertinentes pour un modèle en examinant les relations et les corrélations entre différentes variables.

- Comprendre comment les variables sont liées les unes aux autres aide à sélectionner celles qui seront les plus utiles pour la prédiction et à transformer les données de manière à optimiser la performance de l'algorithme d'IA.

Les statistiques descriptives pour normaliser et standardiser

- Les statistiques descriptives révèlent souvent le besoin de normaliser ou de standardiser les données, surtout quand les variables sont sur des échelles différentes ou suivent différentes distributions.

- Ces transformations sont essentielles pour de nombreux algorithmes d'apprentissage automatique, notamment ceux qui sont sensibles à la magnitude des données, comme la régression logistique, les SVM, et les réseaux de neurones.

- On appelle série statistique la suite des valeurs prises par une variable X sur les unités d'observation.
- Le nombre d'unités d'observation est noté n
- Les valeurs de la variable X sont notées x_1, x_2,.....x_n

Mesures de tendances centrales : le mode

- Le mode est la valeur distincte correspondant à l'effectif le plus élevé
- Le mode peut être calculé pour tous les types de données, quantitatives et qualitatives

Etat civil	Effectifs
Célibataire	19
Marié	7
Veuf	2
Pacsé	5

Mesures de tendances centrales : la moyenne

- La moyenne est la somme des valeurs observées divisées par leur nombre
- La moyenne ne peut être définie que pour une variable quantitative

Participant	Distance de la résidence à Saint-Brieuc
A	0
B	86
C	24
D	5
E	0

Mesures de tendances centrales : la médiane

- La médiane est la valeur centrale de la série statistique
- 50% des valeurs sont en dessous de la médiane, 50% au dessus

Participant	Distance de la résidence à Saint-Brieuc
A	0
B	86
C	24
D	5
E	0

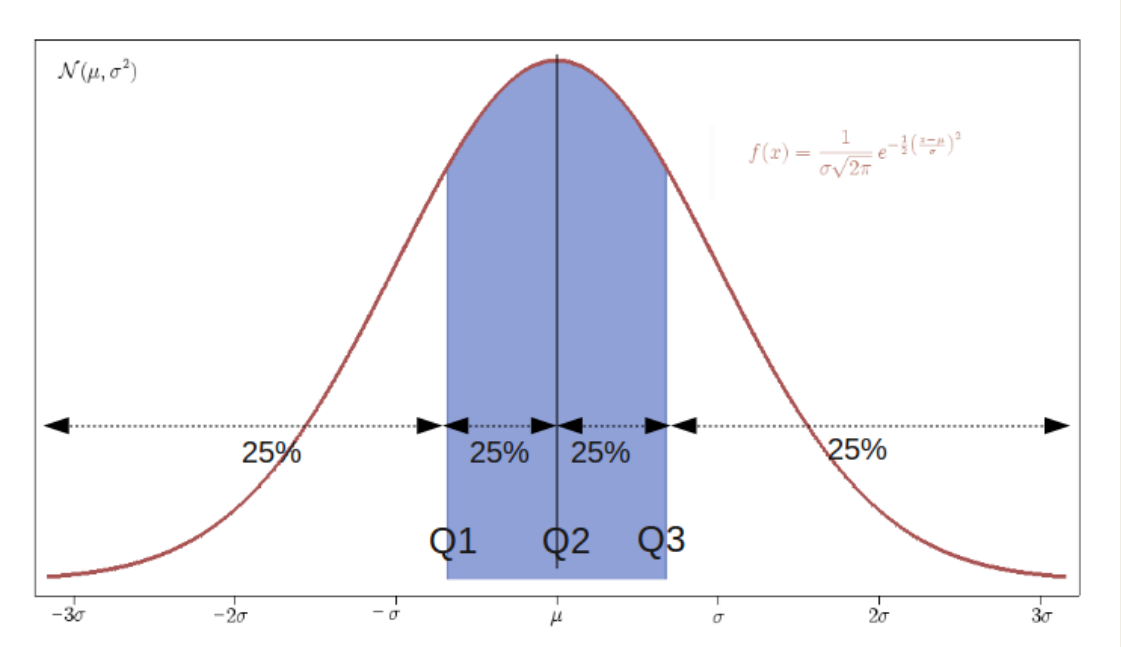
Quelle différence entre la moyenne et la médiane ?

Calculons ces valeurs sur la série de revenus mensuels au sein d'une entreprise :

1300, 1322, 1850, 2540, 1345, 12450, 1540, 1630

Les quantiles

Les quantiles sont les valeurs qui divisent un jeu de données en intervalles de même probabilité
Exemple : les quartiles sont les 3 quantiles qui divisent un jeu de donnée en 4 groupes de même probabilité



Les quantiles

Les quantiles sont les valeurs qui divisent un jeu de données en intervalles de même probabilité
Exemple : les quartiles sont les 3 quantiles qui divisent un jeu de donnée en 4 groupes de même probabilité

Participant	Distance de la résidence à Saint-Brieuc
A	0
B	86
C	24
D	5
E	0

Comment interpréter les données avec ces grandeurs?

- Salaire médian en France : 2100€
- Salaire moyen : 2550€
- 1ier décile : 1370€
- 9ieme décile : 4010€
- 99ieme centile : 9600€

Les mesures de dispersion : l’écart-type et la variance

- Ces mesures caractérisent la dispersion des mesures autour de la moyenne
- On peut y voir un reflet de la variabilité des données
- La variance est la somme des carrés des écarts à la moyenne divisée par le nombre d’observations
- L’écart-type est sa racine carrée et est de la même unité que la variable

Variance: $\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$

Écart-Type: $\sigma = \sqrt{\sigma^2}$

Où x_i sont les valeurs individuelles, μ est la moyenne, et N le nombre total d'observations.

Les mesures de dispersion : la distance interquartile

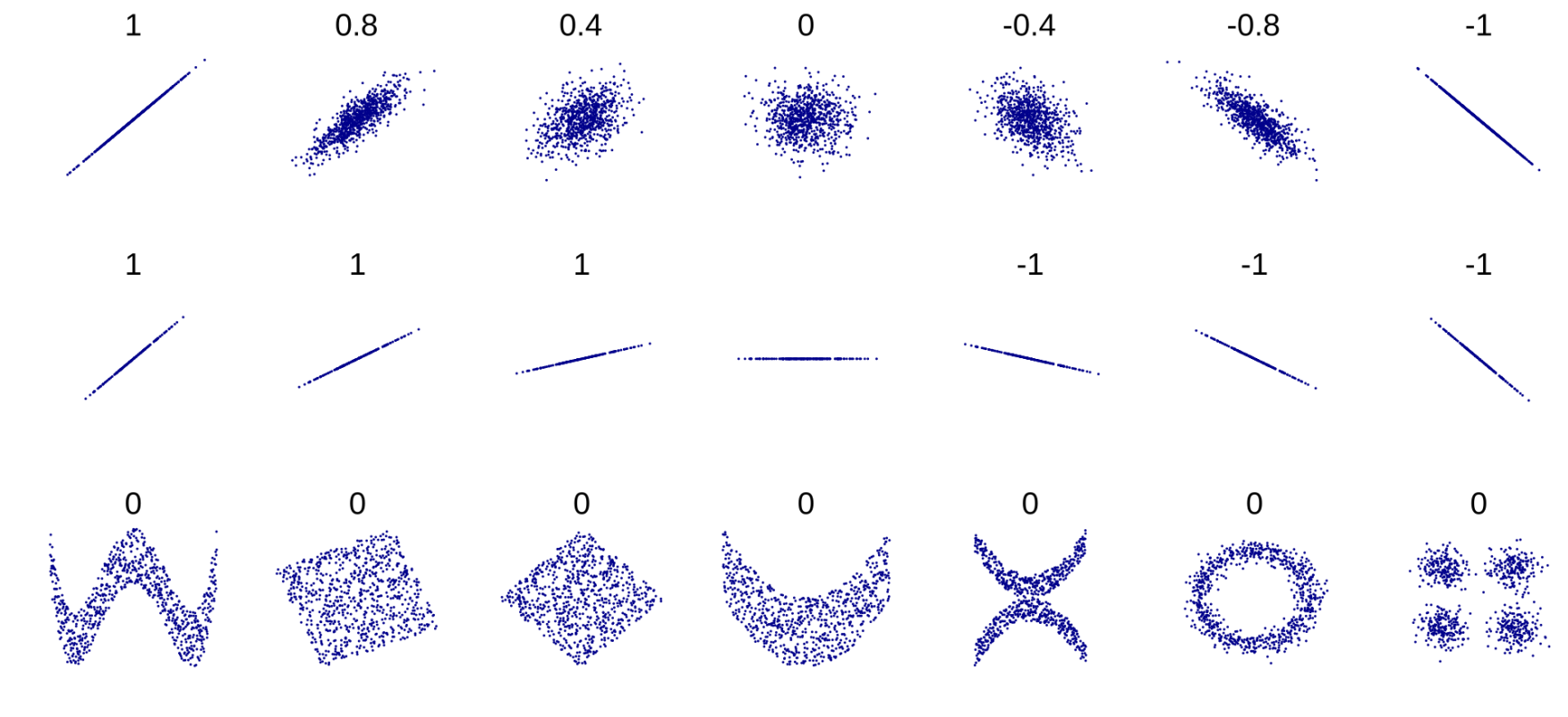
- La distance interquartile est la différence entre le troisième et le premier quartile

INTRODUCTION AUX DATA POUR L'IA

Partie 2 : La corrélation

La corrélation

- la **corrélation** entre plusieurs **variables aléatoires** ou statistiques est une notion de **liaison** qui contredit leur indépendance.



La corrélation linéaire

- Le coefficient de corrélation linéaire est égal à 1 lorsque l'une des variables est une fonction affine croissante de l'autre variable et à −1 lorsque l'une des variables est une fonction affine décroissante de l'autre.

- Les valeurs intermédiaires renseignent sur le degré de dépendance linéaire entre les deux variables. Plus le coefficient est proche des valeurs extrêmes −1 et 1, plus la corrélation linéaire entre les variables est forte ; on emploie simplement l'expression « fortement corrélées » pour qualifier les deux variables.

- Une corrélation égale à 0 signifie que les variables ne sont pas corrélées linéairement, elles peuvent néanmoins être corrélées non-linéairement

La corrélation linéaire

- Le coefficient de corrélation n’est pas sensible aux unités de chacune des variables. Ainsi, par exemple, le coefficient de corrélation linéaire entre l’âge et le poids d’un individu sera identique que l’âge soit mesuré en semaines, en mois ou en années.

- En revanche, ce coefficient de corrélation est extrêmement sensible à la présence de valeurs aberrantes ou extrêmes (ces valeurs sont appelées des « déviants ») dans notre ensemble de données (valeurs très éloignées de la majorité des autres, pouvant être considérées comme des exceptions).

La corrélation de Pearson

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples

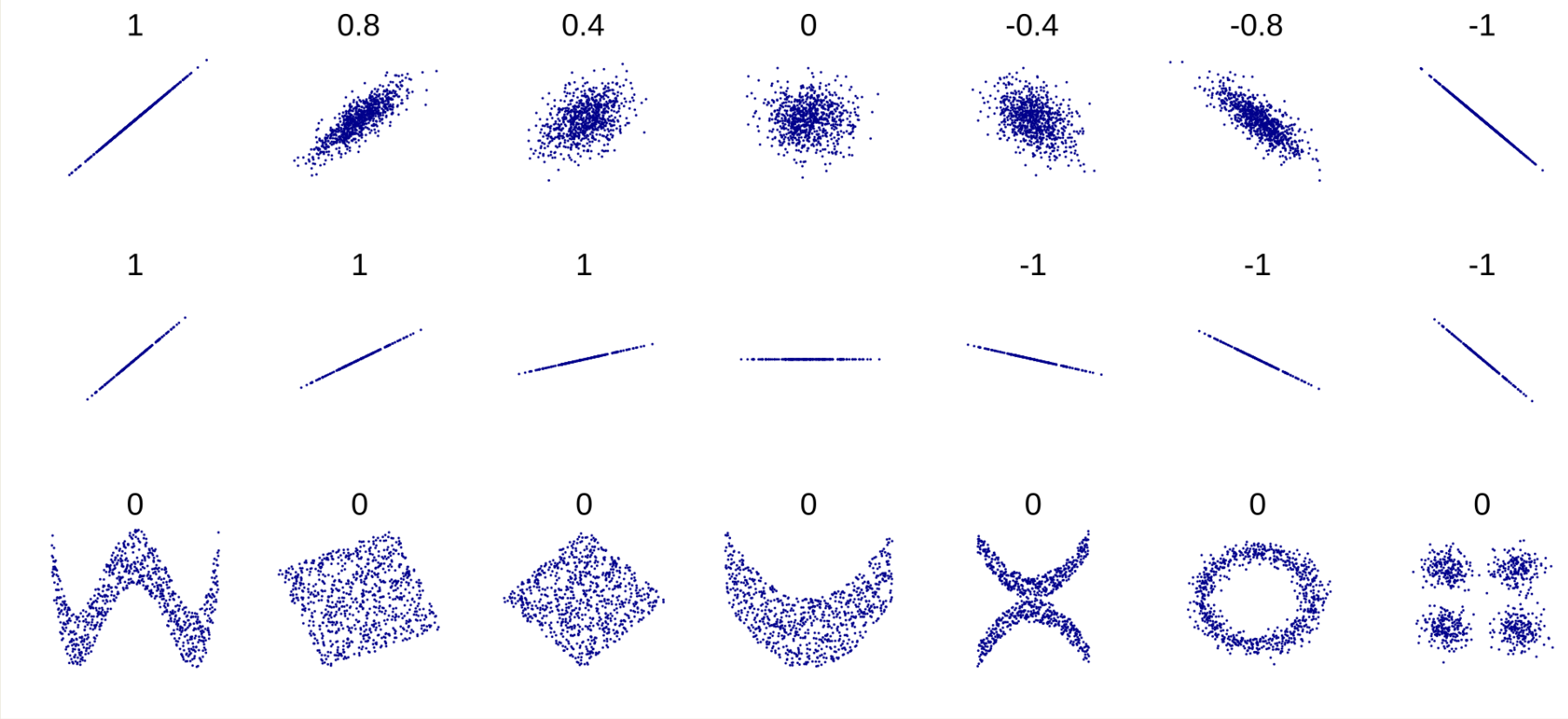
y_i = y variable sample

$\bar{x}$ = mean of values in x variable

$\bar{y}$ =mean of values in y variable

Le diagramme de dispersion

- Le diagramme de dispersion ou de corrélation est un **outil de contrôle et d'aide à la décision pour vérifier l'existence de corrélation ou d'une relation entre variables de nature quantitative.**



INTRODUCTION AUX DATA POUR L'IA

Partie 3 : Les valeurs aberrantes

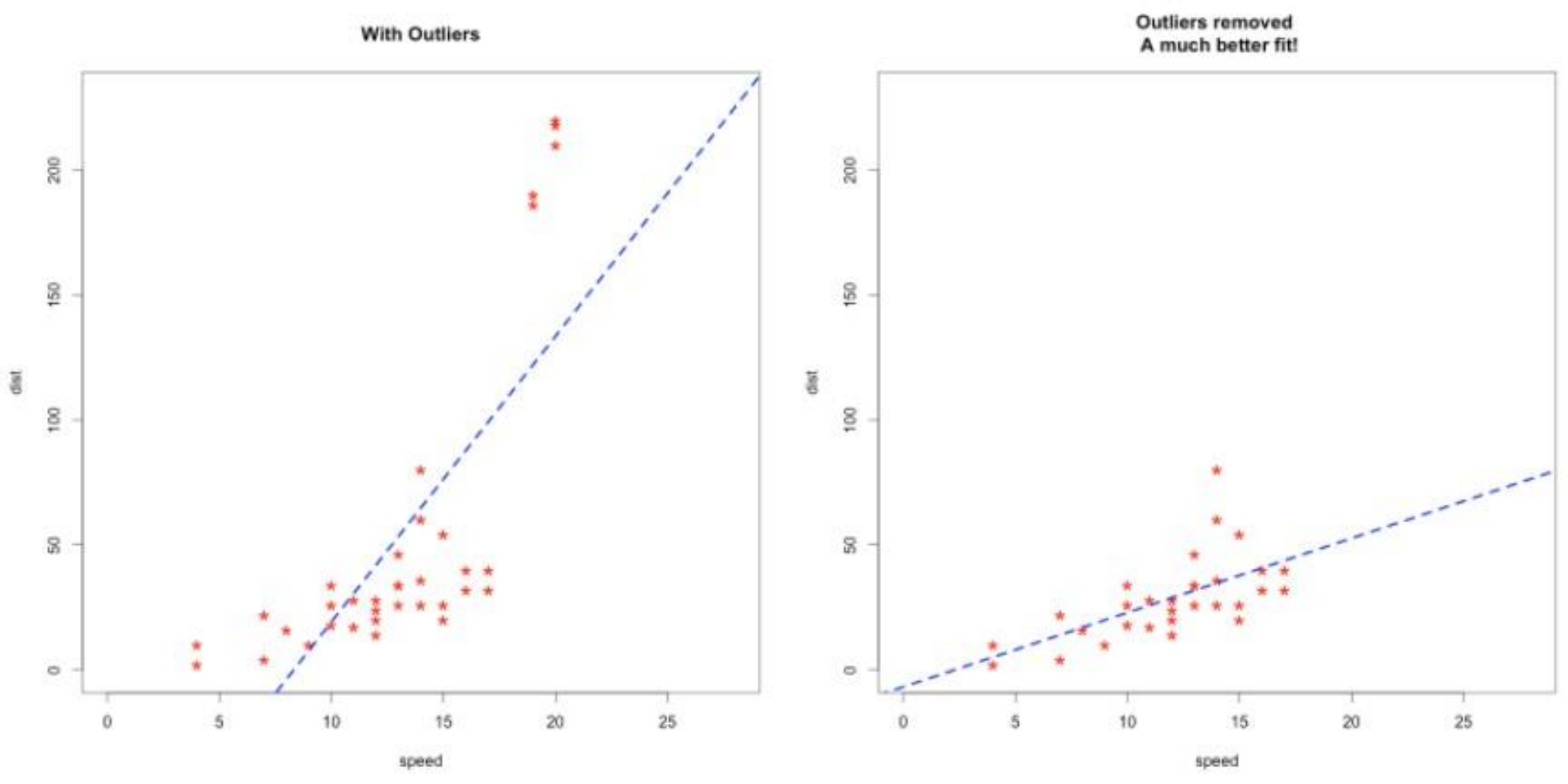
Les outliers

- Un outlier est une valeur ou observation qui est « distante » des autres observations effectuées sur le même phénomène.

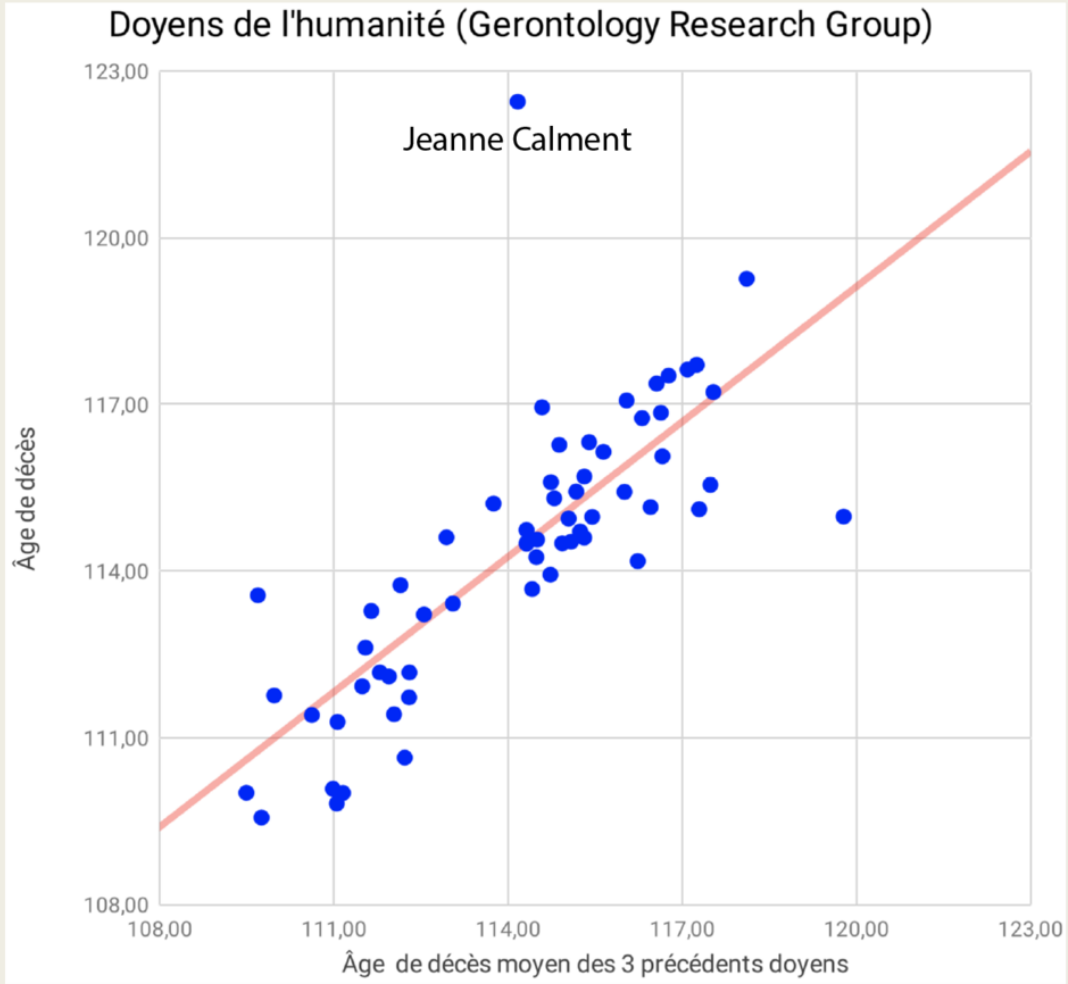
- L'outlier peut être dû à la variabilité inhérente au phénomène observé, ou constituer une aberration expérimentale. Il faut alors chercher à les retirer des analyses, car elles ne comportent pas d'information pertinente pour la modélisation

- Remarque : la valeur médiane est bien plus robuste aux outliers que la moyenne

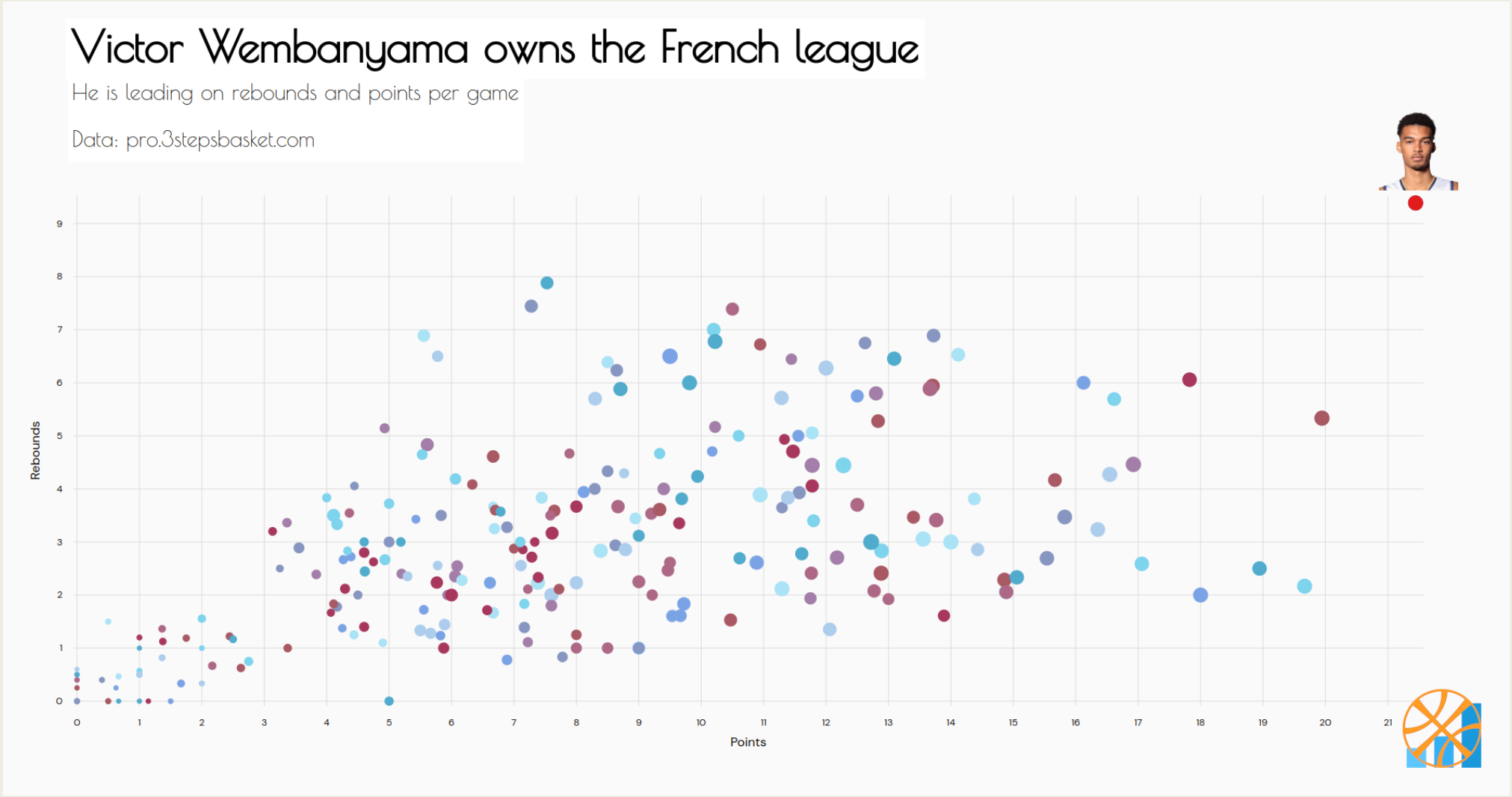
Les outliers



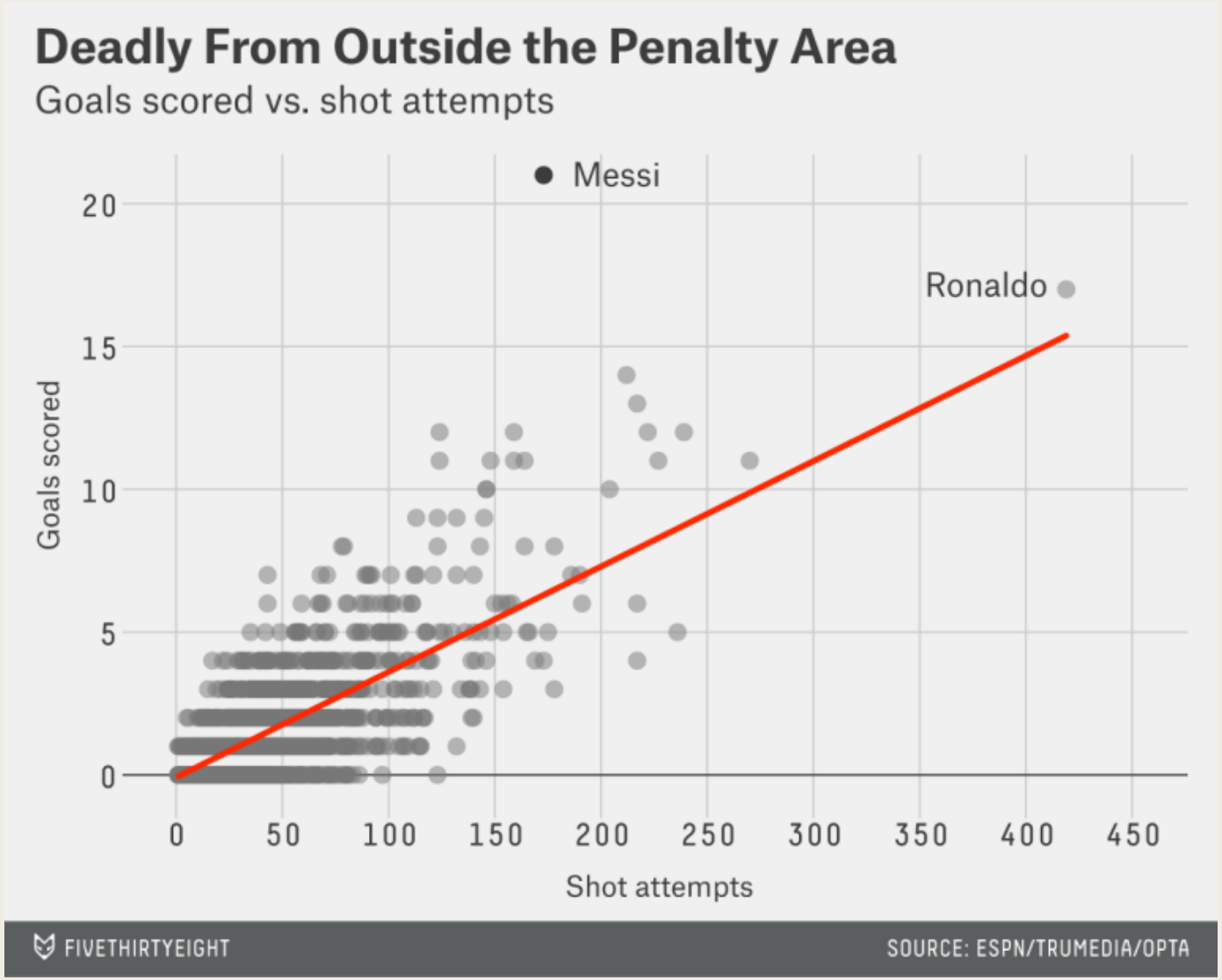
Comment caractériser un outlier ?



Comment caractériser un outlier ?

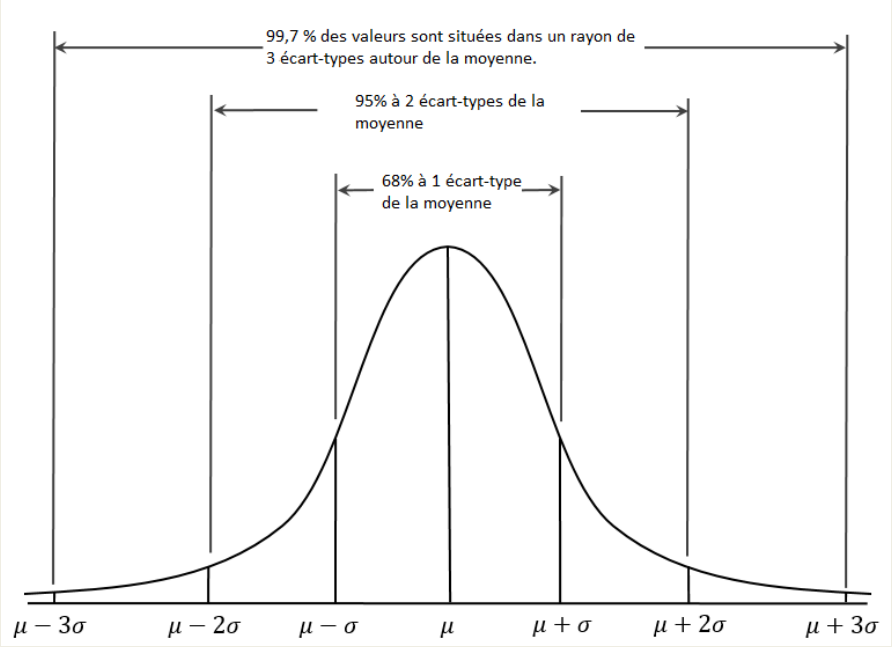


Comment caractériser un outlier ?



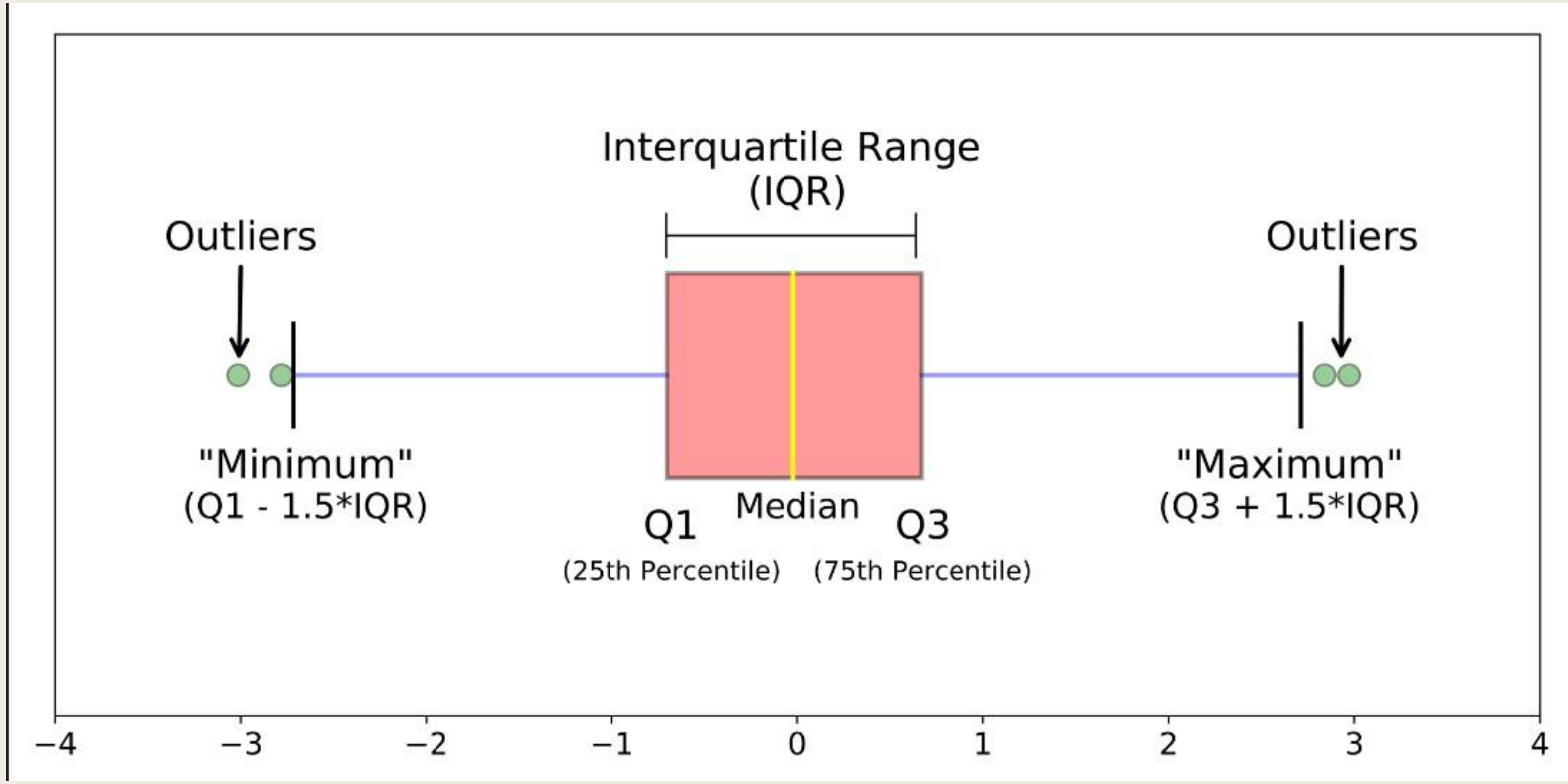
Comment caractériser un outlier ?

- Quand la distribution d'une variable suit une loi normale, 68% des mesures se situe dans l'intervalle [moyenne-1sigma, moyenne + 1 sigma]
- **1 observation sur 22** aura un écart à la moyenne > 2 fois l'écart type
- **1 observation sur 370** aura un écart à la moyenne > 3 fois l'écart type
- Une première méthode simple pour détecter un outlier et de considérer toute valeur au-delà de 2 écarts types



Comment caractériser un outlier ?

- La méthode de l'interquartile
- Pour rappel : IQR = Q3-Q1



Comment caractériser un outlier ?

- Le test de Grubbs (suppose une distribution normale) identifie un outlier en mesurant à quel point la mesure dévie de la moyenne relativement à l'écart-type.

1. **For the Maximum Value:**

   The test statistic G is calculated for the suspected maximum outlier $X_{\max}$ as follows:

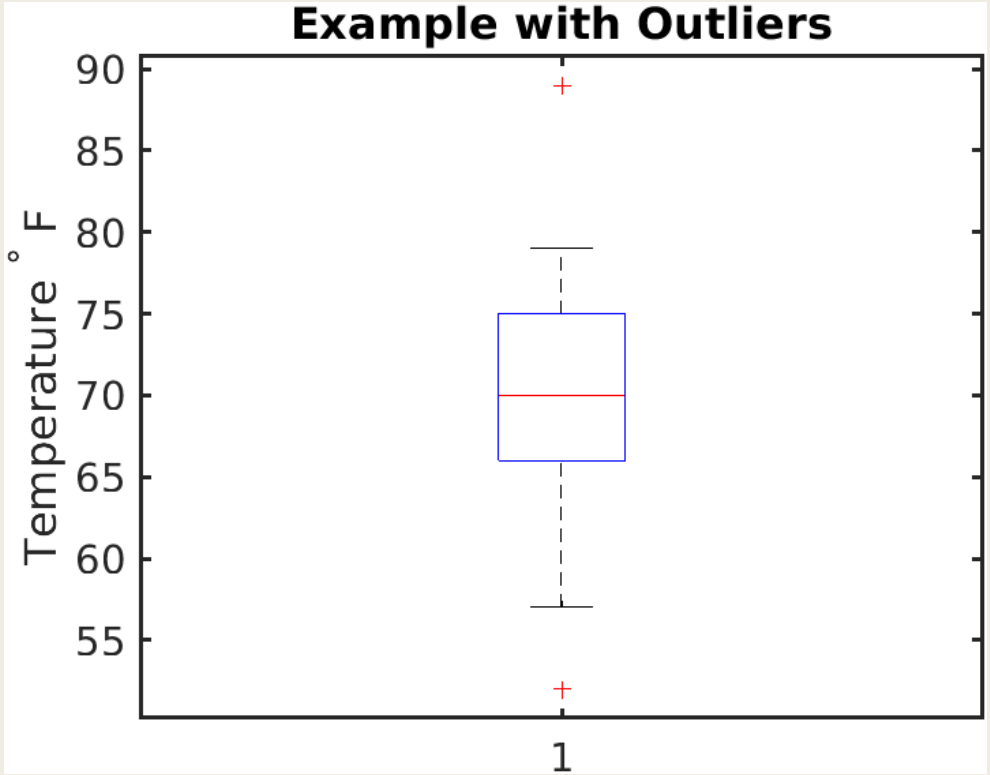
$$G_{\max} = \frac{|X_{\max} - \bar{X}|}{s}$$

   - $X_{\max}$ is the maximum value in the dataset.
   - $\bar{X}$ is the sample mean.
   - s is the sample standard deviation.

   If $G_{\max}$ exceeds the critical value from the Grubbs' distribution table for a given significance level, $X_{\max}$ is considered an outlier.

Comment visualiser les outliers ?

- L'outil des boxplot ou des boites à moustaches
- Les outliers sont considérés comme au-delà de Q3 + 1,5 IQR, ou en deca de Q1-1,5 IQR



INTRODUCTION AUX DATA POUR L'IA

Partie 4 : Le nettoyage des données

Le nettoyage des données

- Le nettoyage des données est le processus d'identification et de correction des erreurs dans les données pour améliorer leur qualité.
- Il garantit la fiabilité des analyses
- Il permet des décisions éclairées basées sur des données précises.

Les données à nettoyer

- Problème 1 : les valeurs aberrantes
- Problème 2 : les valeurs dupliquées
- Problème 3 : les données incorrectes

Le processus de nettoyage des données

1. **Exploration des données :** Analyser les données pour comprendre leur structure et détecter des problèmes.

2. **Nettoyage des données :** Corriger ou supprimer les valeurs erronées

3. **Validation des données :** Vérifier que les données nettoyées sont correctes

INTRODUCTION AUX DATA POUR L'IA

Partie 5 : La gestion des données manquantes

Les données manquantes

- **Type 1 : MCAR : Missing completely at random**
  - Les valeurs manquantes ne sont pas liées à d'autres variables ou à la valeur manquante elle-même.
  - **Exemple** : si des données sont perdues en raison d'une erreur de saisie aléatoire.
  - **Les analyses effectuées sur ces données sont généralement valides**, car le manque d'information est aléatoire.

Les données manquantes

- **Type 2: MAR : Missing at random**
  - Les données sont manquantes, mais ce manque peut être expliqué par d'autres variables observées.
  - **Exemple** : les répondants ayant des revenus élevés peuvent être moins susceptibles de fournir des informations sur leurs dépenses.
  - **Des méthodes d'imputation appropriées peuvent être utilisées**, car le modèle de données manquantes peut être pris en compte dans l'analyse.

Les données manquantes

- **Type 3: NMAR : Not Missing at random**
  - Le manque de données est lié à la valeur même de la donnée manquante.
  - **Exemple** : une personne qui ne souhaite pas révéler son salaire peut avoir une valeur manquante pour cette variable.
  - Plus difficile à traiter, car il peut introduire des biais significatifs. **Des méthodes spécifiques doivent être appliquées** pour tenter de résoudre ces problèmes.

Identifier les données manquantes

```
import pandas as pd
data = pd.read_csv('data.csv')
missing_values = data.isnull().sum()
print(missing_values)
```

```
import seaborn as sns
import matplotlib.pyplot as plt
sns.heatmap(data.isnull(), cbar=False, cmap='viridis')
plt.show()
```

Identifier les données manquantes

- **Détection précoce :** Identifier les valeurs manquantes dès le début permet de planifier comment les gérer.

- **Choix des méthodes :** Comprendre le type de données manquantes aide à choisir la bonne méthode de traitement.

- **Impact sur l'analyse :** Les valeurs manquantes peuvent affecter la précision des modèles prédictifs et des analyses statistiques.

Enlever les valeurs manquantes

df.dropna() => enlève les lignes avec au moins 1 Nan

df.dropna(how = ‘all’) => enlève les lignes avec que des Nan ou des valeurs manquantes

df.dropna(axis =1) => enlève les colonnes avec au moins 1 Nan

Remplacer les valeurs manquantes

df.fillna(0) => les Nan sont remplacés par des 0

df.fillna(value=mean_value, inplace=True) => les Nan sont remplacés par la moyenne de la colonne

df.fillna(value=median_value, inplace=True) => les Nan sont remplacés par la mediane de la colonne

df.fillna(method = 'pad') => les Nan sont remplacés par la valeur précédente

df.fillna(method = 'bfill') => les Nan sont remplacés par la valeur suivante

df.interpolate(method = 'linear', limit_direction = 'forward') => les Nan sont interpolés linéairement

INTRODUCTION AUX DATA POUR L'IA

Partie 6 : La normalisation ou la standardisation des données

Pourquoi Normaliser et Standardiser ?

- **Importance :**
  - Les algorithmes de Machine Learning sont sensibles à l'échelle des données.
  - Améliore la convergence des algorithmes d'optimisation.
  - Réduit le biais introduit par des unités de mesure différentes.

- **Conséquences d'une mauvaise échelle :**
  - Les modèles peuvent ne pas converger ou donner des résultats biaisés.

Normalisation des Données

- La normalisation (ou mise à l'échelle) consiste à redimensionner les valeurs d'une variable pour qu'elles se situent dans un intervalle spécifique, souvent [0, 1].

$$X' = \frac{X - \min(X)}{\max(X) - \min(X)}$$

- Idéal pour les algorithmes basés sur la distance (ex. : K-means, K-NN).

Normalisation des Données

- **Facilite l'apprentissage :** Les modèles convergent plus rapidement.
- **Élimine les biais d'échelle :** Toutes les variables contribuent de manière équitable à la distance.
- **Utilisation dans les réseaux de neurones :** Aide à éviter le problème du gradient éclatant.

Standardisation des Données

- La standardisation consiste à centrer les données autour de la moyenne et à les redimensionner en fonction de l'écart-type.

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

- Appropriée pour les modèles qui supposent que les données sont distribuées normalement (ex. : régression linéaire, SVM).

Standardisation des Données

•**Interprétation des coefficients :** Permet de comparer l'importance relative des variables.

•**Préparation pour les modèles paramétriques :** Utile lorsque les algorithmes reposent sur des hypothèses de distribution.

•**Gestion des valeurs aberrantes :** Moins sensible aux valeurs extrêmes par rapport à la normalisation.

Récapitulatif

Critère	Normalisation	Standardisation
Echelle	[0,1] ou [-1,1]	Distribution avec moyennne=0 et std=1
Méthode	Basée sur min/max	Basée sur la moyenne et l'écart type
Quand l'utiliser	Sur des données non normalement distribuées	Sur des données normalement distribuées
Sensibilité aux outliers	Importante	moyenne

INTRODUCTION AUX DATA POUR L'IA

Partie 7 : Le feature engineering

Le feature engineering (ingénierie des caractéristiques)

- choisir, extraire et remodeler les caractéristiques les plus appropriées pour créer des modèles de Machine Learning précis et performants.

- L'efficacité d'un modèle d'apprentissage est liée à la **qualité des caractéristiques utilisées** pour le former.

L’ingénierie des caractéristiques

- Etape 1 : nettoyer les données, imputer les valeurs manquantes
- Etape 2 : le codage catégoriel : convertir les valeurs catégorielles en valeurs numériques : pd.get_dummies()

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

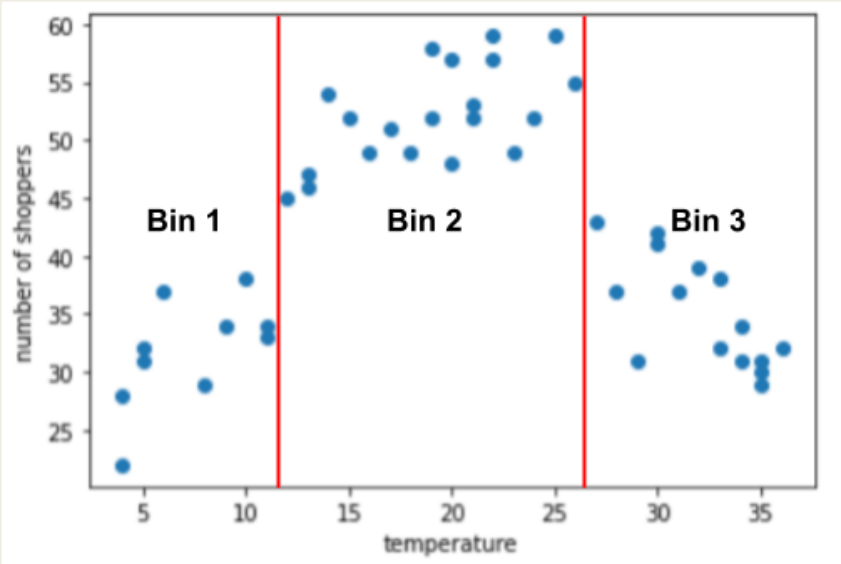
→

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

L’ingénierie des caractéristiques

- Etape 1 : nettoyer les données, imputer les valeurs manquantes
- Etape 2 : le codage catégoriel : convertir les valeurs catégorielles en valeurs numériques
- Etape 3 : le binning ou catégorisation : `from scipy.stats import binned_statistic`



L’ingénierie des caractéristiques

- Etape 1 : nettoyer les données, imputer les valeurs manquantes
- Etape 2 : le codage catégoriel : convertir les valeurs catégorielles en valeurs numériques
- Etape 3 : le binning ou catégorisation
- Etape 4 : la normalisation ou la standardisation