

# Apprentissage Supervisé : Visualisation Supervisée des Données

Nicolas PASQUIER  
Laboratoire I3S (UMR-7271 UCA/CNRS)  
Département Informatique  
Université Côte d'Azur  
<http://www.i3s.unice.fr/~pasquier>

## Objectifs de l'Analyse Exploratoire des Données

---

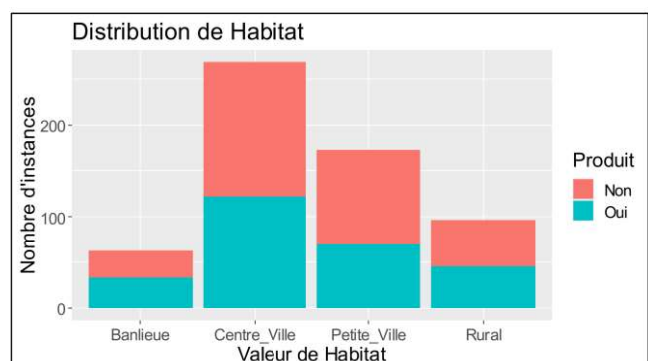
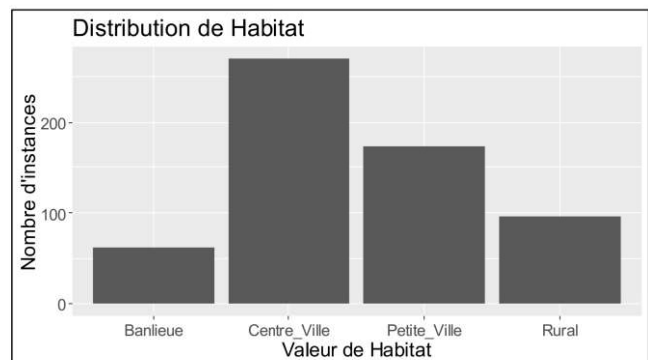
1. Identifier les problèmes de qualité des données
2. Découvrir des indicateurs généraux sur les données (structures, relations, etc.)
3. Identifier les tâches de prétraitement des données nécessaires à la construction de l'ensemble de données :
  - Sélection des données (variables et instances pertinentes)
  - Nettoyage des données (traitement des données bruitées, unification des mesures et représentations, etc.)
  - Transformations des données (rééquilibrage des classes, normalisation ou discrétisation de données numériques, etc.)
- Techniques d'exploration des données
  - Visualisation des données (graphiques mono et multi-dimensionnels)
  - Requêtes sur les données (sélections, dénombrements, etc.)
  - Statistiques descriptives (quantiles, tests de dépendance, etc.)

# Types de Variables

- Outils d'analyse exploratoire utilisées dépendent du type de données
- Variable discrète : variable à domaine de valeur réduit
  - Exemples : Genre  $\in \{\text{Masculin, Féminin}\}$ , Statut\_Civil  $\in \{\text{Marié, Célibataire, Divorcé, Veuf}\}$ , Classement  $\in [1..N]$
  - Différencier codage (e.g. numérique) et sémantique des variables
    - $N^{\circ}\_\text{Dépt\_MC} \in [1..95] \Leftrightarrow \text{Nom\_Dépt\_MC} \in \{\text{Ain, ..., Val-d'Oise}\}$
  - Variables numériques discrètes : moyenne, variance, écart type, ... non-significatifs
  - Sous-types de variables discrètes
    - Variables binaires : deux valeurs possibles (e.g. True/False)
    - Variables ordinales : valeurs ordonnées (e.g. classement d'un concours, valeurs « faible » < « moyen » < « élevé »)
- Variable continue : variable numérique à domaine de valeurs important
  - Exemples : Age  $\in \mathbb{N}$ , Température  $\in \mathbb{Z}$ , Vitesse  $\in \mathbb{R}$

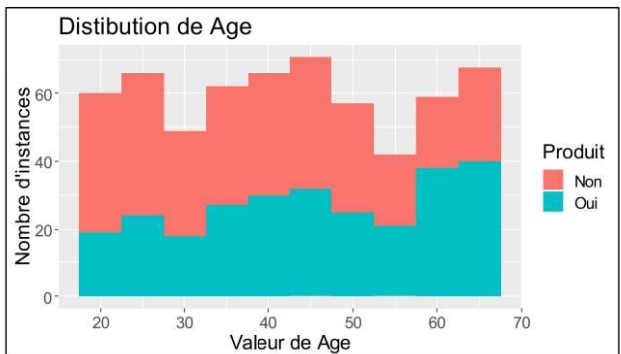
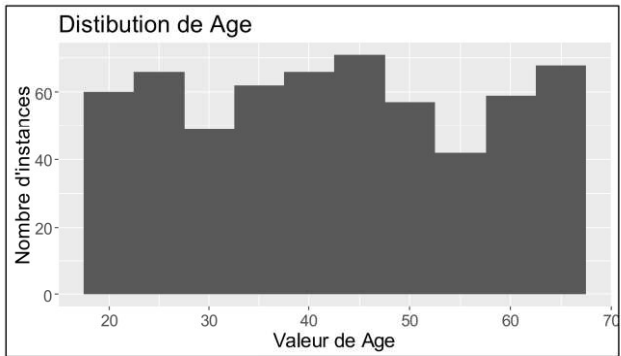
## Histogrammes d'Effectifs de Variables Discrètes

- Chaque barre représente le nombre d'exemples pour une valeur
- Hauteur de la barre : nombre d'exemples avec la valeur
- Valeurs très peu fréquentes
  - Peu d'information fournie
  - Plus difficiles à analyser
- Classes en couleurs
  - Une couleur par classe
  - Part de chaque couleur dans une barre : proportion d'exemples de cette classe avec la valeur
- Interprétation : différences de proportions?



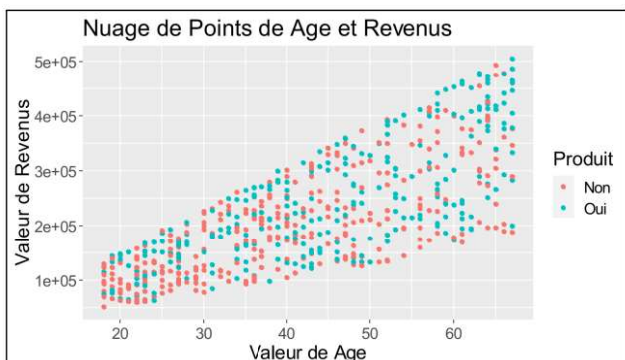
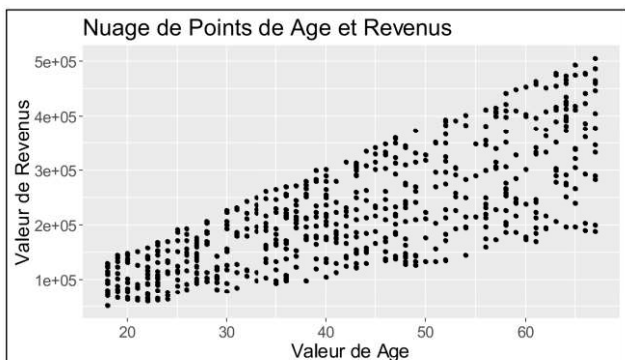
# Histogrammes d'Effectifs de Variables Continues

- Découpage du domaine de valeur par discrétisation
- Barre : intervalle de valeurs
- Hauteur de la barre : nombre d'exemples possédant une valeur dans cet intervalle
- Ajuster la largeur (amplitude) des intervalles ou le nombre d'intervalles
- Classes en couleurs
  - Part de chaque couleur dans une barre : proportion d'exemples de cette classe dans l'intervalle de valeurs correspondant



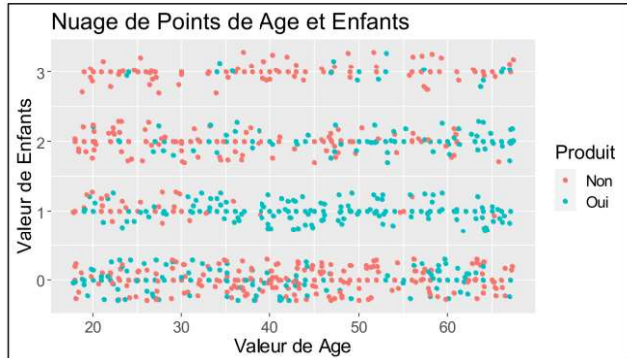
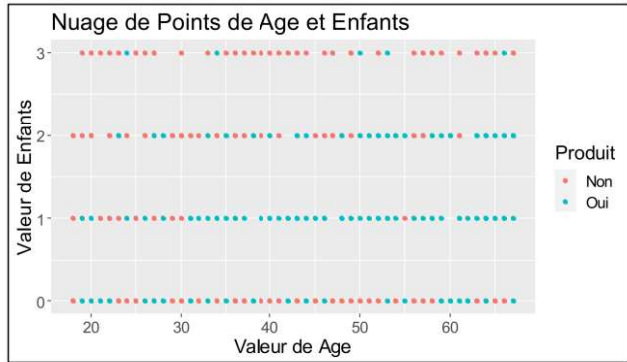
# Nuages de Points de Variables Continues

- Visualisation bidimensionnelle
  - Deux dimensions : deux variables
  - Exemples : représentés comme des points dans l'espace bidimensionnel des données
  - Points positionnés en fonction des valeurs des deux variables
- Identifier des liens (e.g. covariance)
- Classes en couleurs
  - Couleur du point : classe de l'exemple
- Sous-espaces de densité des couleurs différentes : variations des proportions des classes



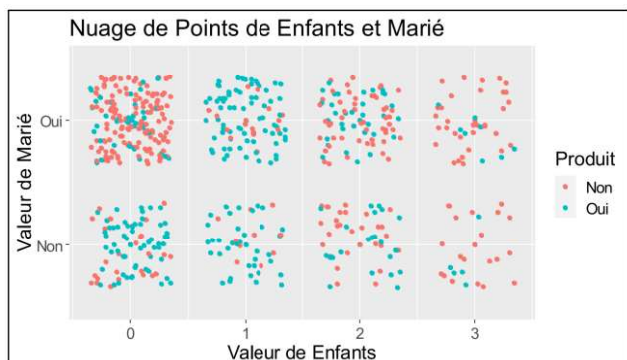
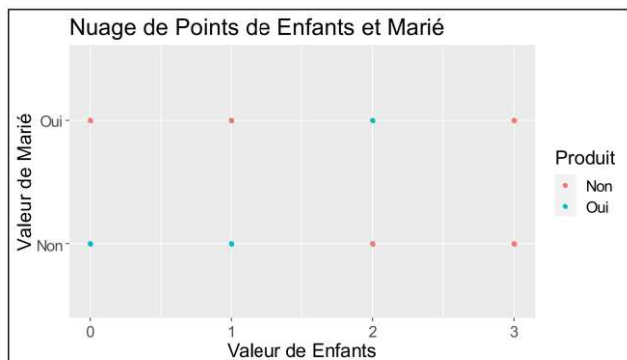
# Nuages de Points de Variables Continues et Discrètes

- Variables discrètes : nombre de valeurs réduit
  - Exemples de même valeur : points sur la même ligne
  - Problème : points superposés dans la représentation graphique (indiscernables)
- Solution : ajouter un léger déplacement aléatoire des points
  - Permet de les distinguer
  - Paramètre appelé Jitter
  - Amplitude du déplacement vertical et/ou horizontal paramétrable



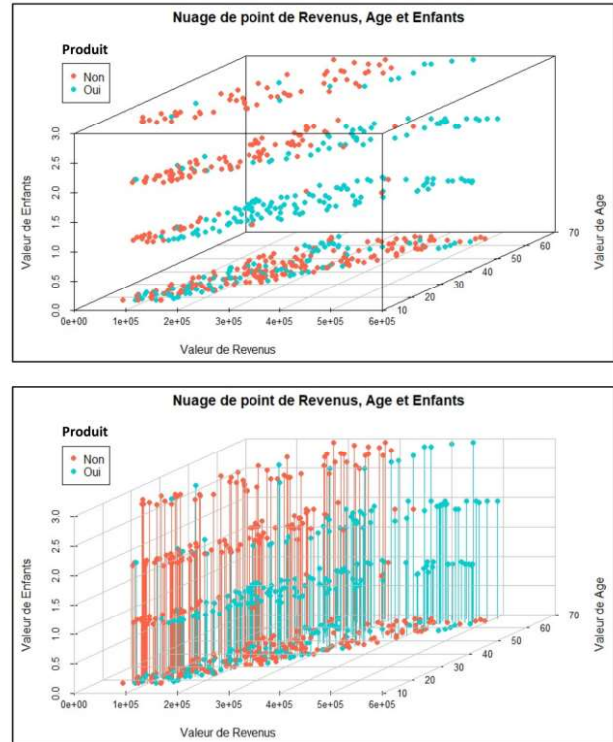
## Nuages de Points de Variables Discrètes

- Deux variables discrètes
  - Nombre de positions des points : produit des cardinalités des domaines de valeurs des variables
  - La couleur affichée pour chaque point est celle du dernier exemple à cette position
  - L'existence ou non d'exemples des deux classes pour chaque combinaison de valeurs est indétectable
- Déplacement aléatoire : Jitter
  - Amplitudes verticale et horizontale paramétrables



# Nuages de Points Tridimensionnels

- Trois dimensions : trois variables
  - Exemples : points dans l'espace tridimensionnel des données
  - Points positionnés en fonction des valeurs des trois variables
  - Classes en couleur
- Sous-espaces tridimensionnels de densité de couleurs variables peuvent être observés
- Selon l'outil utilisé
  - Paramétrage : affichage des grilles, d'indicateurs, etc.
  - Manipulation interactive : rotations, zooms, etc.



## Quantiles : Distribution d'une Variable Numérique

- Variable quantitative : valeur numérique qui quantifie l'une des caractéristiques de l'exemple (e.g. âge d'une personne, vitesse d'un véhicule, etc.)
- Quartiles : les valeurs de la variable sont ordonnées et divisées en 4 partitions de même taille (i.e. en 4 quantiles)
- Exemple : variable prédictive Age de la matrice Buyer (20 exemples)



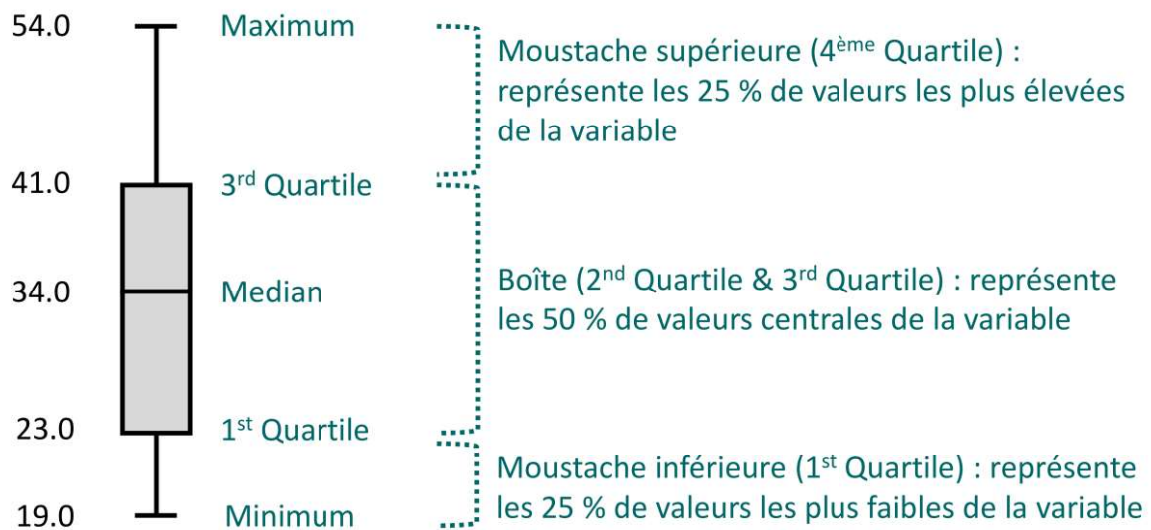
- Chaque quartile représente 25 % des exemples

Minimum	1 <sup>st</sup> Quartile	Median	3 <sup>rd</sup> Quartile	Maximum
19.0	23.0	34.0	41.0	54.0

25% des valeurs
25% des valeurs
25% des valeurs
25% des valeurs

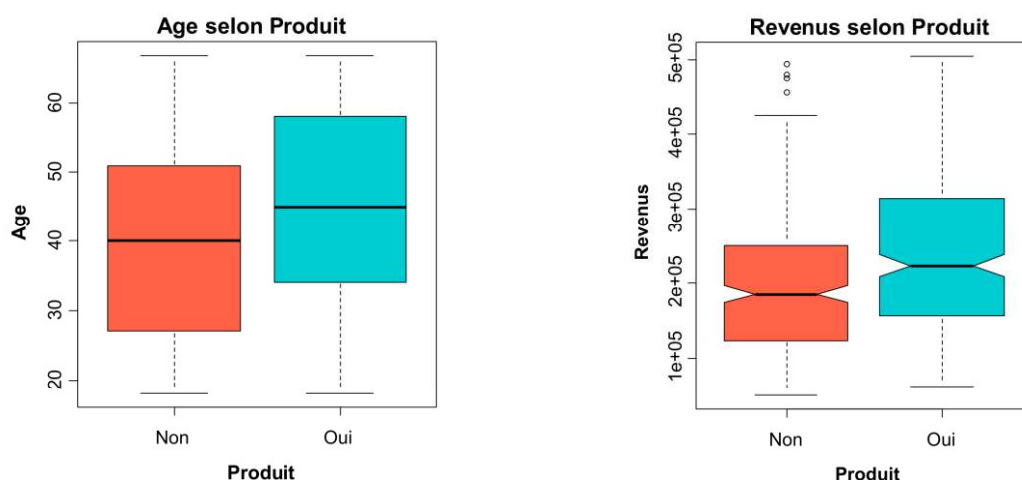
# Boxplot : Représentation Graphique des Quartiles

- Quartiles : représentés graphiquement par un boxplot (boîte à moustaches)
- Exemple : variable prédictive Age de la matrice Buyer (20 exemples)



## Boxplots : Comparaison des Distributions des Classes

- Un boxplot par classe : comparaison des positions et tailles des quartiles



- Identification des exceptions : points au-dessus ou en-dessous des moustaches
- Paramètre *notch* : le non-chevauchement des entailles indique une différence statistiquement significative des valeurs médianes

# Tables de dénombrement et Graphiques Sectoriels

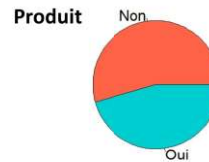
- Dénombrements des cooccurrences des valeurs de variables discrètes
  - Effectifs : comptage des cooccurrences
  - Pourcentages : proportion des cooccurrences
- Quantifier la représentation de l'information dans les données
- Tables et graphique sectoriel

**Effectifs**

Produit	
Non	Oui
326	274

**Pourcentages**

Produit	
Non	Oui
54.3 %	45.7 %



- Table de contingence : cooccurrences des valeurs de deux variables

**Table de Contingence (effectifs)**

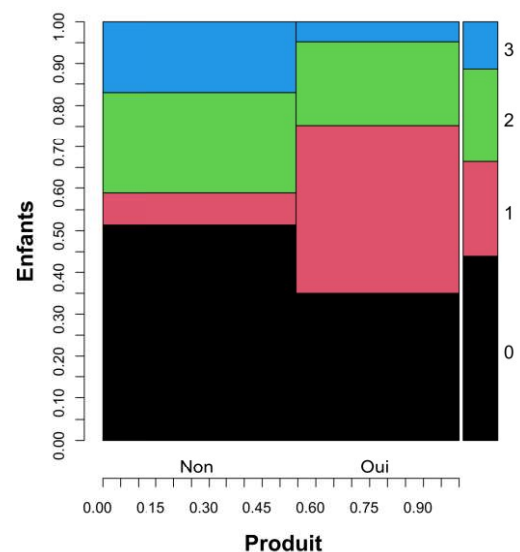
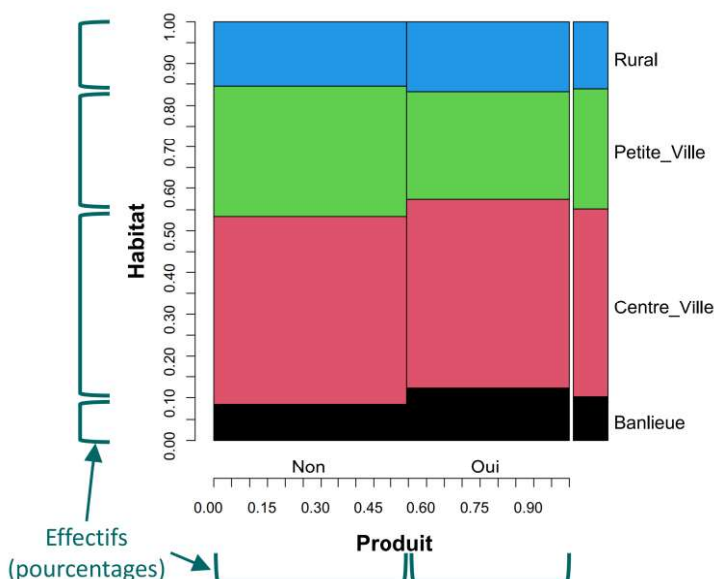
Habitat	Produit	
	Non	Oui
Banlieue	28	34
Centre_ville	146	123
Petite_ville	102	71
Rural	50	46

**Table de Contingence (pourcentages)**

Habitat	Produit	
	Non	Oui
Banlieue	4.7 %	5.7 %
Centre_ville	24.3 %	20.5 %
Petite_ville	17.0 %	11.8 %
Rural	8.3 %	7.7 %

## Graphique en Mosaïque des Proportions de Valeurs

- *Mosaic Plot* de représentation des proportions des cooccurrences
- Largeur et hauteur des boîtes : proportionnelles au nombre d'exemples



# Références et Bibliographie

---

- Librairies R
  - [ggplot2](#) : fonctions avancées de visualisation graphique des données (*Grammar of Graphics* – H. Wickham)
  - [scatterplot3d](#) : traçage de nuages de points tridimensionnels
  - [plot3D](#) : visualisations en 2D et 3D (perspective, coupe, surface, etc.)
  - [plot3Drgl](#) : visualisations interactives des graphiques générés par plot3D
  - [graphics](#) (R Base) : affichage de graphiques en mosaïque
  - [MASS](#) : fonctions de calcul de statistiques descriptives et de visualisation graphiques de données