Exploration des données

CM1-1 : Introduction

Mickaël Martin Nevot

V1.0.0

Exploration des données

Rappel : EDA

- Analyse exploratoire des données :
  - Objectifs :
    - Découvrir et comprendre la structure et les carac. des données
    - Les préparer pour modèles ou algorithmes
  - Résumé des données (moyennes, médianes, écart-types, etc.)
  - Identification des valeurs manquantes ou aberrantes
  - Visualisations (graphiques, histogramme, nuages de points, etc.)

Étape clef pour traitement de données / *machine learning*

Rappel : types de données

- Quantitative (numérique) :
  - Continu
  - Discret
- Qualitative (catégorique) :
  - Nominal
  - Ordinal

Variables

Avec une relation d'ordre

Id	gender	age	ever_married	work_type	Residence_type	avg_glucose_level
9046	Male	67	Yes	Private	Urban	228.69
51676	Female	61	Yes	Self-employed	Rural	202.21
31112	Male	80	Yes	Private	Rural	105.92

Catégorique

Numérique

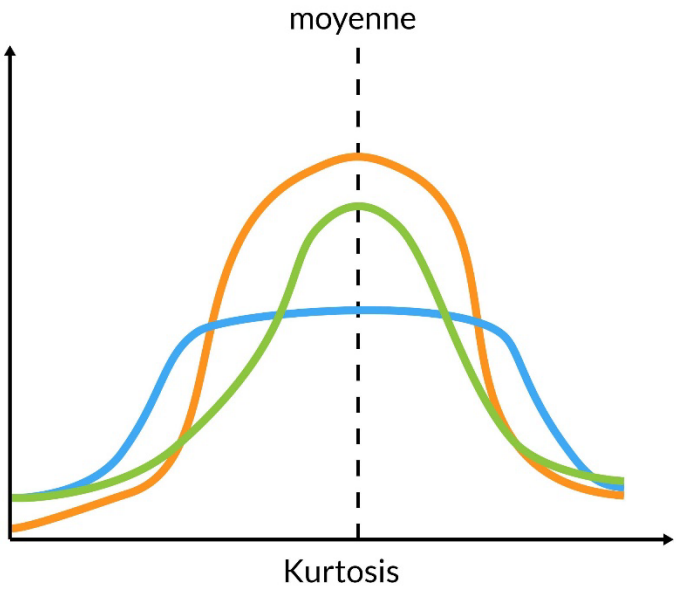
Rappel : statistiques descriptives

- **Synthétiser** : résumer en valeurs clefs
- **Comprendre** : identifier tendances/variabilités/anomalies
- **Préparer l'analyse** : orienter choix/modélisations
- **Comparer des partitions** : identifier des différences

	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	5110.000000	5110.000000	5110.000000	5110.000000	4909.000000	5110.000000
mean	43.226614	0.097456	0.054012	106.147677	28.893237	0.048728
std	22.612647	0.296607	0.226063	45.283560	7.854067	0.215320
min	0.080000	0.000000	0.000000	55.120000	10.300000	0.000000
25%	25.000000	0.000000	0.000000	77.245000	23.500000	0.000000
50%	45.000000	0.000000	0.000000	91.885000	28.100000	0.000000
75%	61.000000	0.000000	0.000000	114.090000	33.100000	0.000000
max	82.000000	1.000000	1.000000	271.740000	97.600000	1.000000

Rappel : stat. univariées

- **Mesures** :
  - **Tendance centrale** : moyenne, médiane, mode
  - **Dispersion** : étendue, variance, écart-type, écart interquartile
- **Forme et distribution** (discrète ou continue) :
  - Symétrie/asymétrie
  - Unimodale/bimodale
  - Multimodale
  - Kurtosis



Rappel : stat. bidimensionnelles

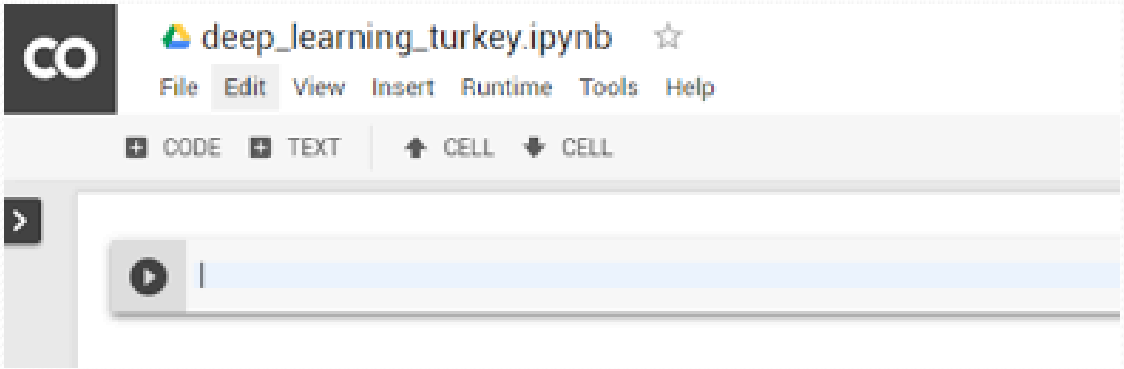
- Données quantitatives :
  - **Covariance** : sens de la variation
  - Corrélation de **Pearson** : mesure de la relation linéaire
  - Corrélation de **Spearman** : éval. basée sur le rang
- Données qualitatives :
  - **Khi -carré** (χ^2) : test d'indépendance

Aussi pour données ordinales

R_X	R_Y	$d_i = R_X - R_Y$	$(d_i)^2$
1	6	-5	25
2	5	-3	9
3	4	-1	1
4	3	1	1
5	2	3	9
6	1	5	25

Pour tester

- Interpréteur en ligne : https://colab.research.google.com
- Documentation : https://pandas.pydata.org/docs

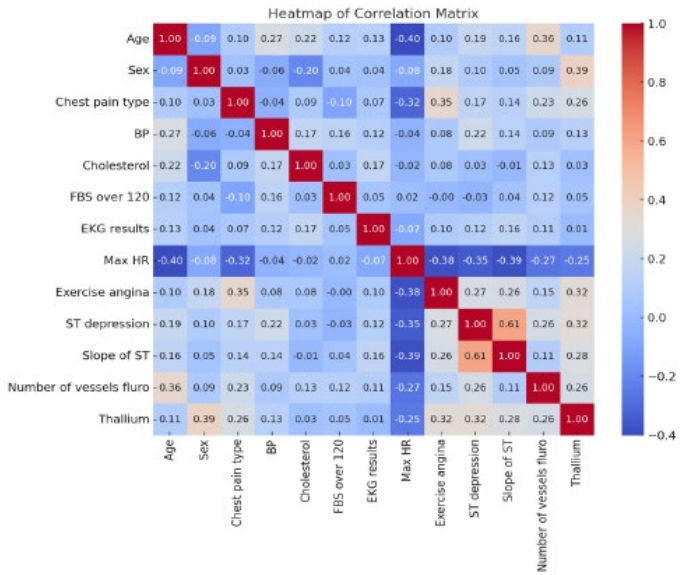


Rappel : visualisation

- **Analyse univariée** :

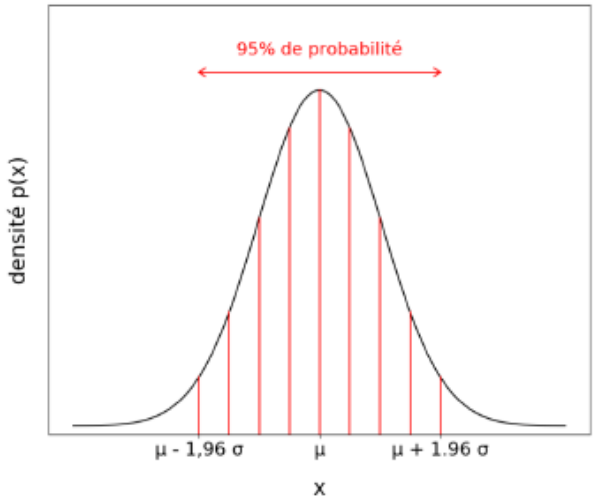
Type de variable	Exemple de visualisation
Quantitative	Histogramme, boîte à moustaches, courbe de densité, diagramme en violon, graphique linéaire (données temporelles)
Qualitative	Diagramme en barres, diagramme circulaire, diagramme de Pareto, nuage de mots (données textuelles)

- **Analyse bivariée** :
  - Nuage de points
  - Carte de chaleur
  - Matrice de corrélation



Rappel : transformation

- Variables quantitatives :
  - **Normalisation** :
    - Statistique *(min-max scaling)* $x' = \frac{x - x_{min}}{x_{max} - x_{min}}$
    - Par mise à l'échelle décimale $x' = \frac{x}{10^k}$
    - Par la somme $x' = \frac{x}{\sum x}$
  - **Standardisation** $x' = \frac{x - \bar{x}}{\sigma}$



De nombreux algorithmes fonctionnent mieux lorsque les variables numériques ont une distribution de probabilité gaussienne : régression linéaire, classification bayésienne

Rappel : transformation

- Variables qualitatives :
  - **Encodage** :
    - Encodage des étiquettes : d'une catégorie à un numérique
    - ***One hot encoding*** : d'une catégorie en variables booléennes
    - *Binary encoding* : d'une catégorie à un binaire
    - *Frequency encoding* : d'une catégorie à une fréquence d'apparition
    - *Target encoding* : d'une catégorie à une moyenne

Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

One-Hot Encoded Data

Team_A	Team_B	Team_C	Points
1	0	0	25
1	0	0	12
0	1	0	15
0	1	0	14
0	1	0	19
0	1	0	23
0	0	1	25
0	0	1	29

Boîte à outils

- Bibliothèques Python :
  - Matplotlib (visualisations 2D)
  - Seaborn (extension de Matplotlib)
  - Ploty (visualisations interactives)
  - YData Profiling (profilage)
- Autres :
  - Tableau : https://www.tableau.com/fr-fr
  - Power BI : https://www.microsoft.com/fr-fr/power-platform/products/power-bi
  - Metabase : https://www.metabase.com
  - Qlikview : https://www.qlik.com/us

Crédits

Auteur

Mickaël Martin Nevot
mmartin.nevot@gmail.com

Carte de visite électronique

Relecteurs

Cours en ligne sur : www.mickael-martin-nevot.com