

M2 Data Science – M2 Intelligence Artificielle

Data Mining

Les bases



Définitions

- « **une donnée** » c'est: « un élément collecté dans le but d'essayer de stocker une information relatant une observation, peut prendre plusieurs formats/type, qui peut être exploité (analyse, traitement), avec un cycle de vie (collecte -> exploitation -> suppression). »
- « **miner** » c'est: « Investiguer/explorer/fouiller profondément un ensemble brut pour y découvrir et en extraire quelque chose »
- Le « **minage de données** » c'est : « Extraction informatique, depuis des ensembles massifs de données brutes (BigData), d'informations afin d'en tirer des connaissances via l'analyse, pour prendre des décision »

Les données réelles, c'est le bordel

Valeurs
manquantes

Valeurs
aberrantes



patient_id	patient_na...	# age	gender	appointme...	booking_d...	doctor	department	billing_am...	follow_up_...
1080	Tammy Williams	76	female	2026/02/26	2024/12/03	Christopher Graham	Cardiology	£425.8	1
1074	Megan Strickland	69	F	05/23/2025	12-Jun-2024	Brandon Lewis	Orthopedics	€344.26	Y
1067	Amanda Schroeder	79	M	30-Nov-2025	August 05, 24	Deanna Edwards	Neurology	€203.34	Y
1072	Anthony Mcpherson	47	F	May 18, 25	09/09/2024	Karen Parsons	General	Rs85.76	No
1092	Benjamin Brown	45		2026/03/07	08/17/2024	Andrea Hernandez	General	\$84.44	1
1057	Joe Sharp	78	1	September 26, 25	08/31/2025	Kimberly Fields	General		Yes
1018	Alexander Harris	22	Female	30-Oct-2025	12/29/2024	Jim Jones	Cardiology	€99.0	No
1074	Raymond Williams	57	M	2025/04/23	28-Mar-2025	James Brooks	Neurology	Rs374.63	No
1041	Dustin Willis	54	female	August 22, 25	2024/11/07	Jamie Hill	General	\$452.37	Yes
1011	Tina Gallagher	45	M	03/16/2026	12-Dec-2025	Tiffany Gates MD	Neurology	€494.97	Yes
1086	Robert Guerrero	54	male	2025/04/24	07-Oct-2024	Shane Brown	Cardiology	£91.91	Y
1041	Karen Walker	68	F	2026/03/07	2024/04/24	Natalie White	General	\$82.88	Yes
1057	Chris Rice	61	1	2025/07/31	May 05, 24	Amy Olson	Neurology	Rs277.14	N

Format inconsistant

Pour bien commencer, apprendre à nettoyer ses données

Place à la pratique !

- Sur [kaggle](#), téléchargez l'un des jeux de données suivants:
 1. [FIFA 21 messy](#)
 2. [Multi-Hospital Lab Results](#)
 3. [Real Estate Raw Dataset](#)
 4. [Messy E-Commerce Sales](#)
 5. [Messy Employee Dataset](#)
 6. [Amazon Laptop Messy dataset](#)
 7. [Customer Messy Data](#)
- Identifiez ce qui empêche une bonne extraction d'informations
- Déterminez ce qui doit être mis en place pour améliorer l'extraction d'informations
- Nettoyez votre jeu de données



Nettoyer ses données: Gérer les redondances d'informations

ID	Nom	Prenom	Nom_Complet	Date_Naissance
006	Trevelyan	Alec	Alec Trevelyan	12/03/1928
007	Bond	James	James Bond	11/11/1920
006	Trevelyan	Alec	Alec Trevelyan	12/03/1928

Informations redondantes

Informations redondantes

Bonnes pratiques:

- Bien vérifier la concordance des lignes et colonnes concernées
- Pour les lignes:
 - Supprimer les doublons
- Pour les colonnes:
 - Si pas de données manquantes:
 - Privilégier le détail
 - Sinon:
 - Privilégier les colonnes sans données manquantes

Nettoyer ses données: Gérer les formats inconsistants

ID	Date	Valeur	Unite
0	2024-03-14 14:52	8,1	Mmol/L
1	02-28-2023	14,3	%
2	12/04/2012	7.58	Mmol/L
3	06-18-2024 0811	5.2	UI/L
4	202207141312	13.2	%

Format
variable

Unité de mesure
variable

Bonnes pratiques:

- Choisir le format le plus précis et convertir les autres valeur de la colonne
- Choisir l'unité de mesure la plus commune et convertir les autres valeurs

OU

- Faire une colonne par unité et convertir les valeurs pour chaque colonne
 - Ex: Valeur_Mmol_L, Valeur_percent, Valeur_UI_L

Nettoyer ses données:

Gérer les variables catégorielles

ID	Date	Label	Valeur	Unite
0	2024-03-14	HbA1c	8,1	Mmol/L
1	2023-02-28	LDL	14,3	%
2	2012-12-04	Hemo Glyc	7.58	Mmol/L
3	2026-06-18	Hb A1c	5.2	UI/L
4	2022-07-14	HDL	13.2	%

Catégories
non-fixes

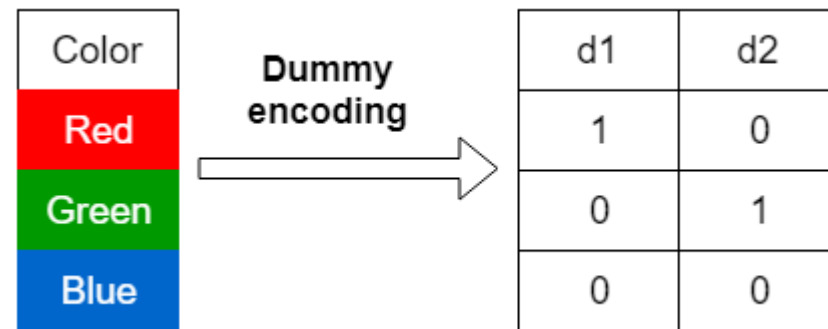
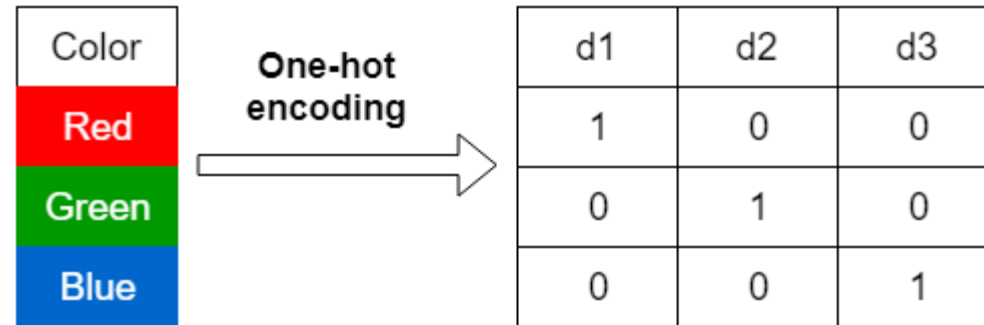
Bonnes pratiques:

1. Identifier les différents noms utilisés pour un même « objet » (plus de détails dans le cours 6)
2. Choisir le label le plus adaptés parmi ces noms
3. Remplacer le reste par le label choisi

=> Il peut aussi parfois avoir besoin d'encoder ces variables

Nettoyer ses données:

Gérer les variables catégorielles – bonus encodage¹



¹[Encoding Categorical Variables: One-hot vs Dummy Encoding | Towards Data Science](#)

Nettoyer ses données:

Gérer les données manquantes et aberrantes (outliers)

ID	Nom	Prenom	Age
0	Bouchard	Gérard	88
1	Moulin	Monique	143
2	Dupont	Yannick	27
3	Dupond	Anne-Laure	-99
4	Columbo	NA	NA

Valeurs
aberrantes

Valeurs
manquantes

Bonnes pratiques:

- Si pas d'autres informations importantes, supprimer les lignes concernées
- Repérer les *outliers* ET les traiter comme données manquantes
- Pour les données manquantes:
 - Inférer leurs valeurs OU considérer que l'absence d'information est une information en soi

Nettoyer ses données

Synthèse

Nettoyer des données => Choisir

Choisir => introduire des biais

Donc, nettoyer des données => de biens les comprendre

(Une variable qu'on ne comprend pas, il vaut mieux éviter de baser des conclusions dessus)

Le fait de « nettoyer » des données peut faire débat.

À lire: [Debates in the Digital Humanities | Against Cleaning](#)

Un *outlier* peut aussi être une anomalie intéressante à repérer (Cf. cours 2)

Comprendre ses données - descriptions

Place à la pratique !

- Reprenez le jeu de données téléchargé précédemment:
 1. [FIFA 21 messy](#)
 2. [Multi-Hospital Lab Results](#)
 3. [Real Estate Raw Dataset](#)
 4. [Messy E-Commerce Sales](#)
 5. [Messy Employee Dataset](#)
 6. [Amazon Laptop Messy dataset](#)
 7. [Customer Messy Data](#)
- **Faites des stats sur chaque colonne** (moyennes, médianes, etc.)
- **Faites des visualisations** (histogrammes, boxplots, etc.)
- **Réfléchissez et débâtez** sur ce qui est le plus pertinent à faire pour comprendre chaque type de variable



Comprendre ses données

Identifiez le type de chaque variable (ou *feature*)

- Numérique

- Discrète / Entiers (int)
- Continue / Réels (float)



- Moyenne ? Écart-type ?
- Loi normal ? Loi de puissance ?

- Nominal

- Ordonnées (ordinal) ou non
- Intervalles numériques



- Valeurs possibles ?
- Répartitions des valeurs ?

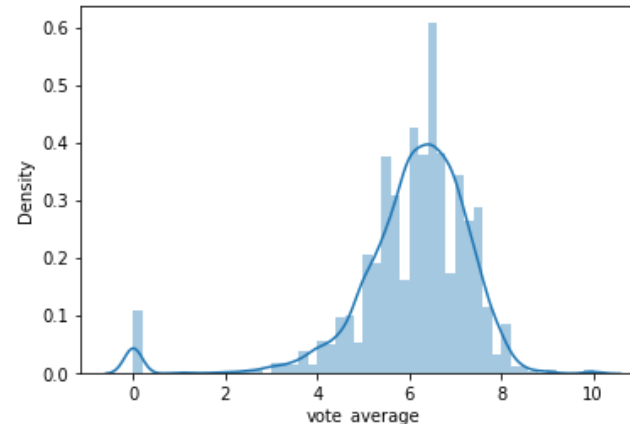
- Dates (timestamp)

- Cf. cours 4

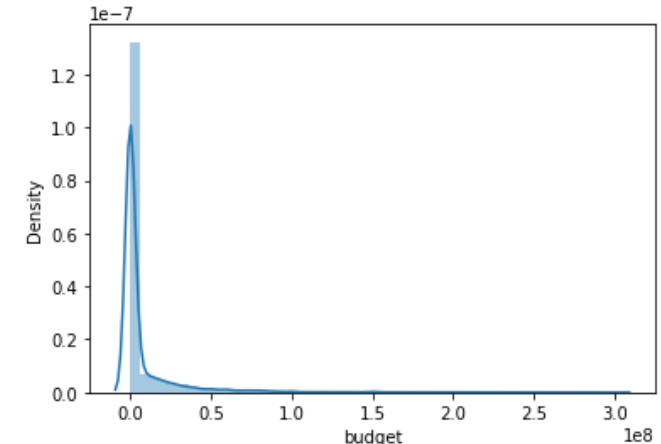
Comprendre ses données

Bonnes pratiques sur les variables numériques

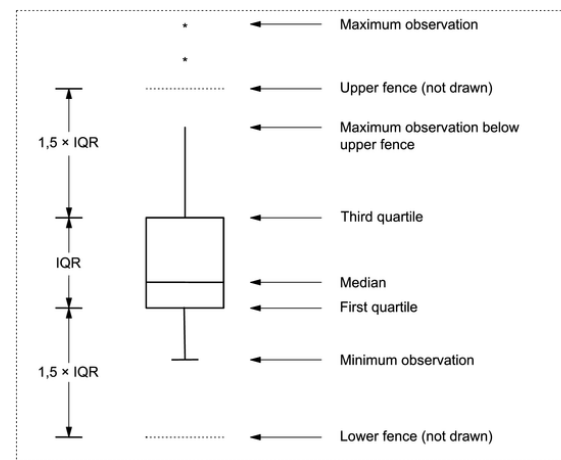
- Identifier si c'est une variable discrète ou continue
- Faire quelques stats générales, mais attention ! Toujours associer:
 - La moyenne avec l'écart-type
 - La médiane avec le min et le max (et quartiles si possible)
- Observer la répartition des valeurs
 - Avec un histogramme (toujours)
 - Avec boxplot ou violonplot



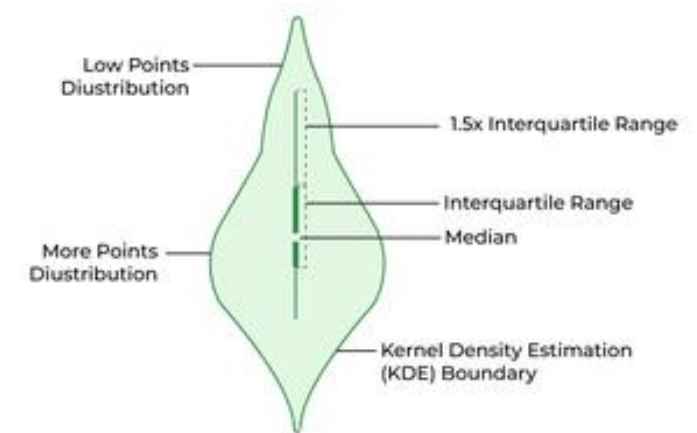
Histogramme (loi normale)



Histogramme (loi de puissance)



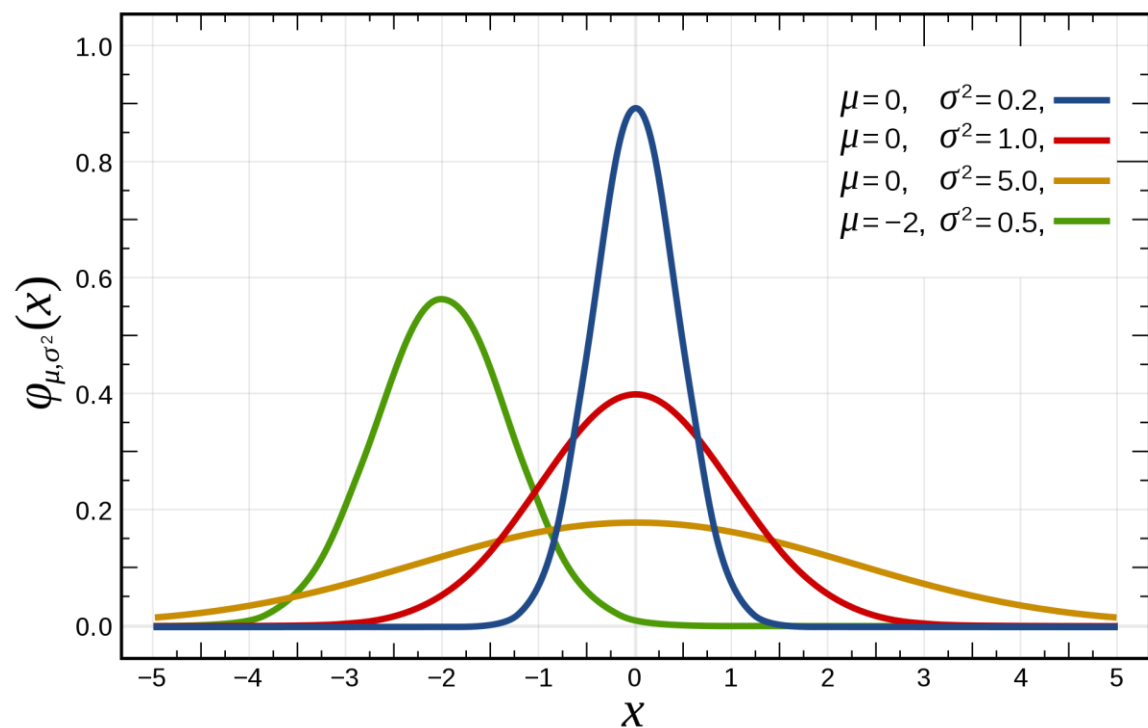
Boîtes à moustache (boxplot)



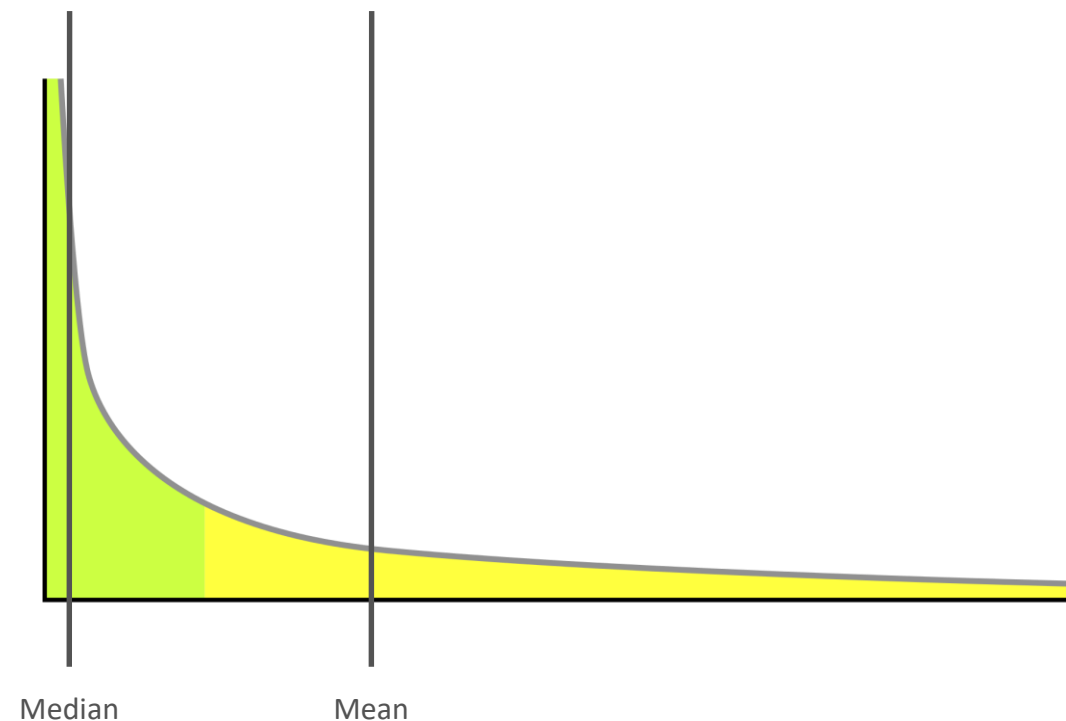
Violon plot

Comprendre ses données

Focus: Loi normale et de puissance



Variables suivant une loi normale



Variable suivant une loi de puissance (power law)



En théorie, toute variable suit une loi normale.
En pratique, la plupart des variables suivent une loi de puissance (ou un mix)

Comprendre ses données

Bonnes pratiques sur les variables nominales

- Identifier si c'est les valeurs sont ordonnées ou non
 - Ordonnées : Petit, Moyen, Grand
 - Non ordonnées : Bleu, Vert, Rouge
- Si les valeurs sont ordonnées, elles **peuvent** être traitées comme une variable numérique
- Dans tous les cas: **visualiser la répartition des valeurs (barplots/histo) !**

Comprendre ses données - corrélations

Place à la pratique !

- Reprenez le jeu de données téléchargé précédemment:
 1. [FIFA 21 messy](#)
 2. [Multi-Hospital Lab Results](#)
 3. [Real Estate Raw Dataset](#)
 4. [Messy E-Commerce Sales](#)
 5. [Messy Employee Dataset](#)
 6. [Amazon Laptop Messy dataset](#)
 7. [Customer Messy Data](#)
- **Cherchez les corrélations** entre les différentes variables
- **Visualisez la relations entre les variables corrélées** (Linéaire ? Monotone ? Etc.)
- **Réfléchissez et débâtez** sur les corrélations trouvées

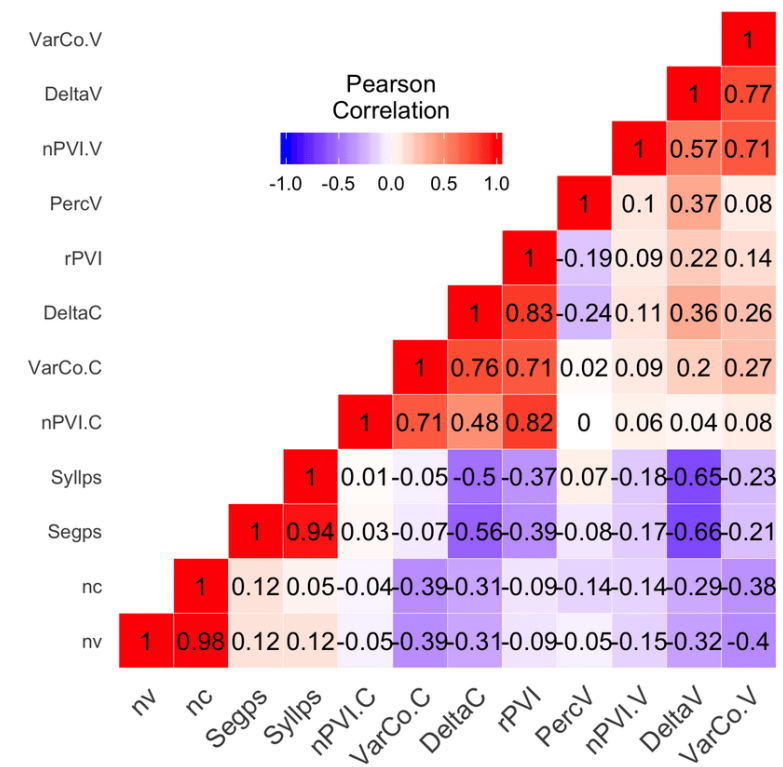
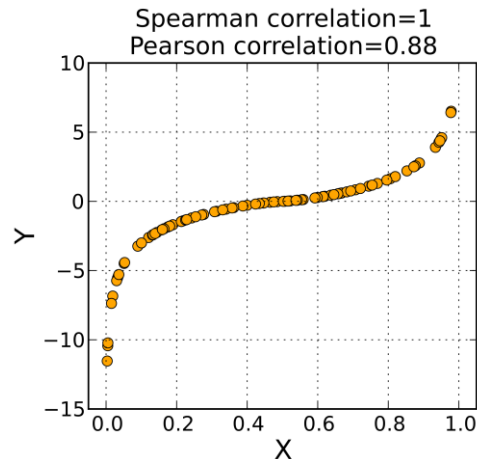
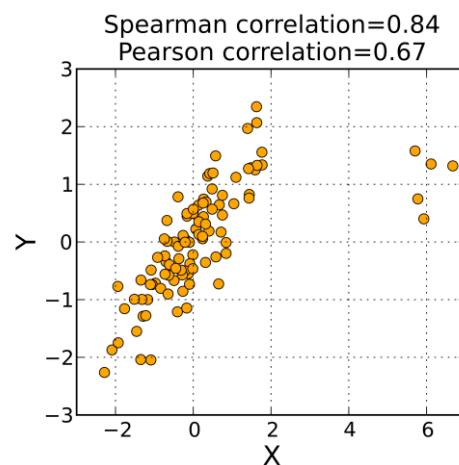
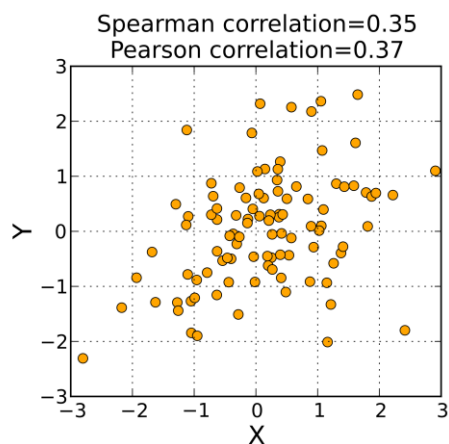


Comprendre ses données

Identifier les corrélations

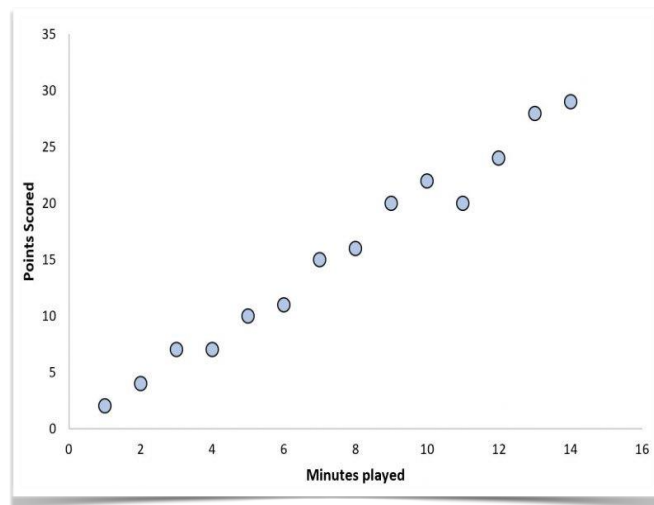
- Deux coefficients souvent utilisés:

- Pearson: $\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$
 - La plus commune, mais assume une relation linéaire
- Spearman : $r_s = \rho_{R(X),R(Y)} = \frac{\text{cov}(R(X),R(Y))}{\sigma_{R(X)} \sigma_{R(Y)}}$
 - Trie les corrélations, plus adaptées aux relations non linéaire

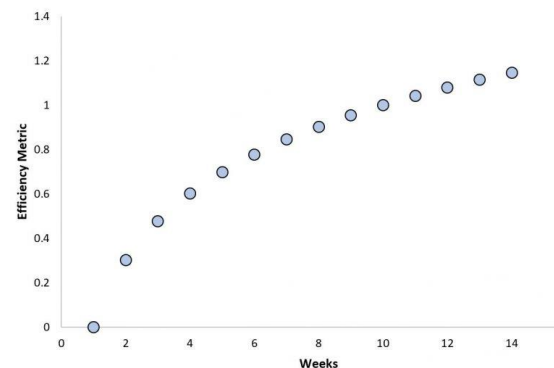
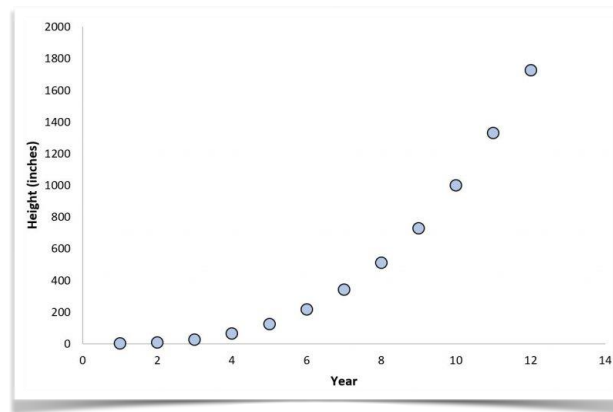


Comprendre ses données

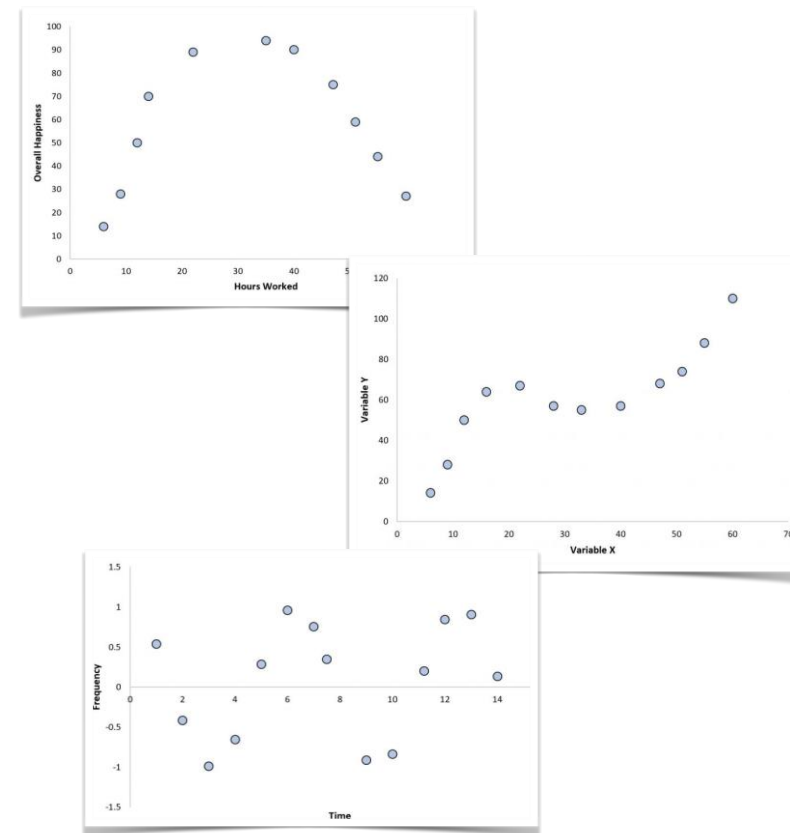
Observer les relations entre variables



Exemple de relation
monotone et linéaire



Exemples de relations
monotone et non-linéaire

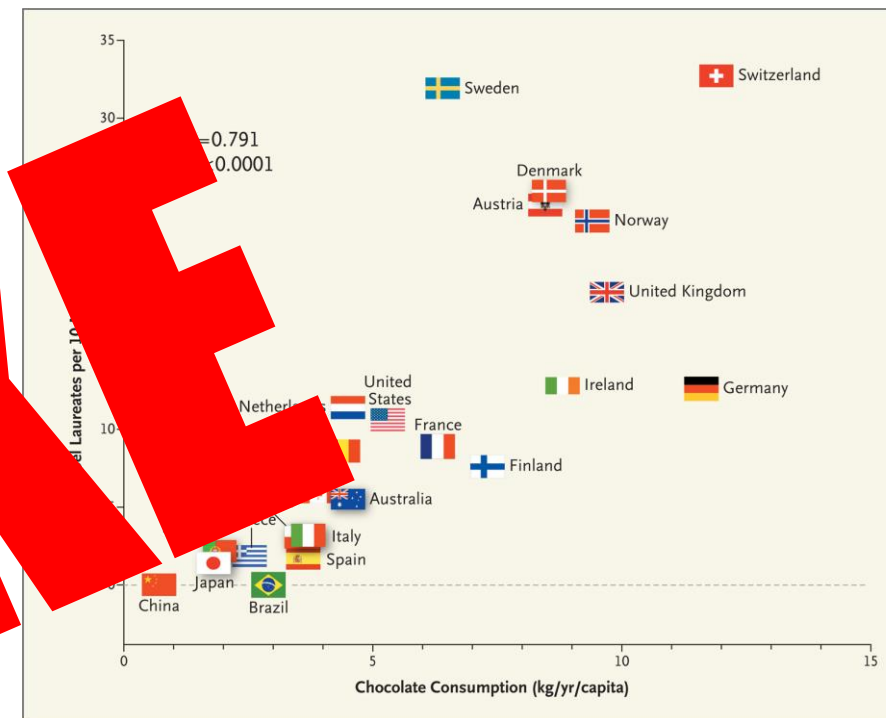


Exemples de relations
non-monotone et non-linéaire

Comprendre ses données

Corrélations et causalités

“There was a close, significant linear correlation ($r=0.791$, $P<0.0001$) between chocolate consumption per capita and the number of Nobel laureates per country with 11 persons in a total of 21 countries [1]



Conclusion: “Chocolate consumption enhances cognitive function” [1]

Comprendre les liens dans les données

Bonnes pratiques:

- Commencer par faire une matrice de corrélations
 - Avec le coefficient de Spearman de préférence
 - Mais toujours calculer Spearman et Pearson
- Visualiser et identifier le type de relation liant deux variables corrélées
- **/!\ CORRÉLATION N'EST PAS CAUSALITÉ /!**

Place à l'action !



Votre Mission:

- Vous rendre sur: <https://github.com/a-t-richard/20262027-UE-DataMining-Project>
- Faire un fork (par groupe) et cloner le repos
- Trouver ce qu'il se cache dans ces données !