

# **IFT870/BIN710**

## **Forage de données**

### **Thème 2 : Exploration de données**

Aida Ouangraoua

Département d'informatique



Université de  
Sherbrooke

# **Partie I : Théorie**

# Type de données

## **□ Données sous forme d'enregistrement**

- ❖ Vecteurs de valeurs d'attributs (ex : matrices numériques)
- ❖ Données de documents (ex: matrices documents-termes)
- ❖ Données de transactions (ex. ensembles d'items)

## **□ Données structurées sous forme de séquences ou graphes**

- ❖ Ordonnées : vidéo (séquence d'images), données séquentielles (séquences de données, ex: séquences biologiques), données temporelles (série de données ordonnées dans le temps)
- ❖ Graphes : réseaux sociaux, données du web

## **□ Données spatiales et de multimédia**

- ❖ Données spatiales (cartes)
- ❖ Images

# Type de données

- ❑ Vecteurs de valeurs d'attributs (ex : matrices numériques)
- ❑ Données de documents (ex: matrices documents-termes)

	équipe	coach	pays	balle	score	jeu	gagné	perdu	saison
<b>Document 1</b>	3	0	5	0	2	6	0	2	0
<b>Document 2</b>	0	7	0	2	1	0	0	3	0
<b>Document 3</b>	0	1	0	0	1	2	2	0	3

- ❑ Données de transactions (ex. ensembles d'items)

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

# Ensemble de données

- ❑ Ensemble d'objets (ou entités, enregistrements, vecteurs, points)
- ❑ Objets représentés par des valeurs d'attributs
- ❑ Ensemble d'objets représentés sous forme de matrice  $M$ 
  - ❖ Lignes : objets
  - ❖ Colonnes : attributs (ou dimensions, variables, caractéristiques, propriétés «features»)
  - ❖  $M(i,j)$  : valeur de l'attribut  $j$  pour l'objet  $i$ .

	équipe	coach	pays	balle	score	jeu	gagné	perdu	saison
Document 1	3	0	5	0	2	6	0	2	0
Document 2	0	7	0	2	1	0	0	3	0
Document 3	0	1	0	0	1	2	2	0	3

# Type d'attributs

## ☐ Nominal ou catégoriel (discret)

- ❖ Valeurs représentant des classes ou catégories

Exemple : continent = {NA, SA, AF, AS, OC, EU}

## ☐ Binaire (discret) : Nominal à 2 valeurs (symétrique ou asymétrique)

## ☐ Ordinal (discret)

- ❖ Valeurs ordonnées

Exemple : Cote : A+, A, A-, B+, B, B-, ...

## ☐ Numérique:

- ❖ Discret : valeurs entières
- ❖ Continu : valeurs réelles

# Descriptions statistiques

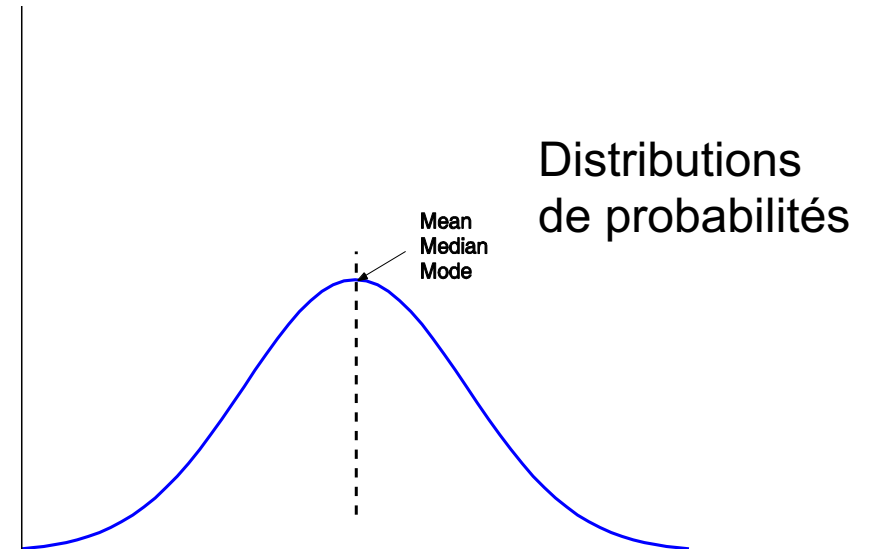
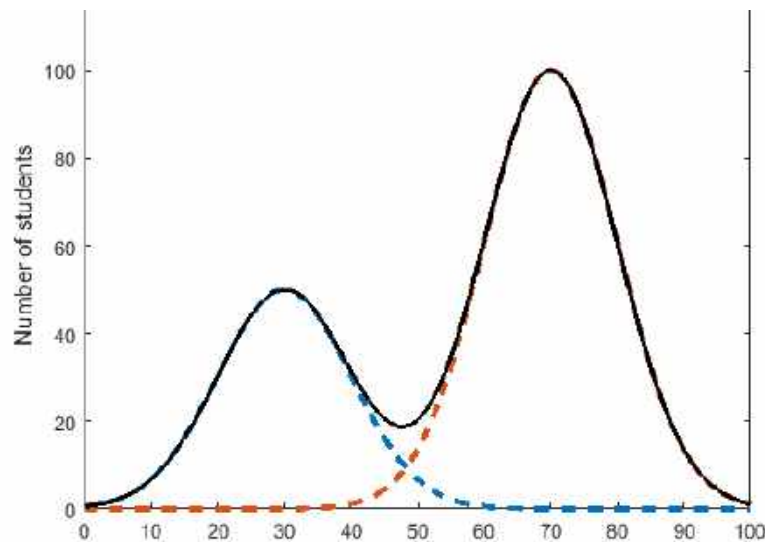
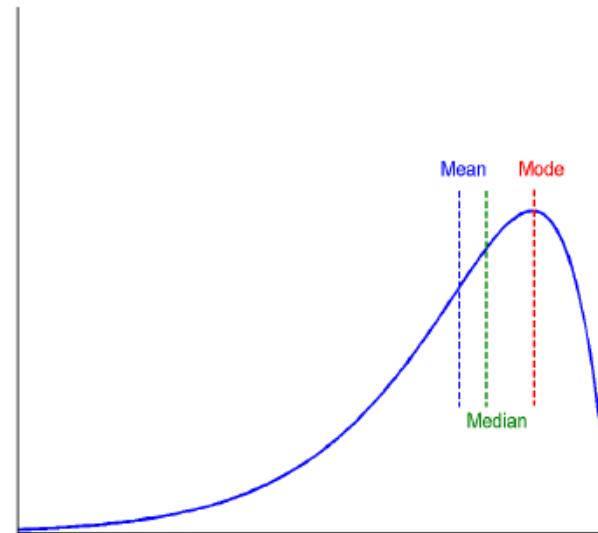
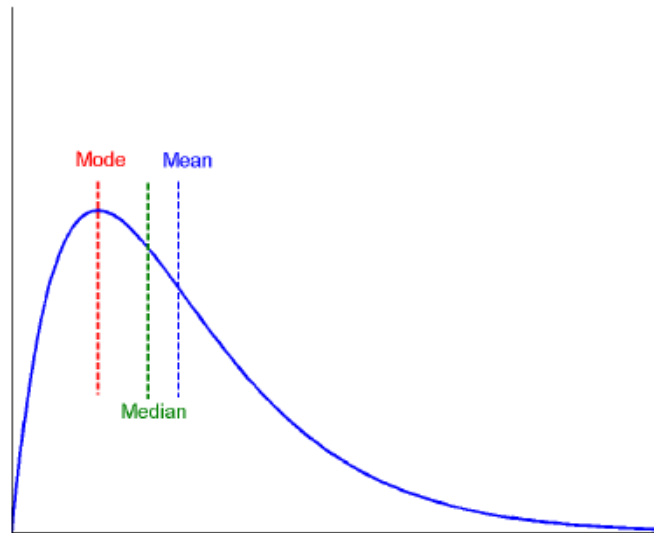
- ❑ Permet de mieux connaître les données pour identifier des questions, et des modèles potentiellement adéquats
- ❑ Synthèse des valeurs d'un attribut
  - ❑ Minimum, maximum, moyenne, médiane, mode, estimation de la probabilité des valeurs (distribution de probabilités)
- ❑ Synthèse de la dispersion des valeurs
  - ❖ Variance, écart-type, quartile, écart inter-quartile, valeurs aberrantes, nombre d'objets par intervalle de valeurs (histogramme), distribution des probabilités

# Descriptions statistiques : Synthèse des valeurs

- ❑ Minimum, Maximum, Moyenne  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
- ❑ Médiane: valeur  $m$  telle qu'il y a autant de valeurs supérieures à  $m$  que de valeurs inférieures à  $m$ .  
Exemple : Médiane([1,2,10,12,13]) = 10 ; Médiane([2,10,12,13]) = 11
- ❑ Mode : valeur la plus fréquente

# Descriptions statistiques : Synthèse des valeurs

- ❑ Mode : valeur la plus fréquente
  - ❖ Peut-être unimodal, bimodal, trimodal



Distributions  
de probabilités

# Descriptions statistiques : Synthèse de la dispersion

❑ Variance:  $\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$  Écart-type:  $\sigma$

❑ Quartiles :

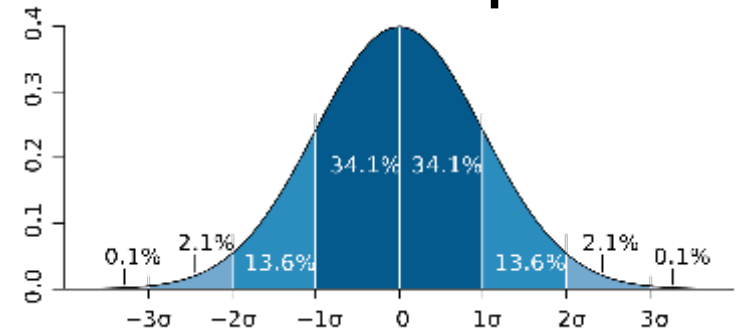
❑ 1<sup>er</sup> quartile Q1 : médiane des valeurs inférieures à la médiane

❑ 2<sup>e</sup> quartile Q2 : médiane

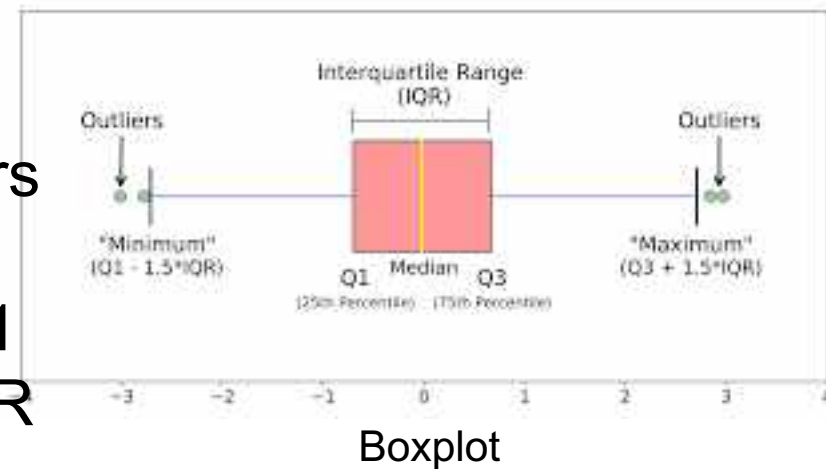
❑ 3<sup>e</sup> quartile Q3 : médiane des valeurs supérieures à la médiane

❑ Écart inter-quartile (IQR) =  $Q3 - Q1$

❑ Valeurs aberrantes :  $< Q1 - 1.5 \cdot IQR$   
ou  $> Q3 + 1.5 \cdot IQR$

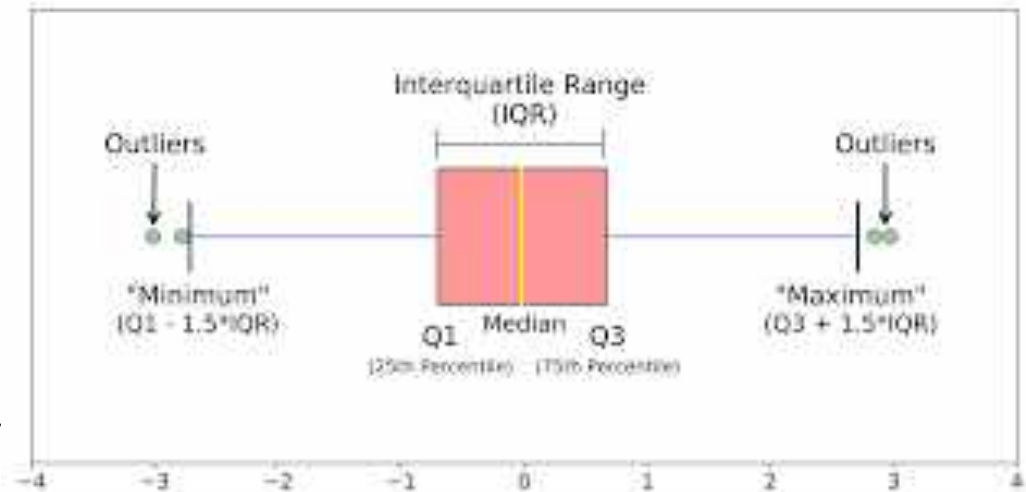
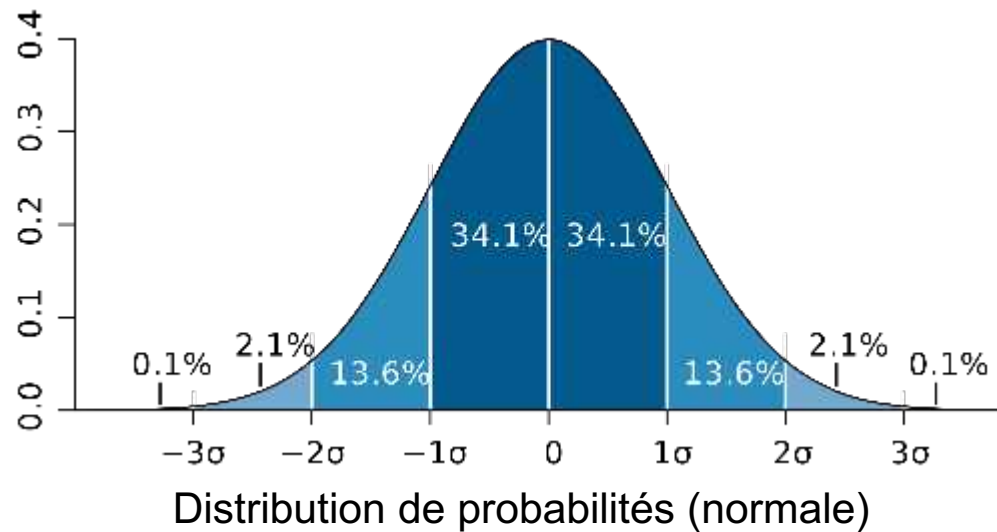


Distribution de probabilités (normale)

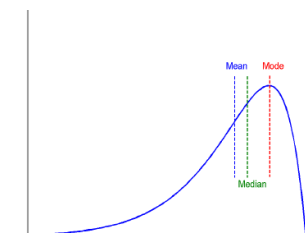
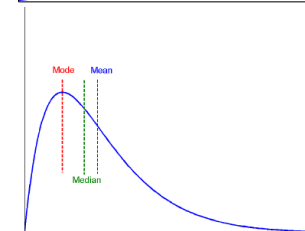
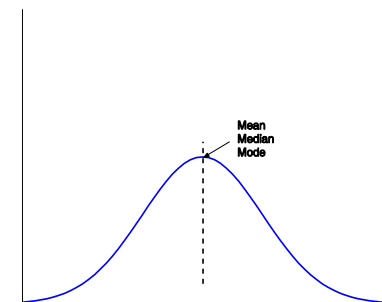
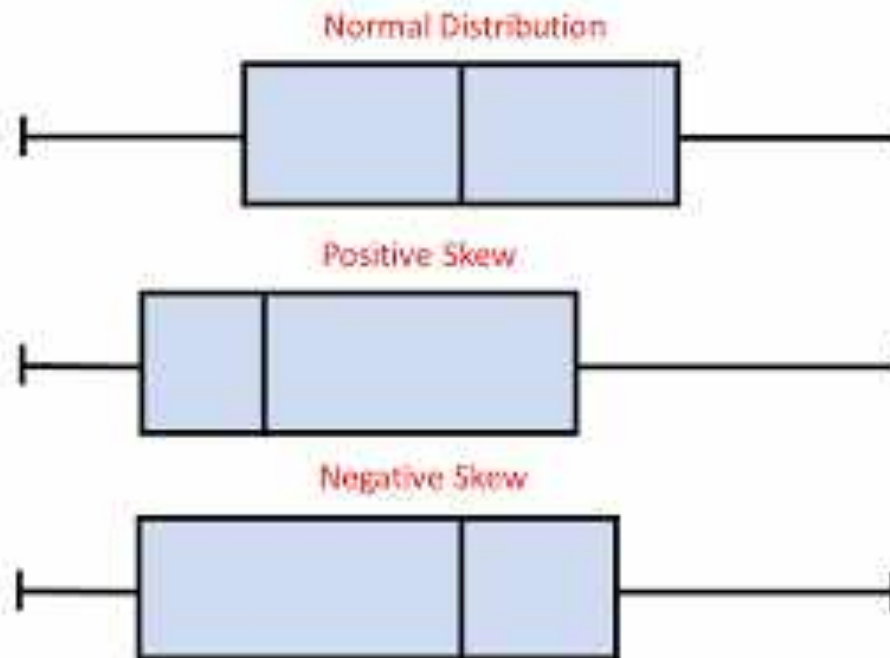


Boxplot

# Descriptions statistiques : Synthèse de la dispersion



Boxplot



# Visualisation des descriptions statistiques : utilité

## □ Pré-traitement

- ❖ Aide à l'exploration des données
- ❖ Donne un aperçu et une vue d'ensemble qualitative de l'espace des données
- ❖ Permet d'identifier des tendances, structures, irrégularités, relations ou modèles entre les données
- ❖ Guide pour trouver des régions intéressantes et des paramètres appropriés pour une analyse quantitative approfondie

## □ Post-traitement

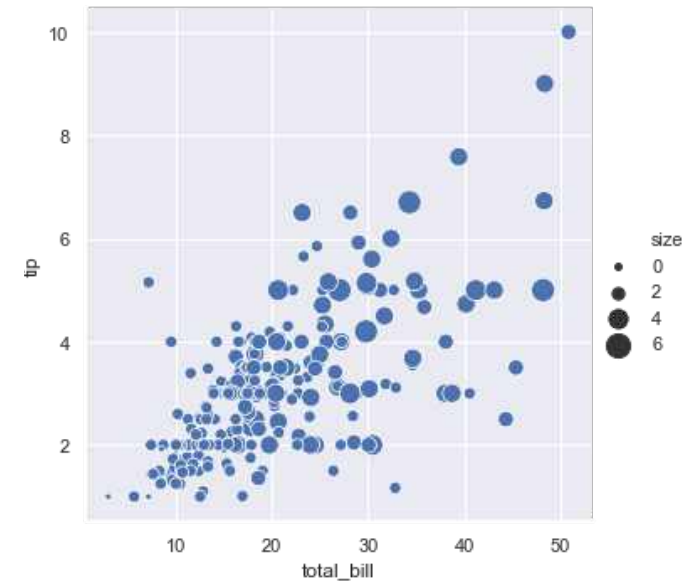
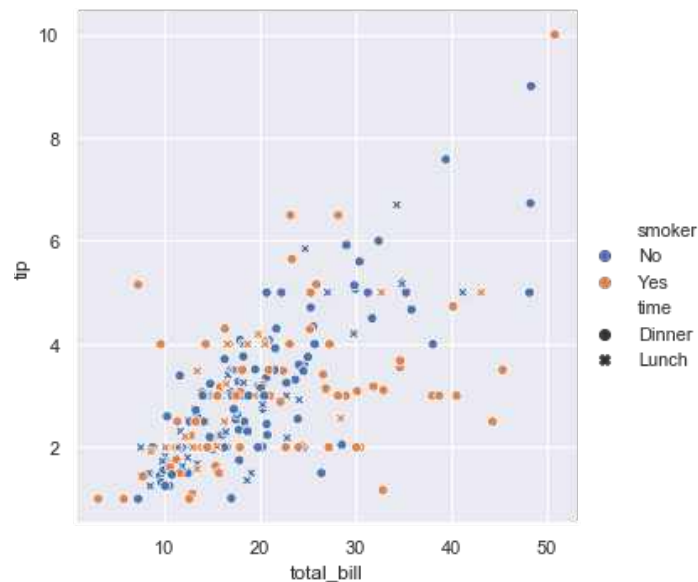
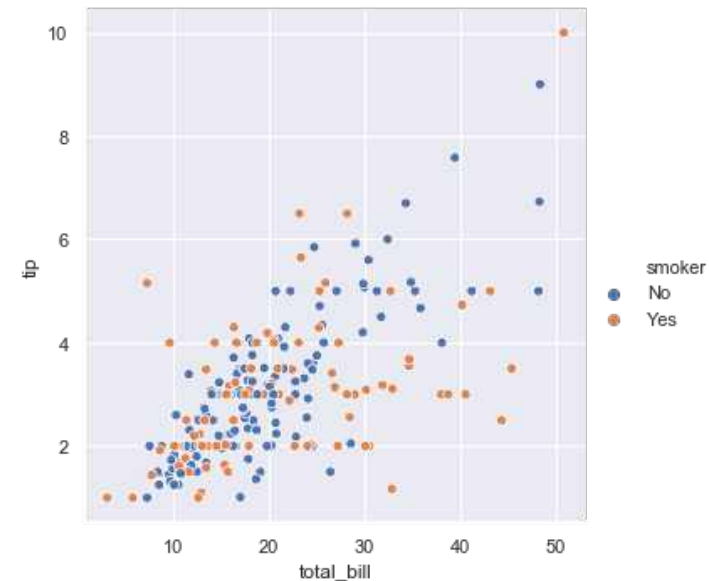
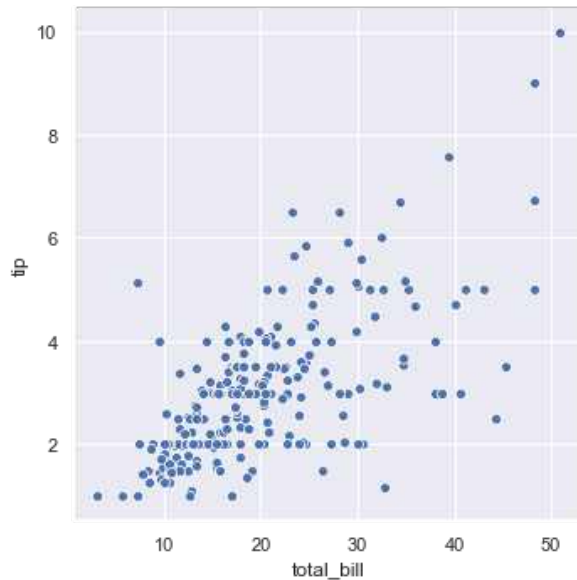
- ❖ Fournit une preuve visuelle des modèles dérivées

# Visualisation des descriptions statistiques : types de graphique

- ☐ Relation entre deux variables (possibilité de distinguer suivant d'autres variables)
- ☐ Régression entre deux variables (possibilité de distinguer suivant d'autres variables)
- ☐ Relation entre deux variables dont une catégorielle (possibilité de distinguer suivant d'autres variables)
- ☐ Distribution univariée et bivariée
- ☐ Matrice de couleurs pour des données rectangulaires
- ☐ Grille multi-graphique
  - ☐ Groupement suivant deux variables (lignes, colonnes)
  - ☐ Grille de relations deux-à-deux entre toutes les variables
  - ☐ Relation ou distribution bivariée couplée aux 2 distributions univariées
- ☐ Visualisation en 3D

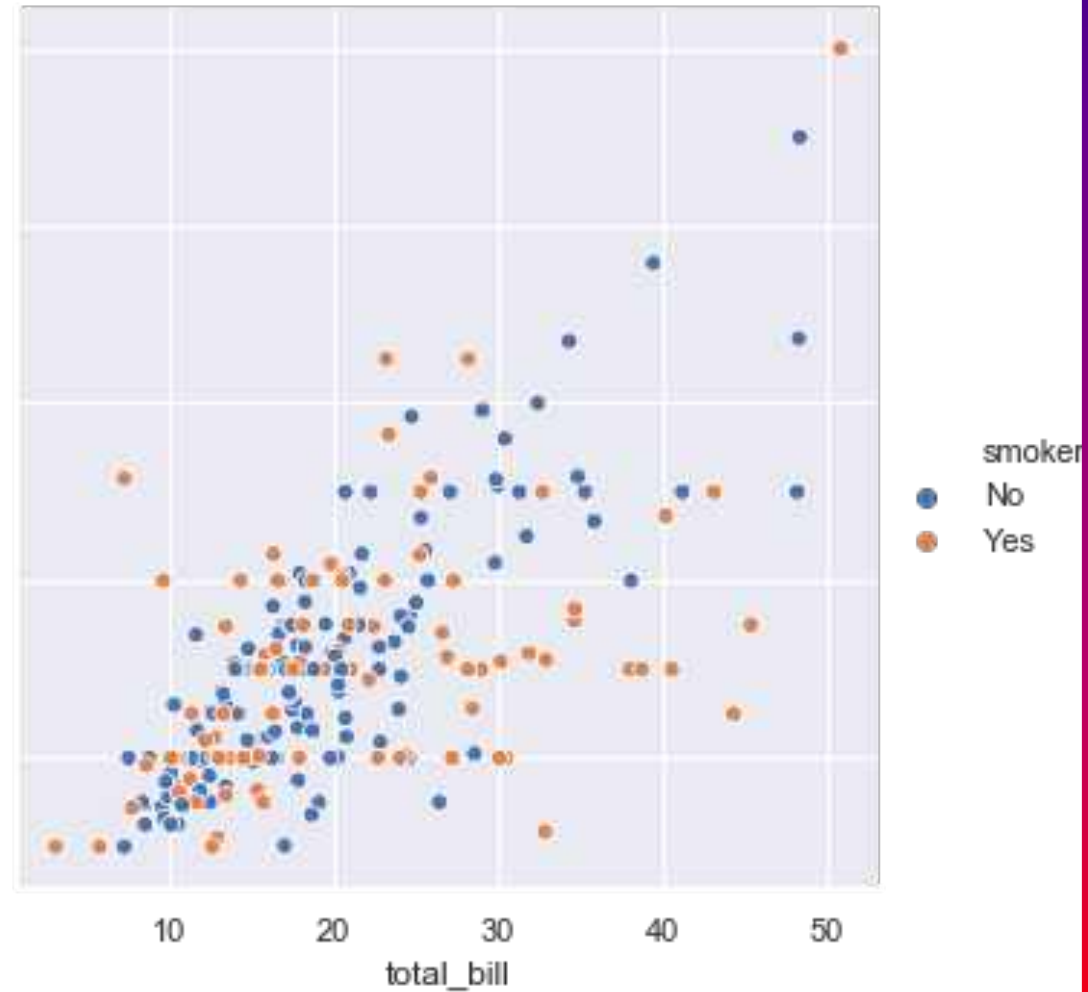
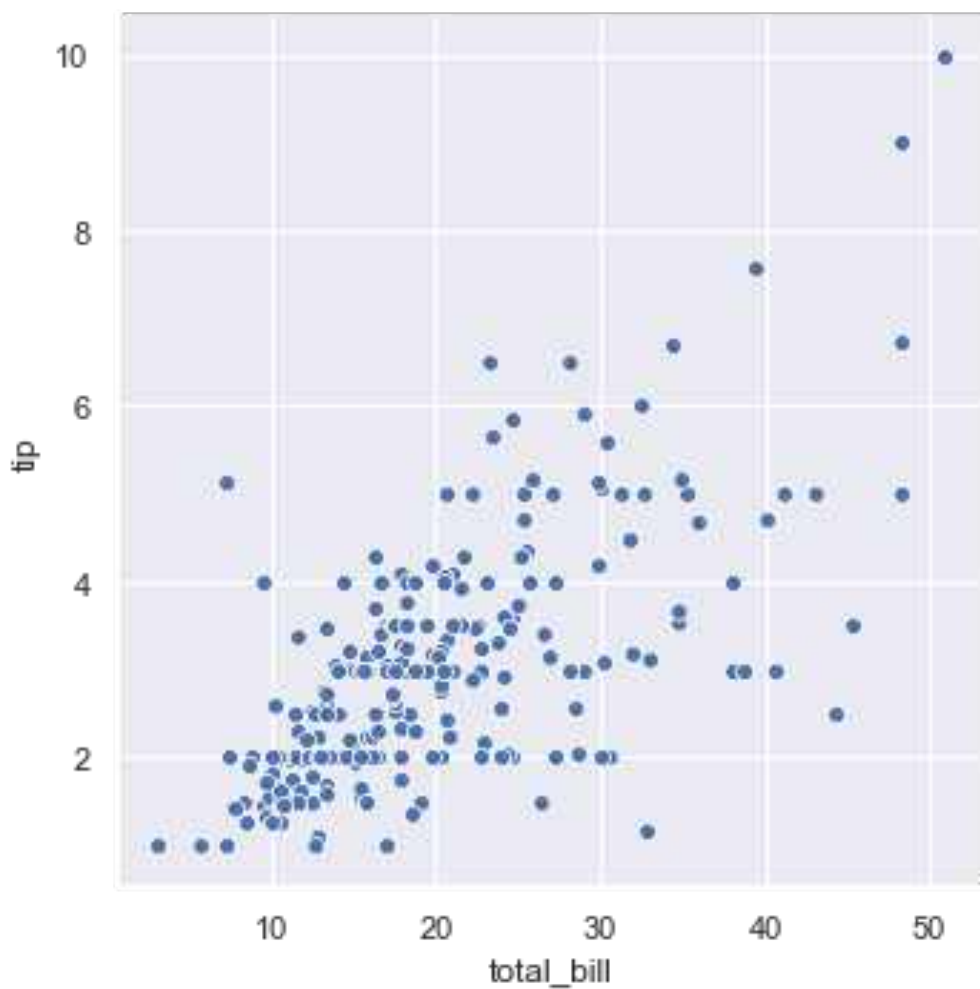
# Visualisation : Relation entre deux variables

❑ Points 2D (**scatter plot**) : tendance globale



# Visualisation : Relation entre deux variables

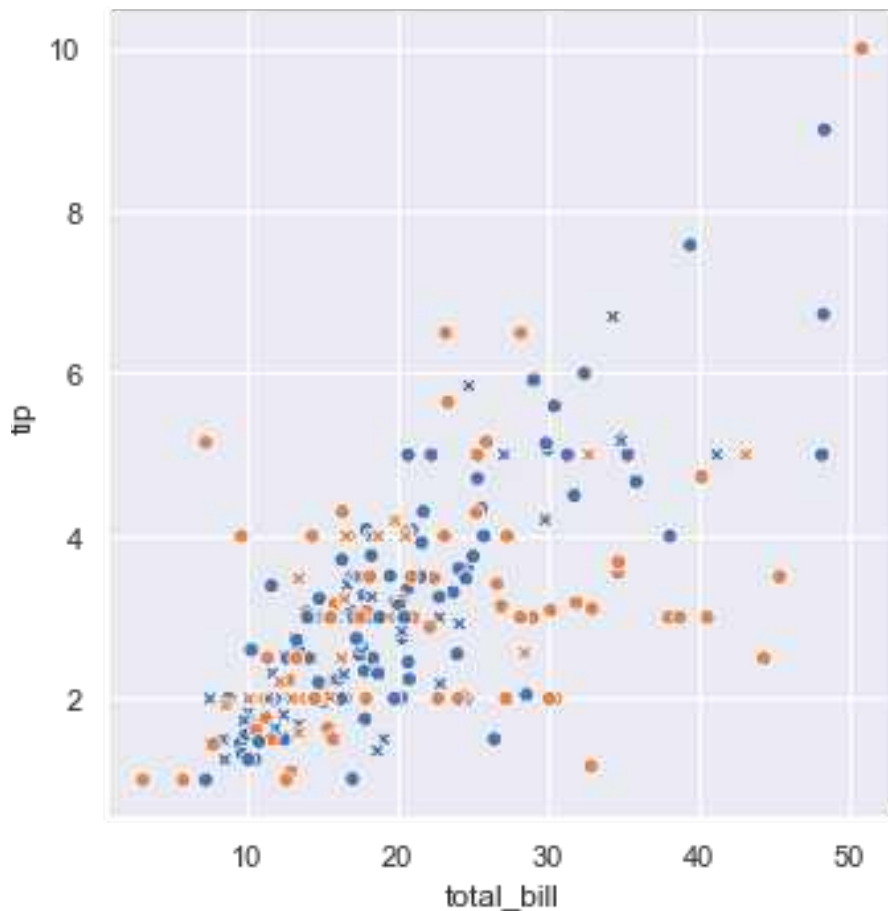
□ Points 2D (**scatter plot**) : tendance globale



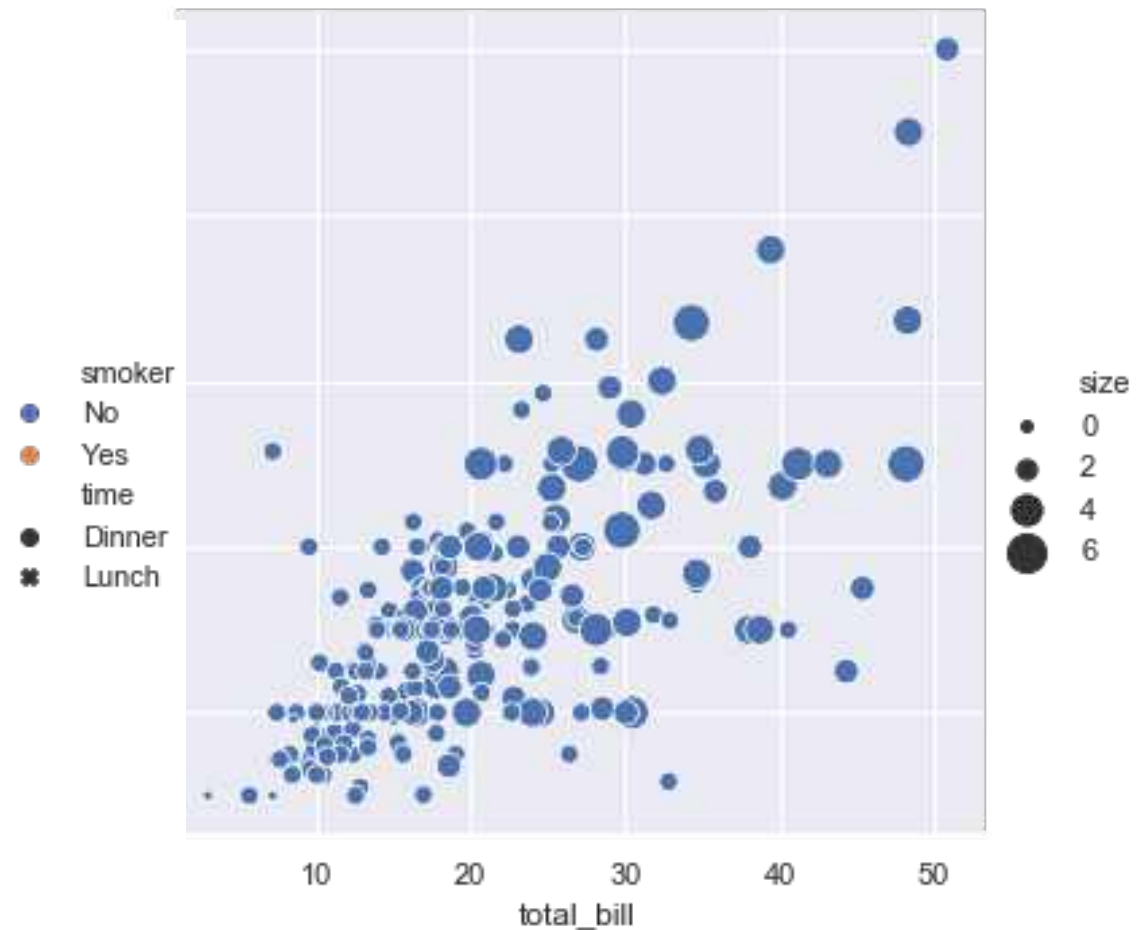
3<sup>e</sup> variable : couleur (hue)

# Visualisation : Relation entre deux variables

□ Points 2D (**scatter plot**) : tendance globale



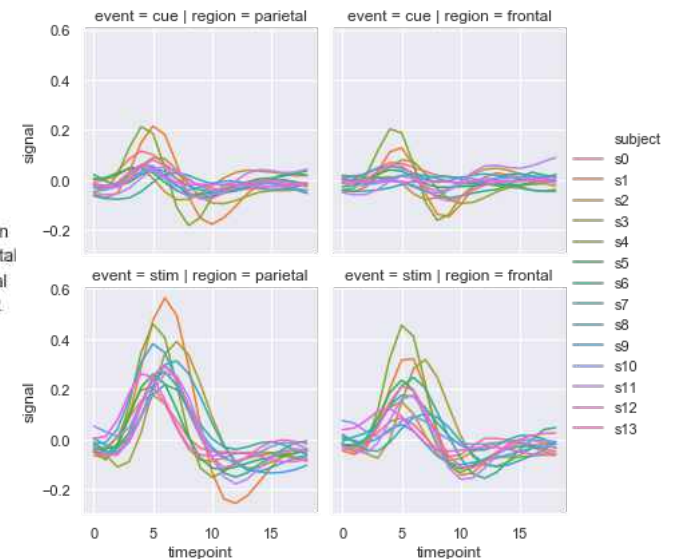
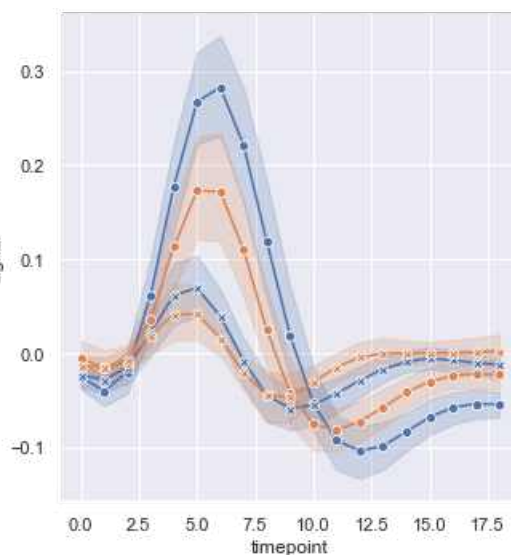
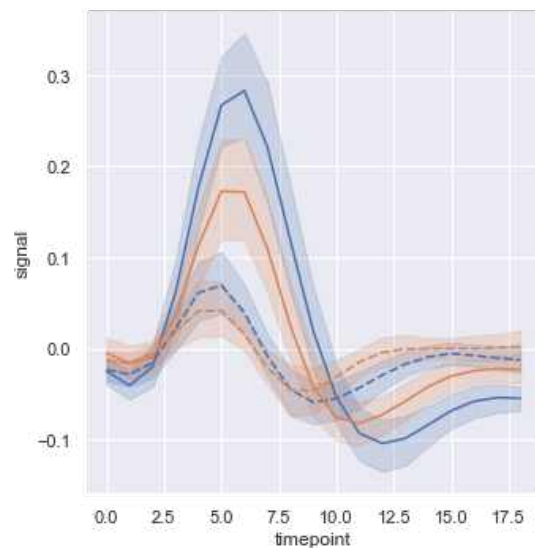
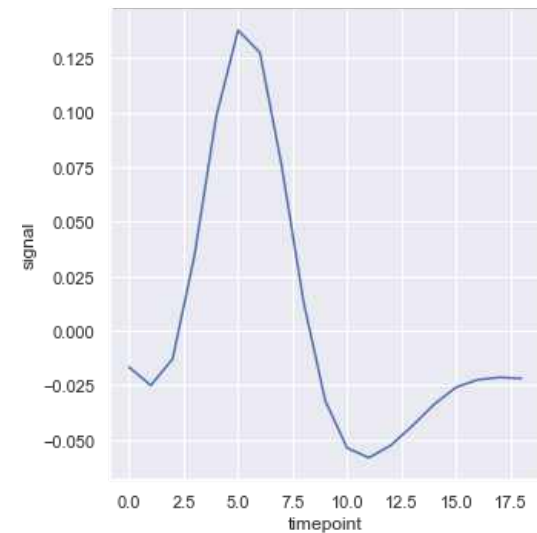
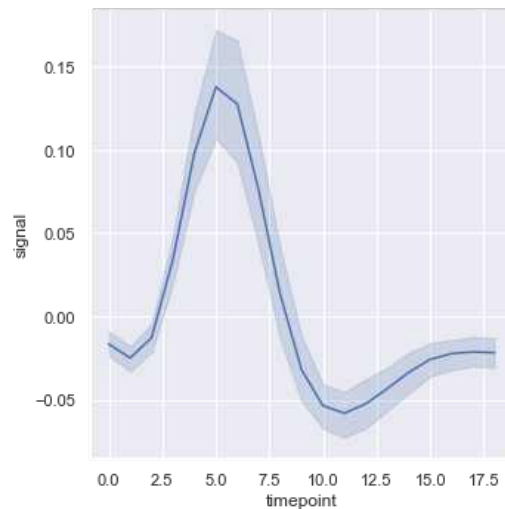
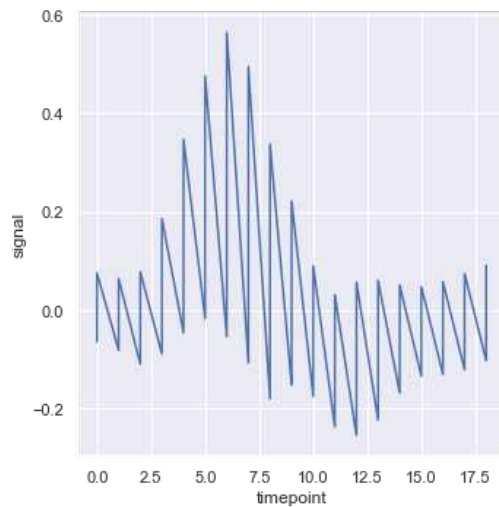
3<sup>e</sup> variable : couleur (hue)  
4<sup>e</sup> variable : style (style)



3<sup>e</sup> variable : taille (size)

# Visualisation : Relation entre deux variables

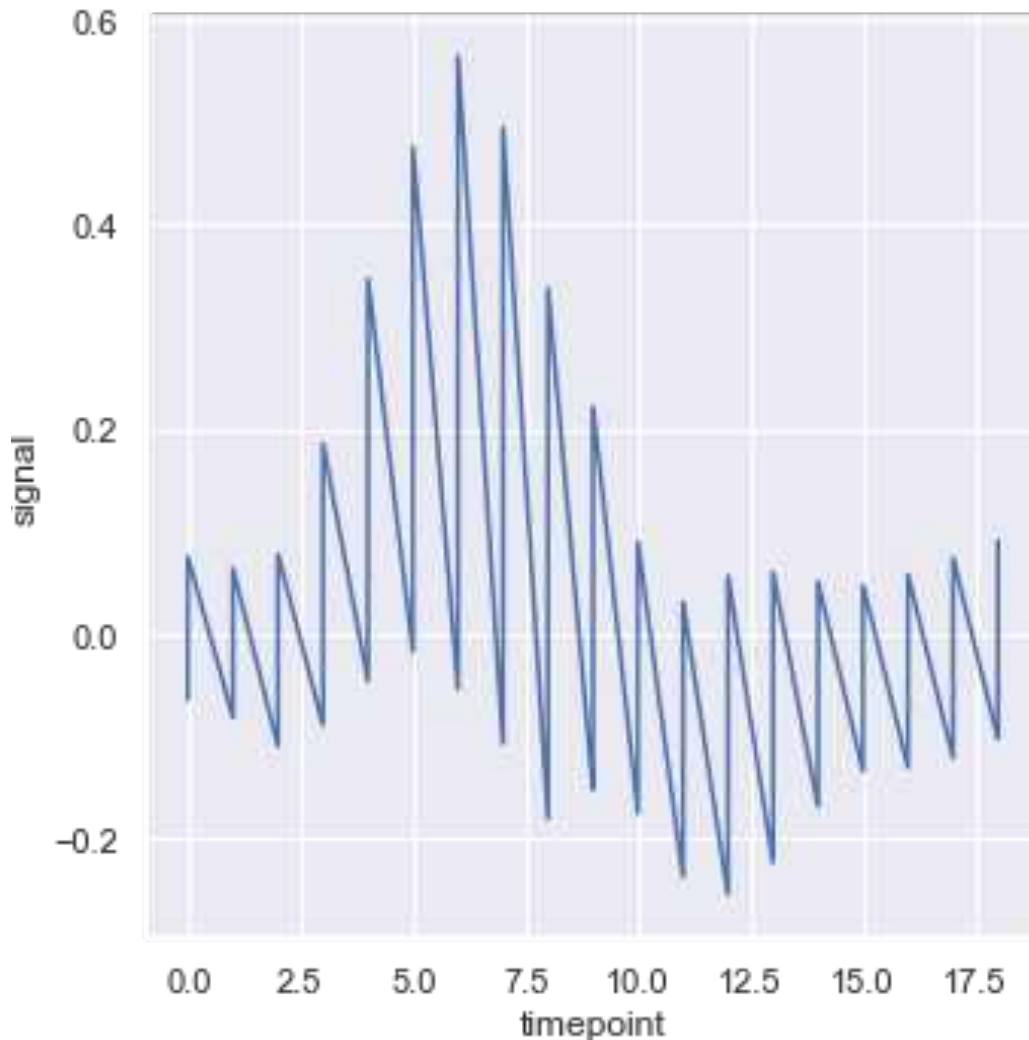
## □ Courbe de fonction (**line plot**) : tendance locale



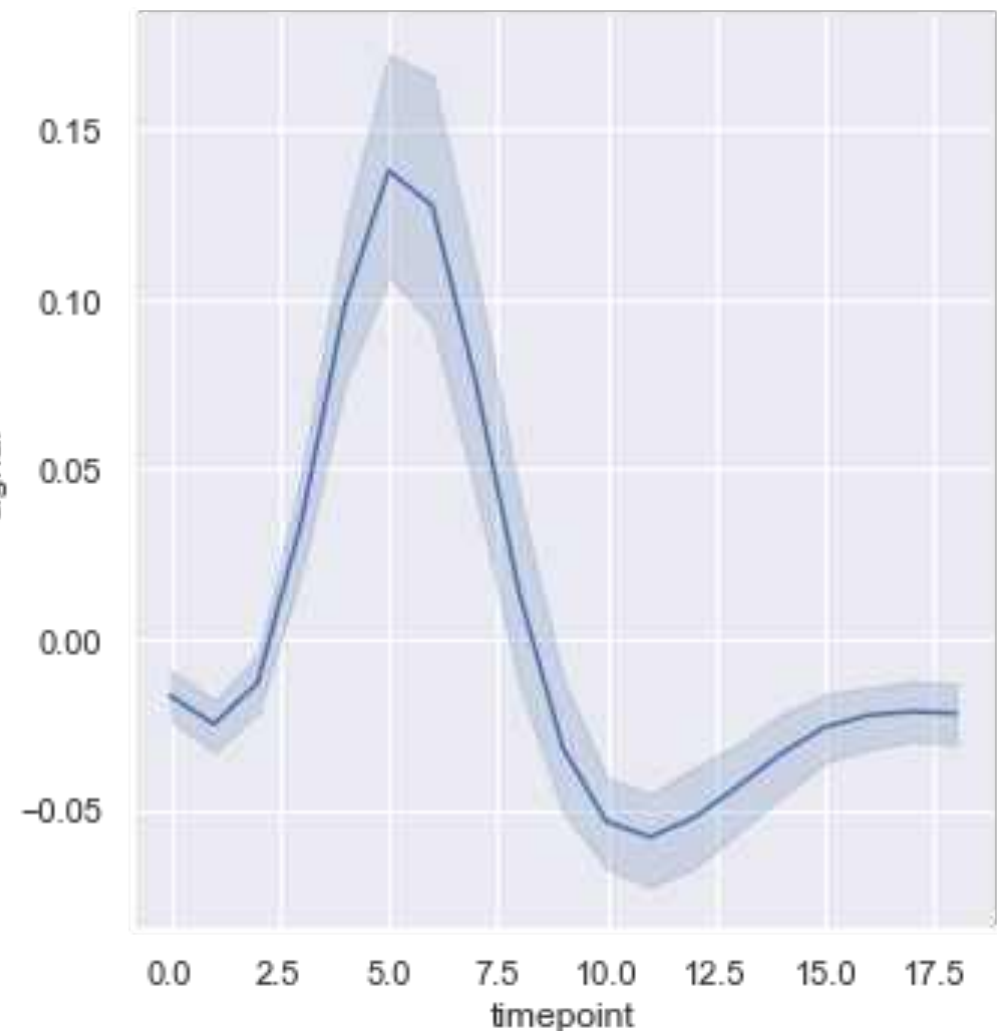
# Visualisation : Relation entre deux variables

- ❑ Courbe de fonction (**line plot**) : tendance locale

Observations sans  
estimation de moyennes

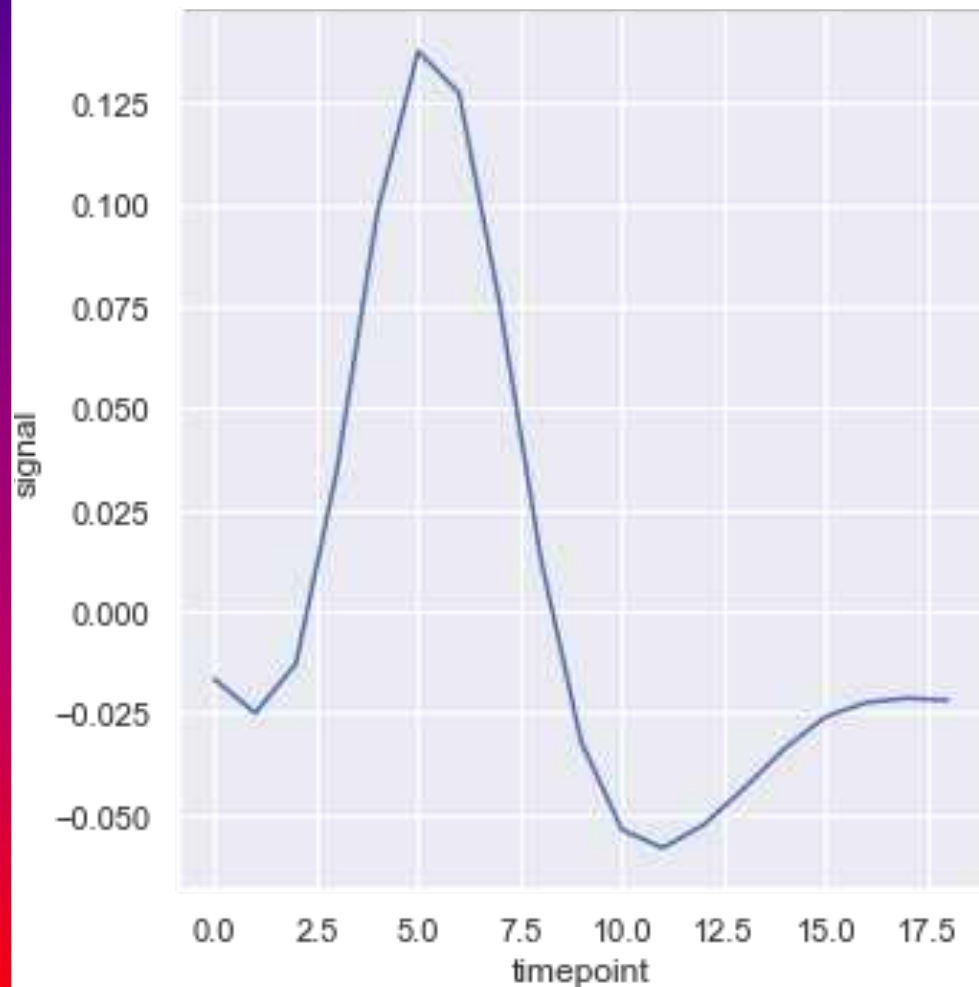


Estimation de moyennes  
et intervalle de confiance (ci)

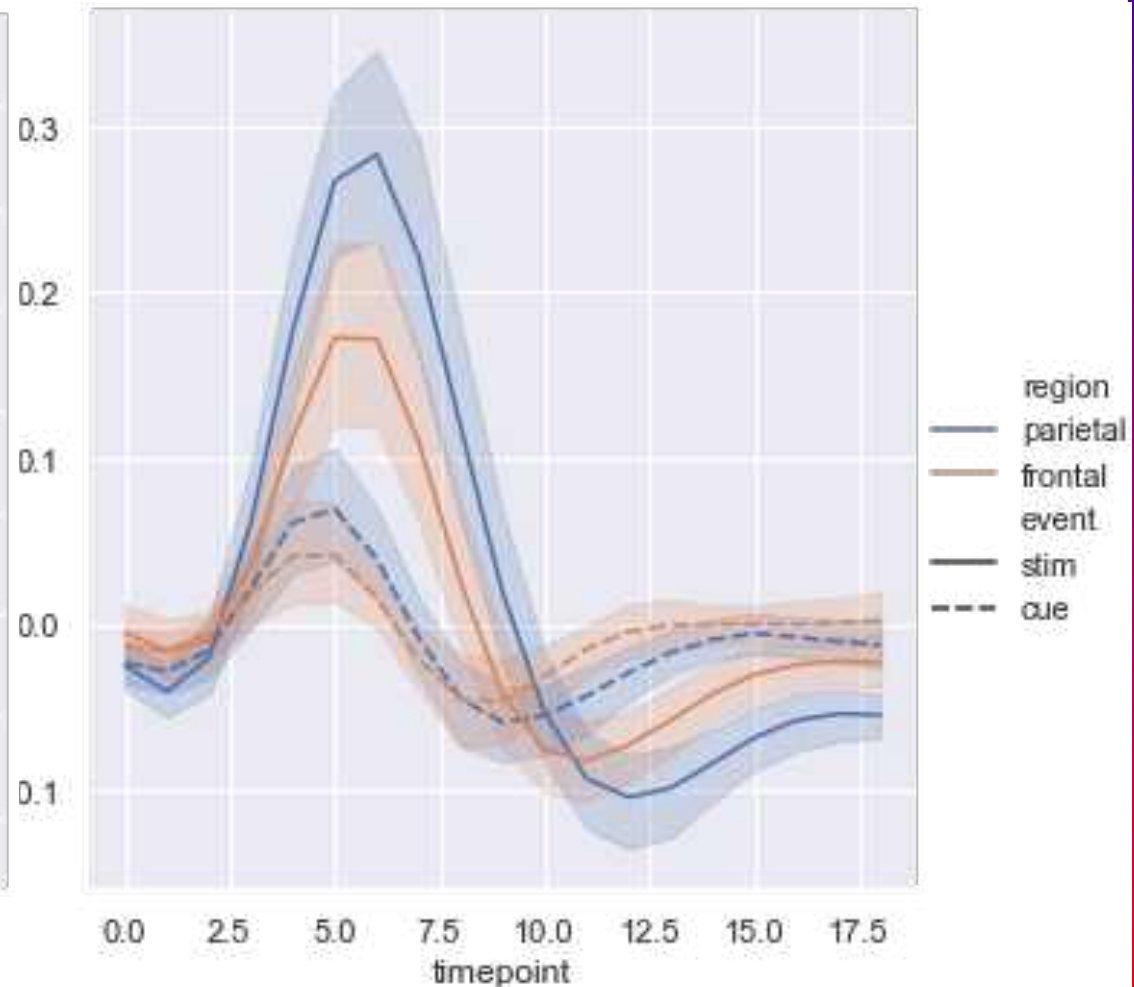


# Visualisation : Relation entre deux variables

❑ Courbe de fonction (**line plot**) : tendance locale



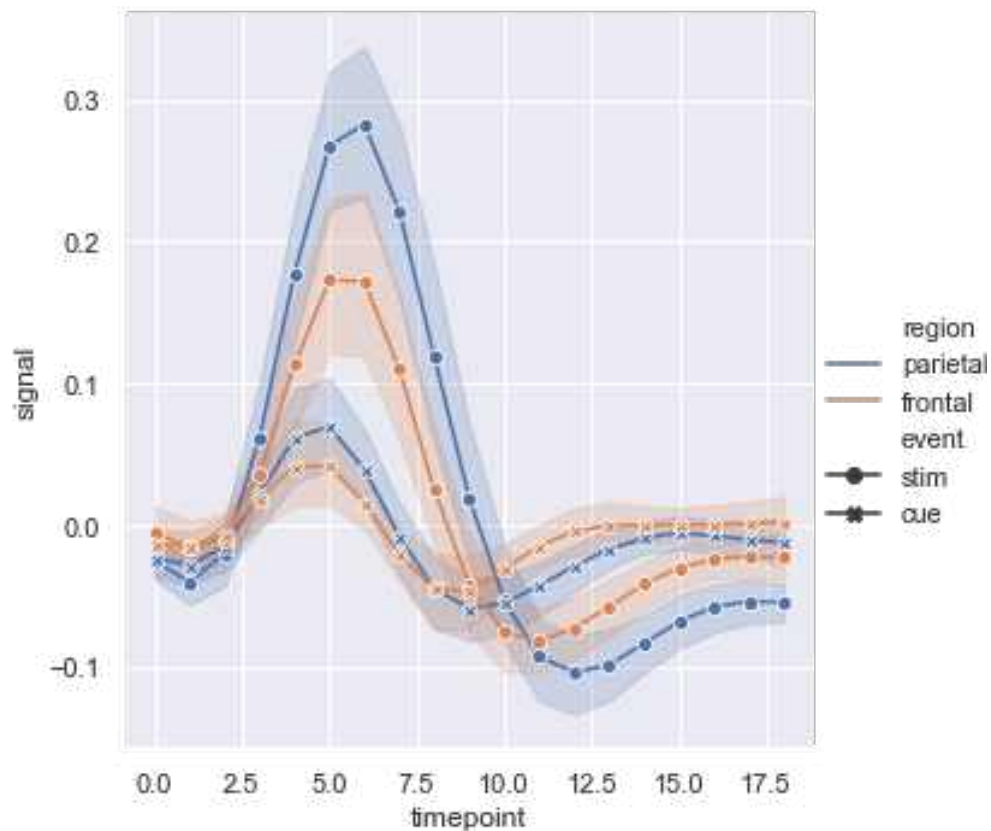
Estimation de moyennes sans  
intervalle de confiance



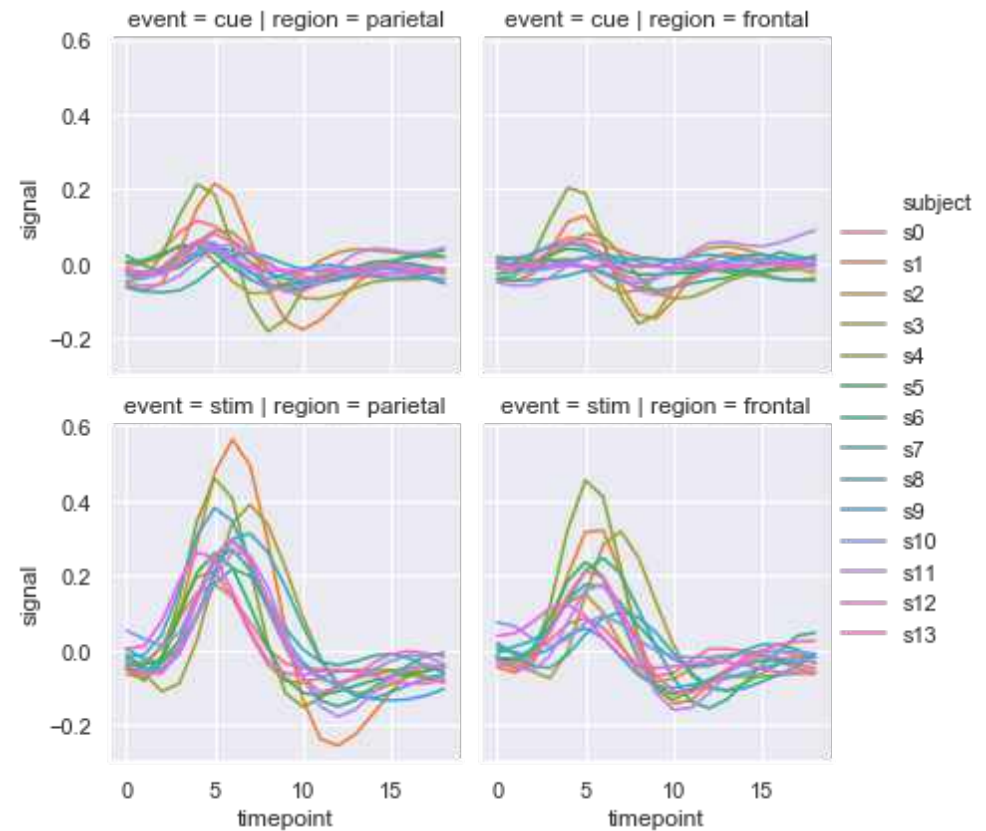
3<sup>e</sup> variable : couleur (hue)  
4<sup>e</sup> variable : style (style: dashes)

# Visualisation : Relation entre deux variables

❑ Courbe de fonction (**line plot**) : tendance locale



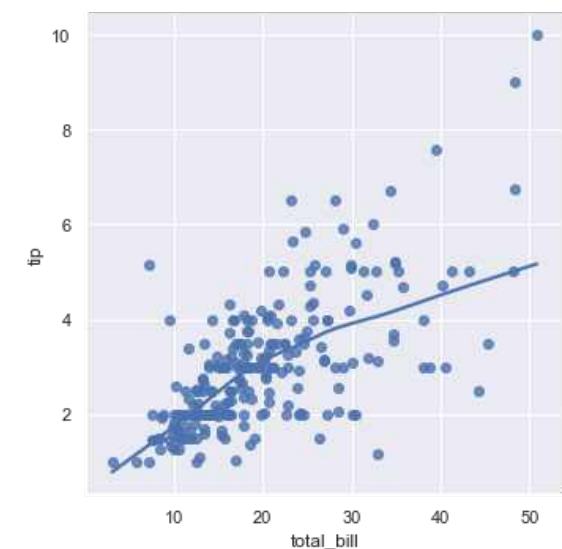
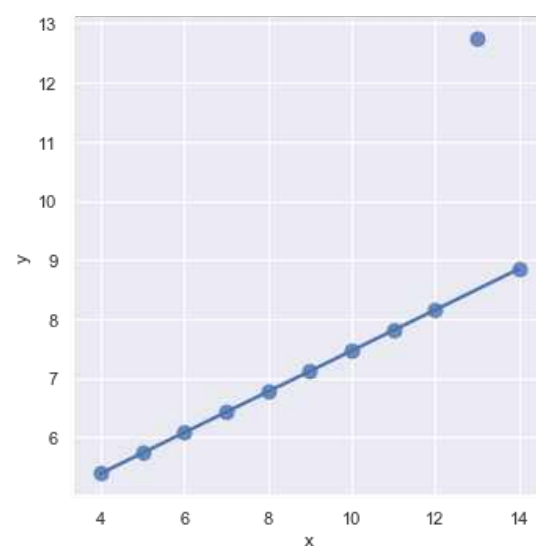
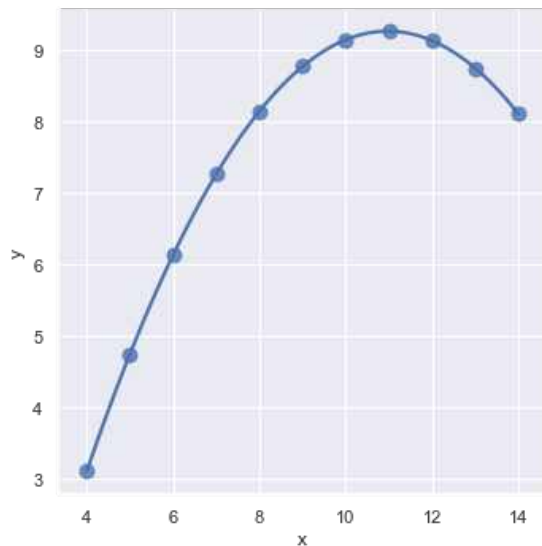
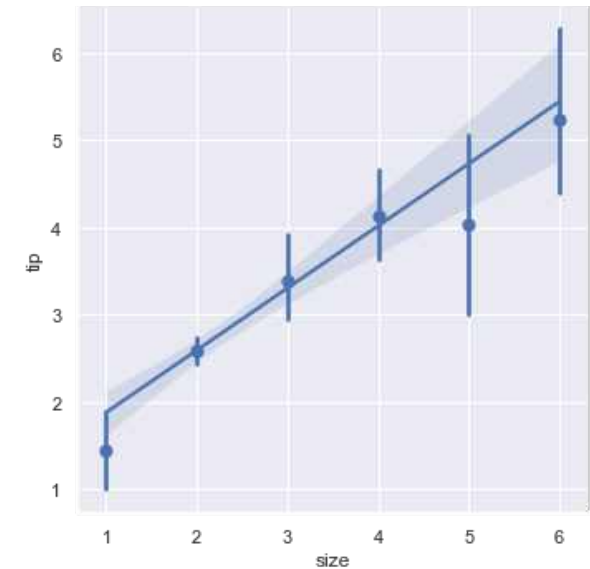
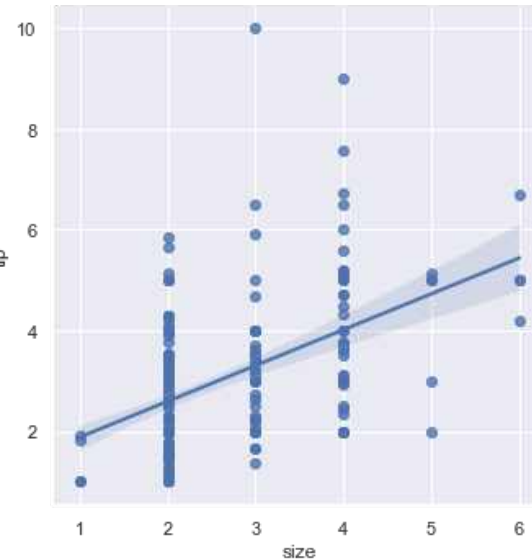
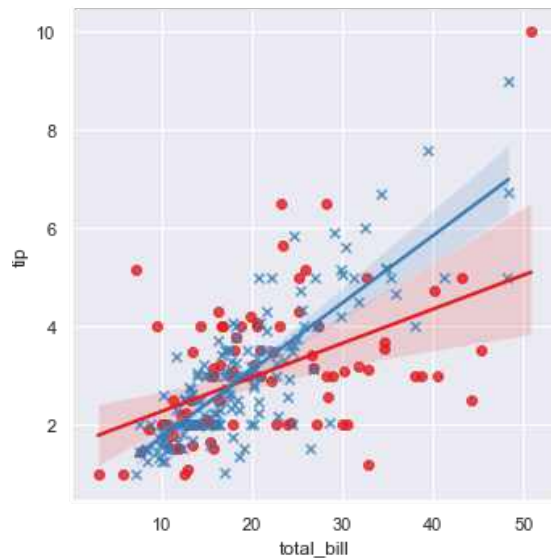
3<sup>e</sup> variable : couleur (hue)  
4<sup>e</sup> variable : style (style: markers)



Regroupement suivant 3<sup>e</sup>  
variable (row) et 4<sup>e</sup> variable (col)

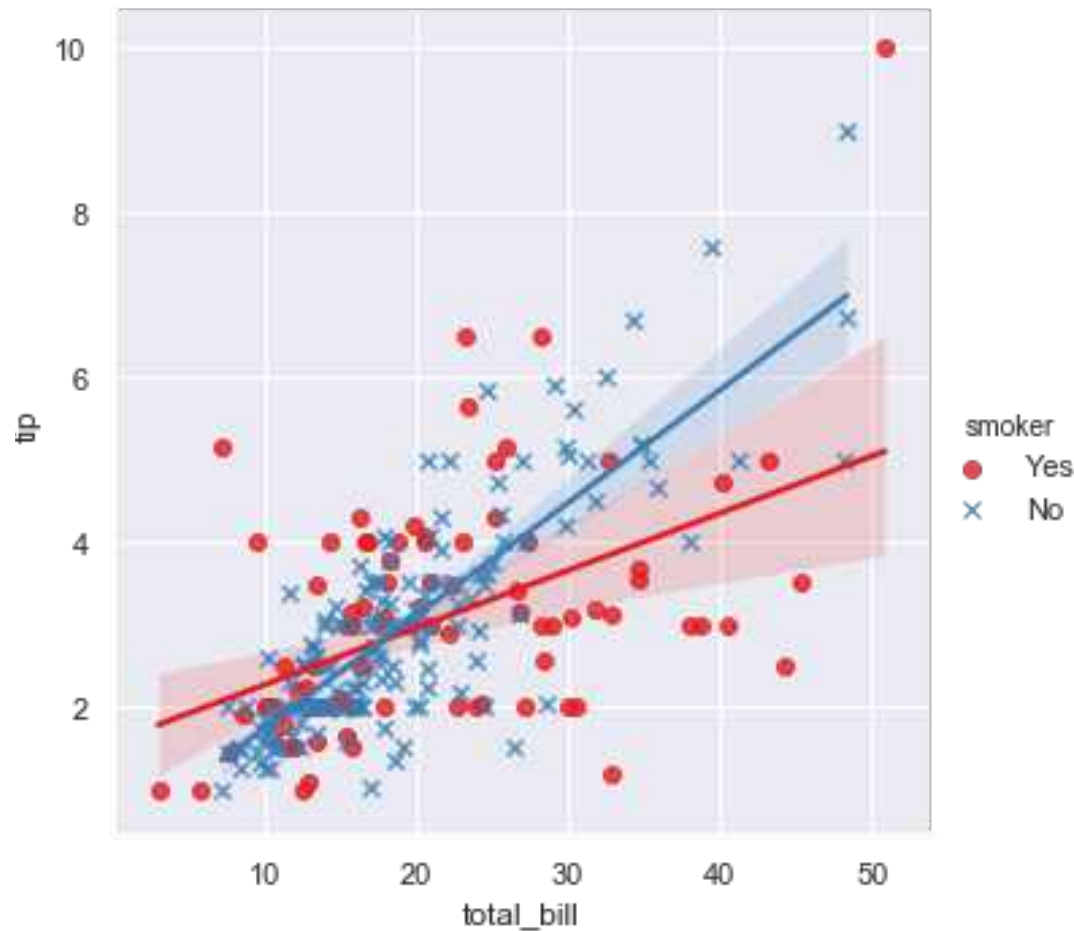
# Visualisation : Régression entre deux variables

❑ Régression (**reg plot**, **lm plot**) : estimation d'une relation simple

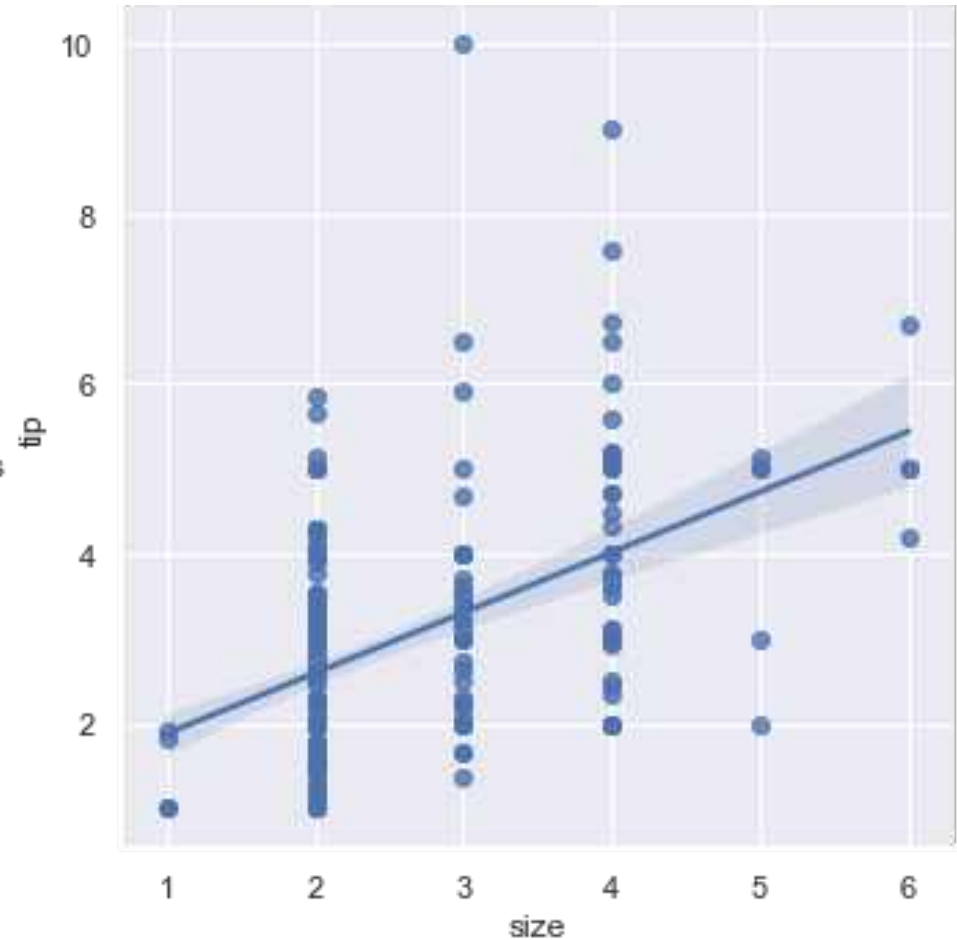


# Visualisation : Régression entre deux variables

- ❑ Régression (**reg plot**, **lm plot**) : estimation d'une relation simple



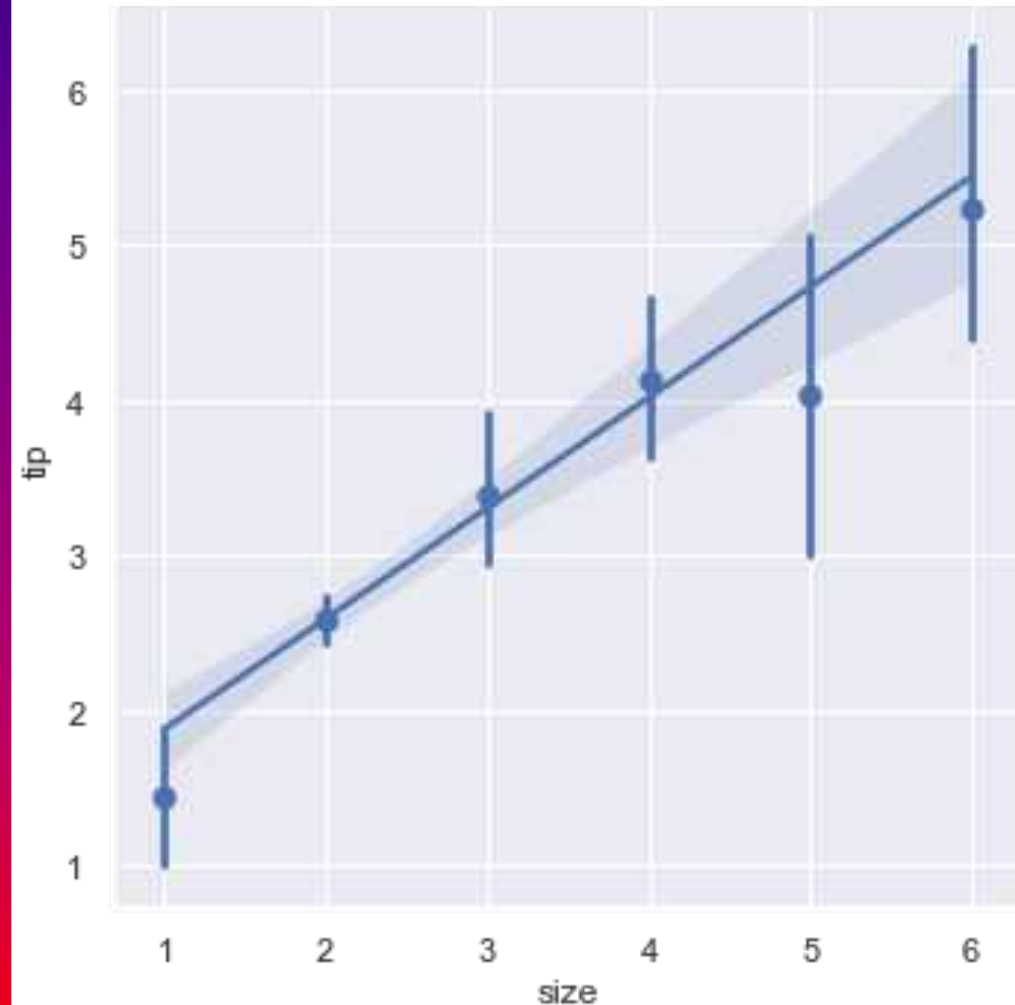
Avec estimation d'intervalle  
de confiance (ci)



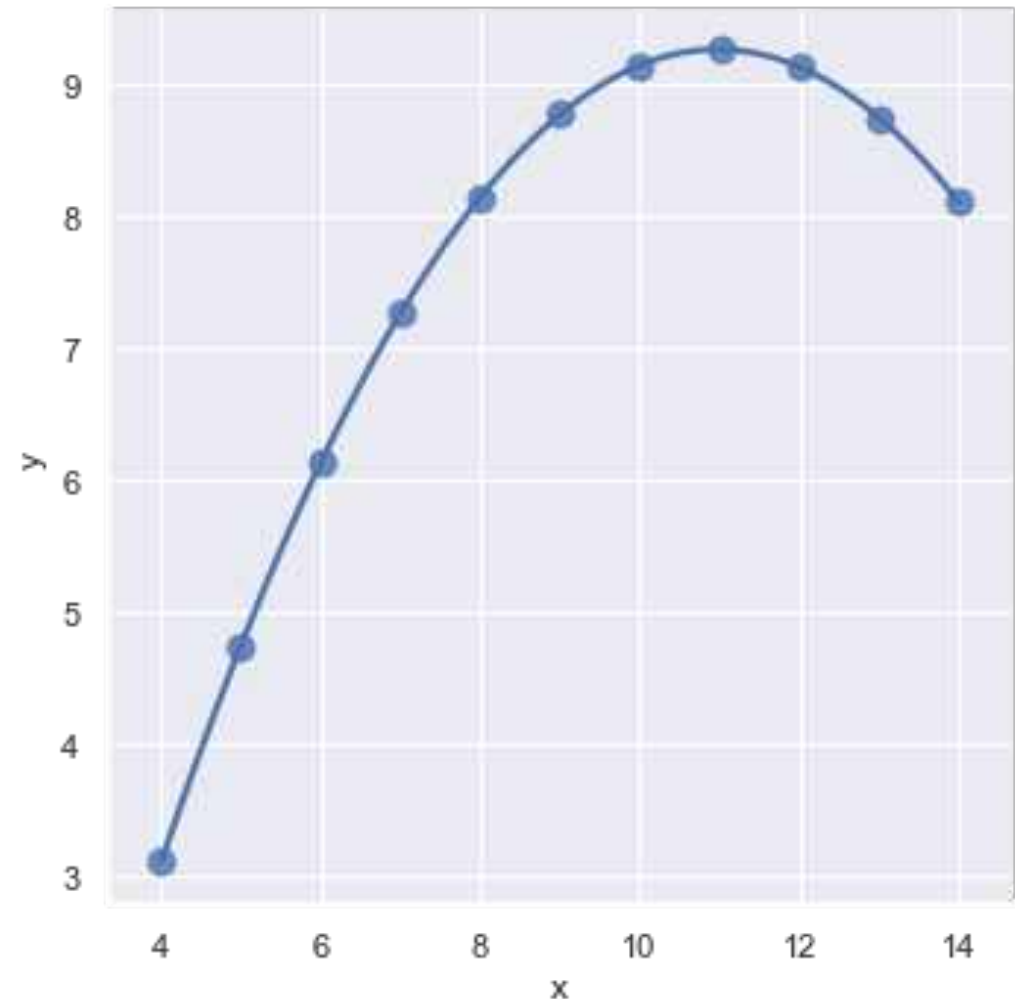
Une des variables  
est catégorielle

# Visualisation : Régression entre deux variables

❑ Régression (**reg plot**, **lm plot**) : estimation d'une relation simple



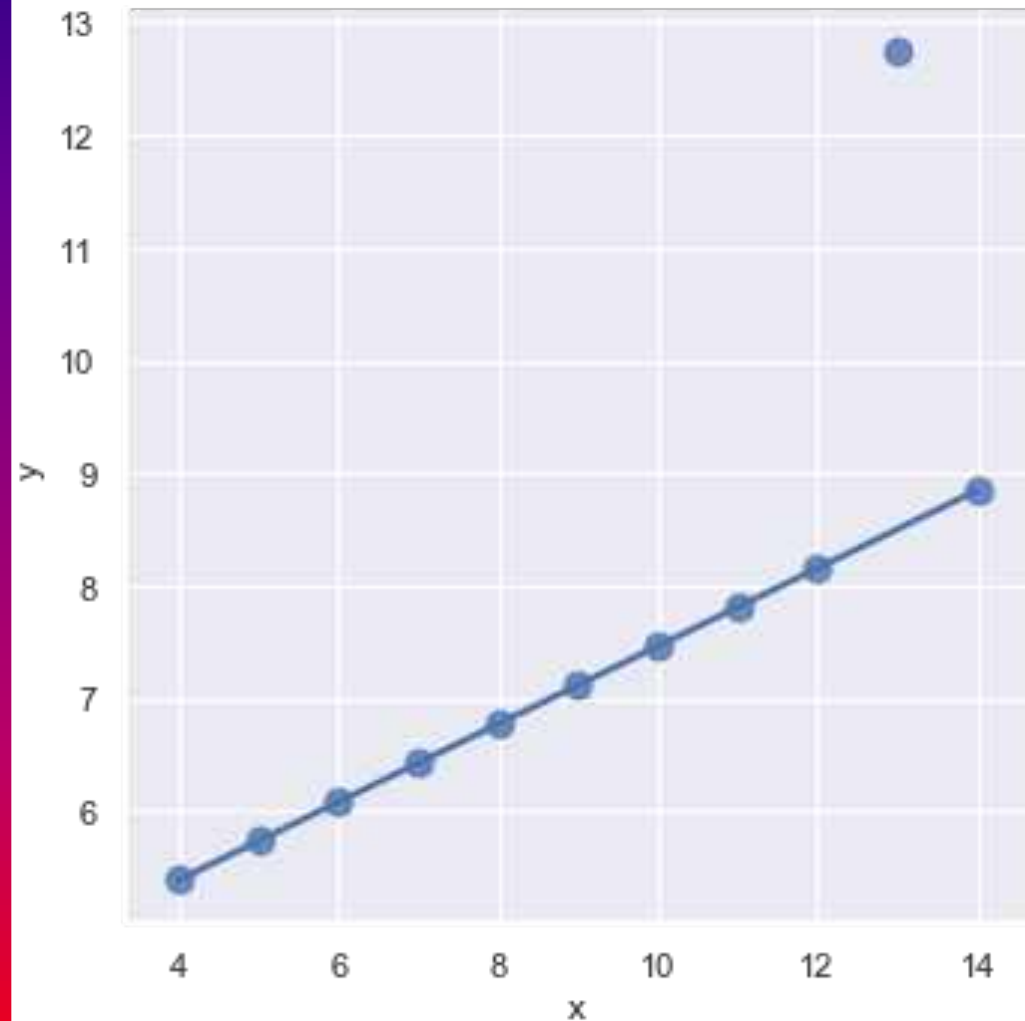
Relation polynomiale  
(order)



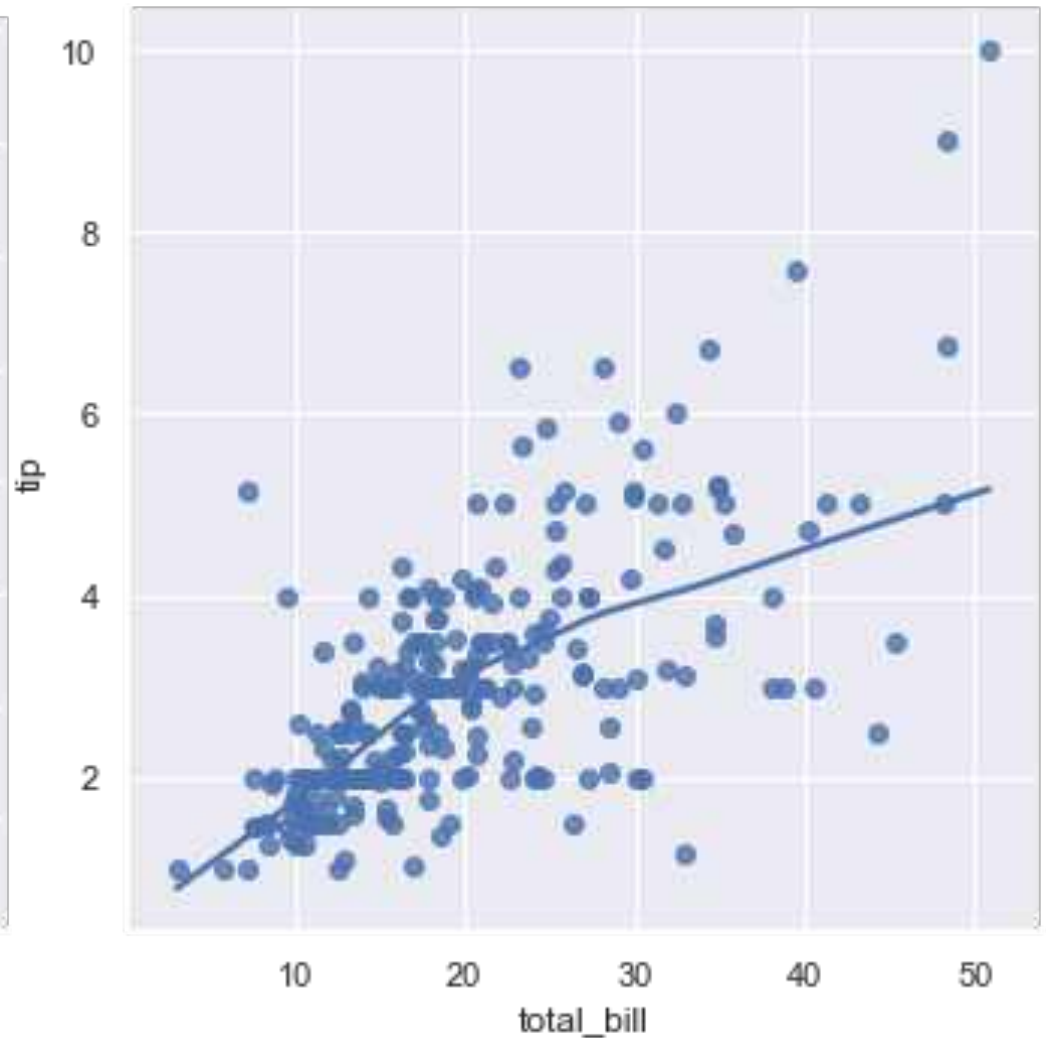
Avec estimation de moyenne et  
intervalle de confiance (ci)

# Visualisation : Régression entre deux variables

❑ Régression (**reg plot**, **lm plot**) : estimation d'une relation simple



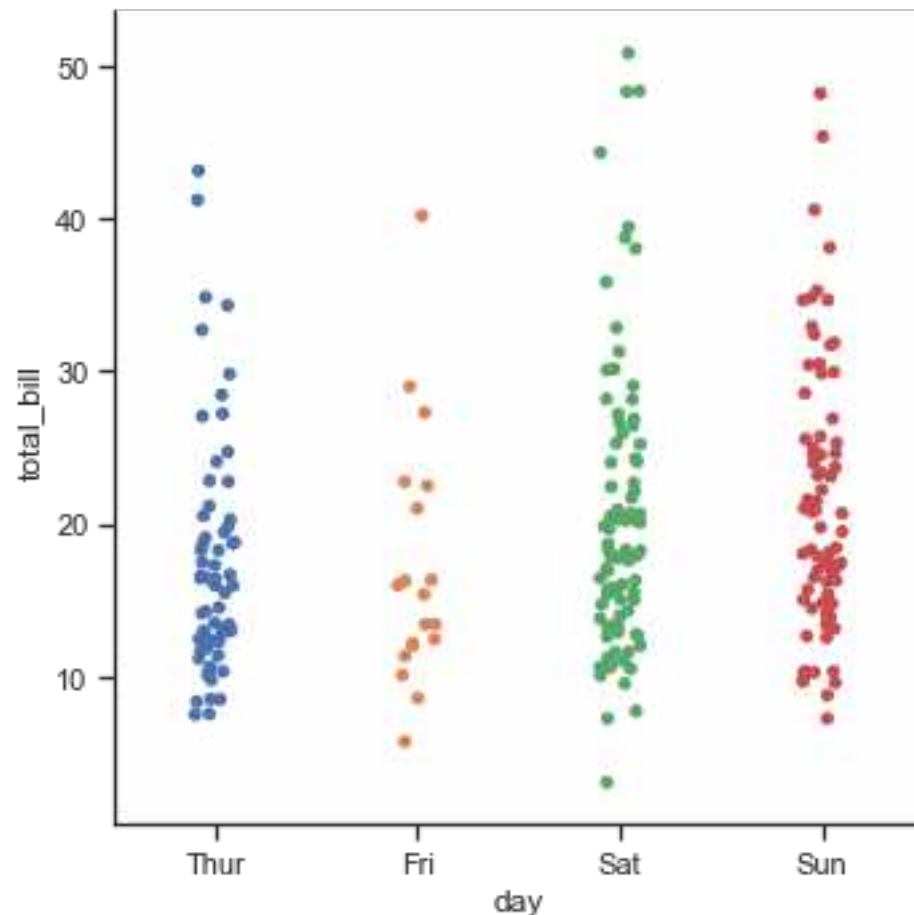
Estimation robuste aux données aberrantes (robust)



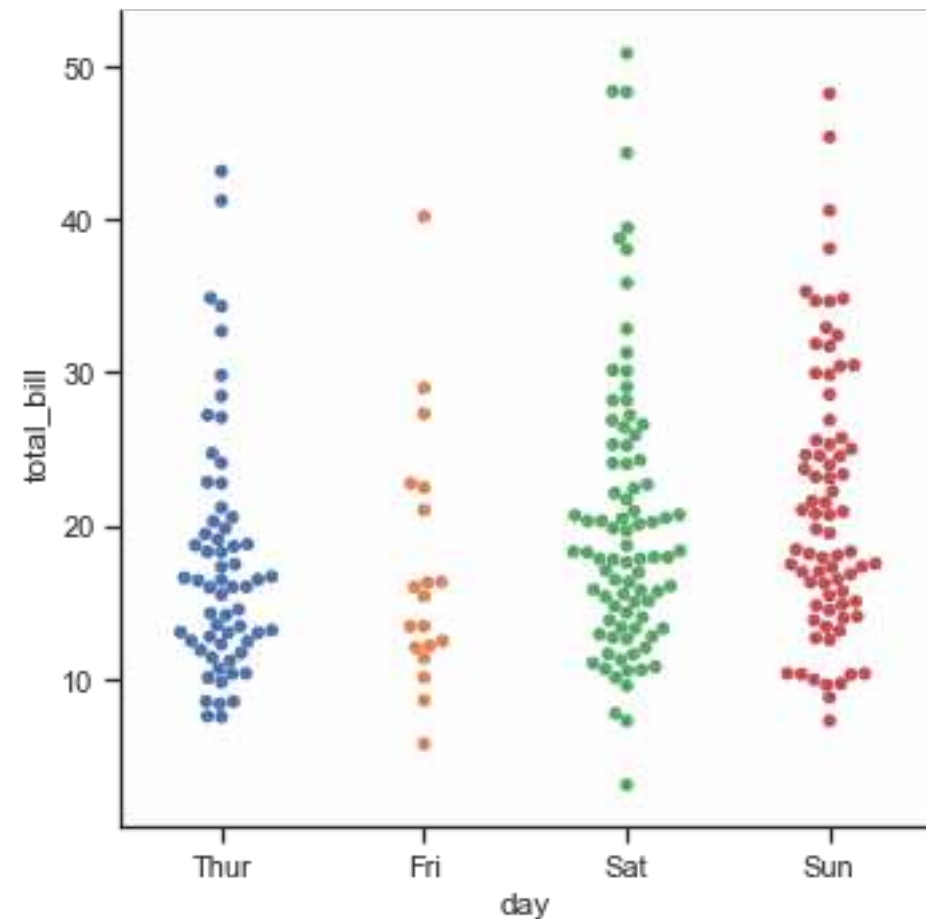
Lissage de la courbe de régression (lowess)

# Visualisation : Relation entre deux variables dont une catégorielle

❑ Point 2D : comparer les tendances globales dans les différentes catégories



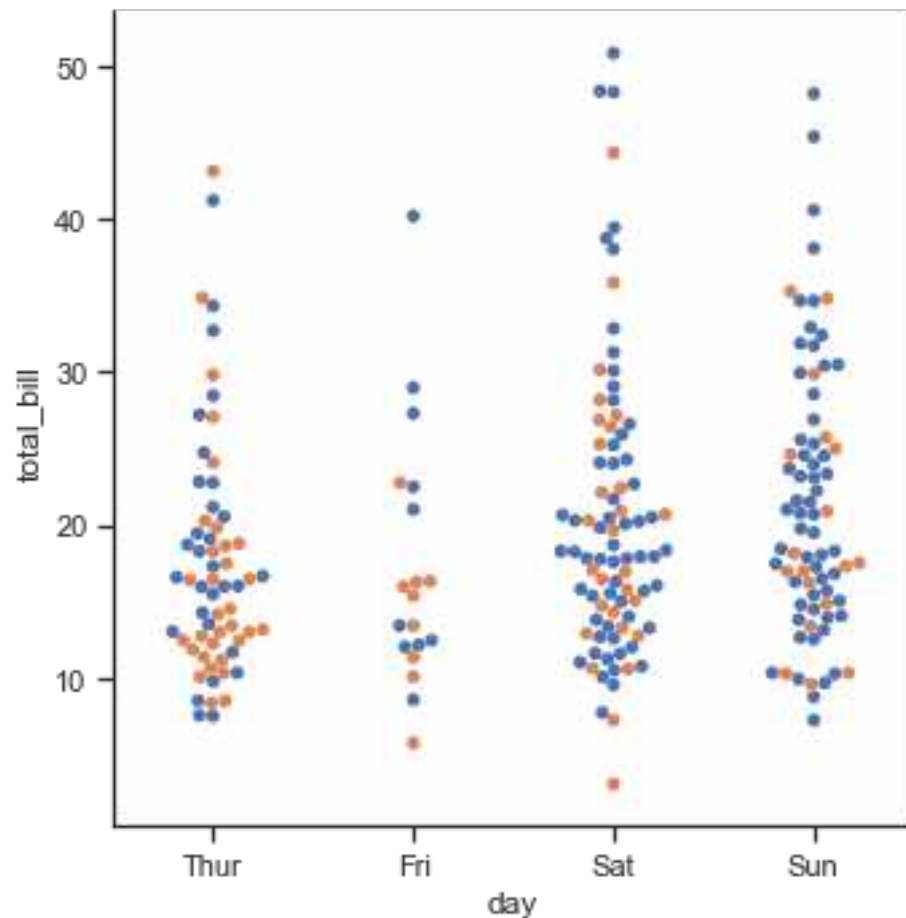
En bande (**strip plot**)



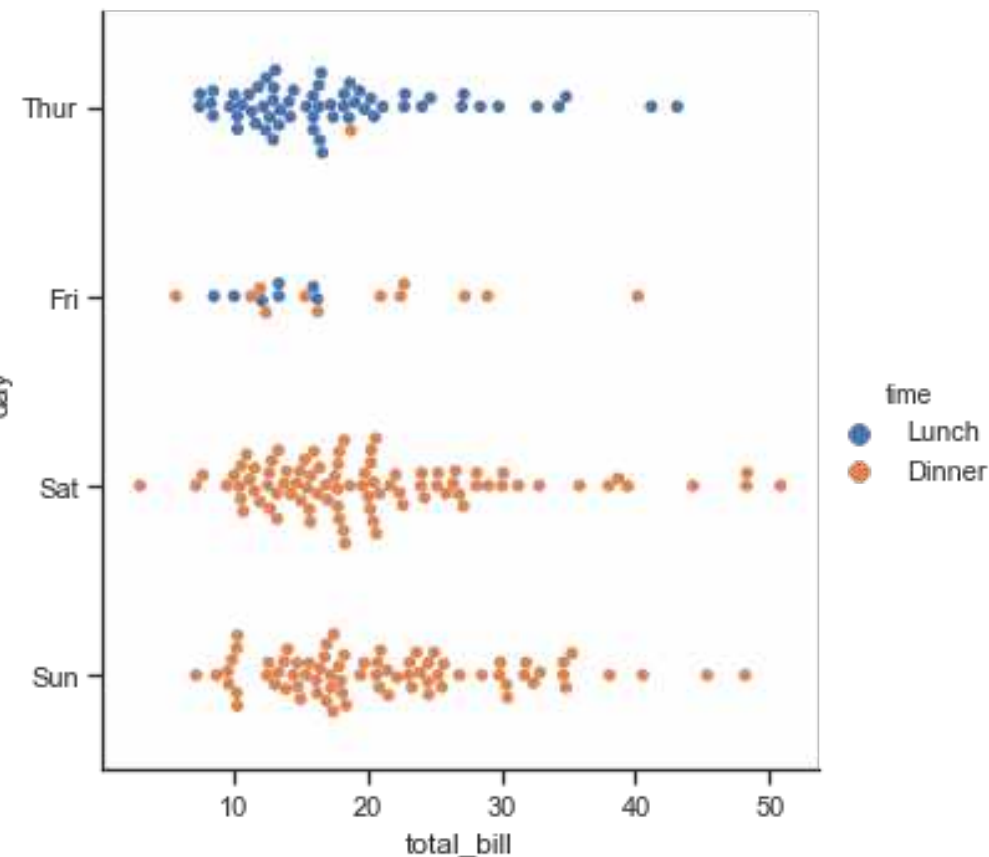
En essaim (**swarm plot**)

# Visualisation : Relation entre deux variables dont une catégorielle

❑ Point 2D : comparer les tendances globales dans les différentes catégories



3<sup>e</sup> variable : couleur (hue)

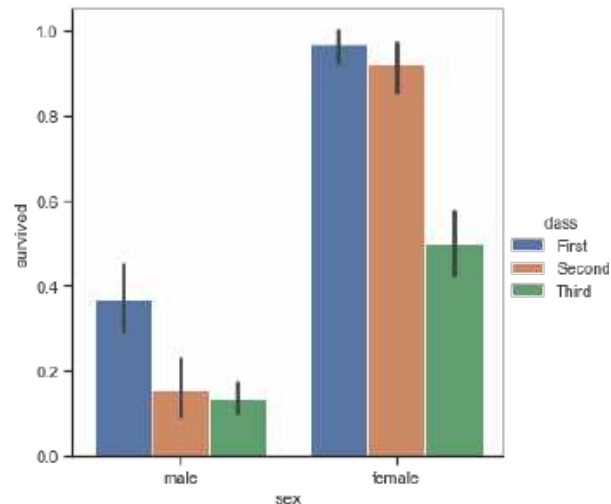


# Visualisation : Relation entre deux variables dont une catégorielle

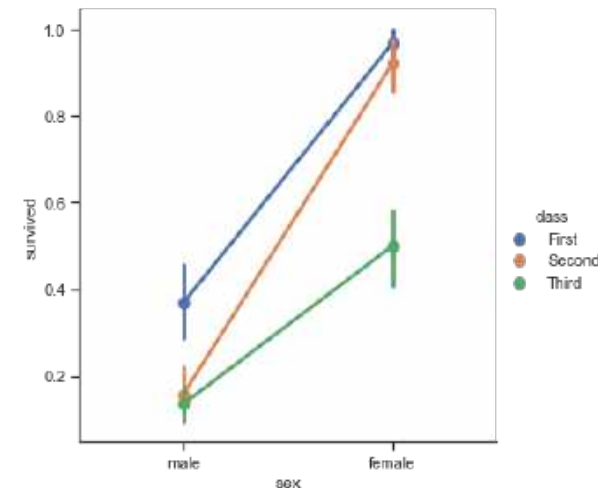
- ❑ Diagramme à barres et dérivées : comparer les tendances centrales dans les différentes catégories

Estimation de moyennes  
En barre (**bar plot**)

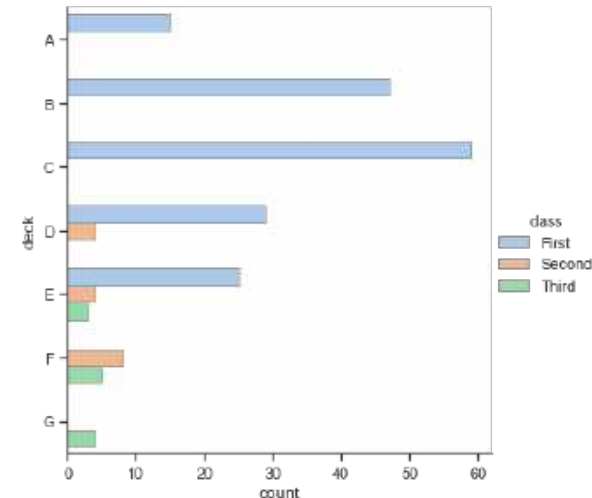
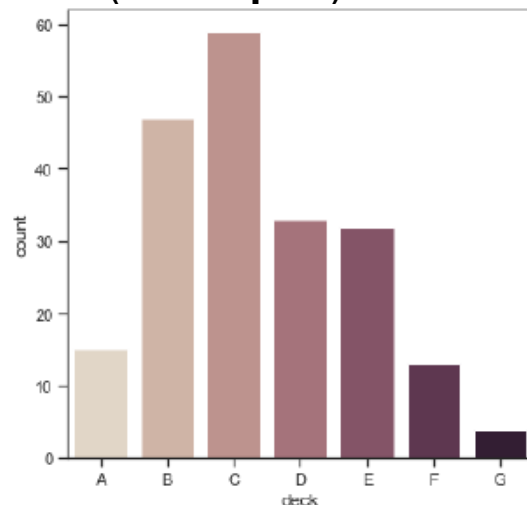
3<sup>e</sup> variable :  
couleur (hue)



Moyennes en  
Point (**point plot**)



Compte pour une seule variable  
(**count plot**)



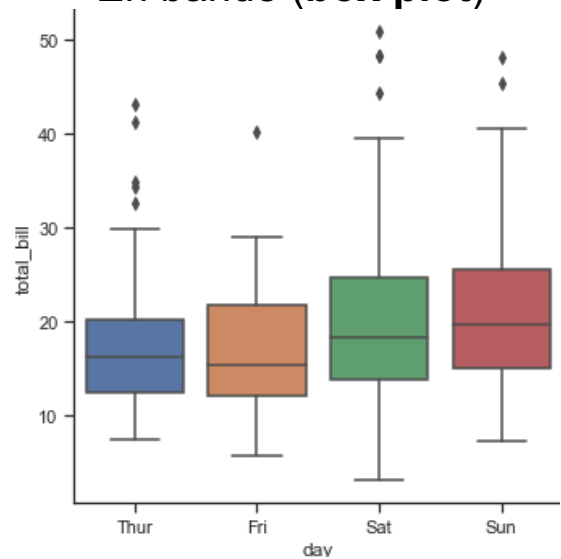
2<sup>e</sup> variable :  
couleur (hue)

# Visualisation : Relation entre deux variables dont une catégorielle

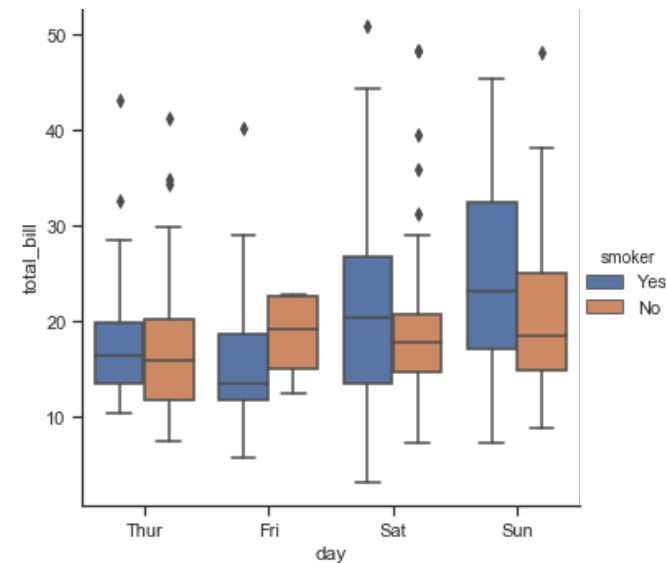
- ❑ Boîtes à moustaches et dérivées : comparer les distributions dans les différentes catégories

Observations : Quartiles 1,2,3.

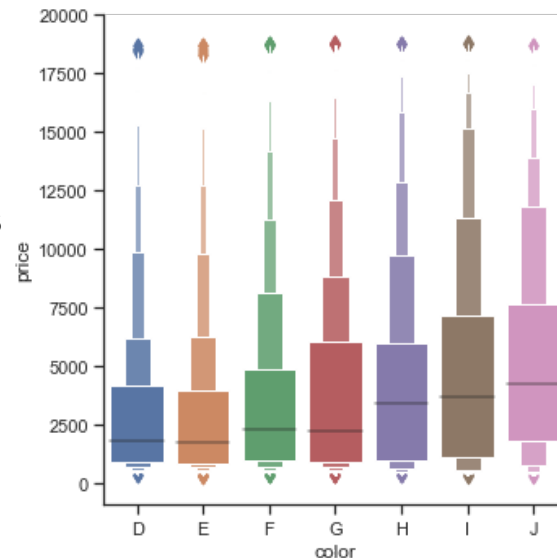
En bande (**box plot**)



3<sup>e</sup> variable : couleur (hue)



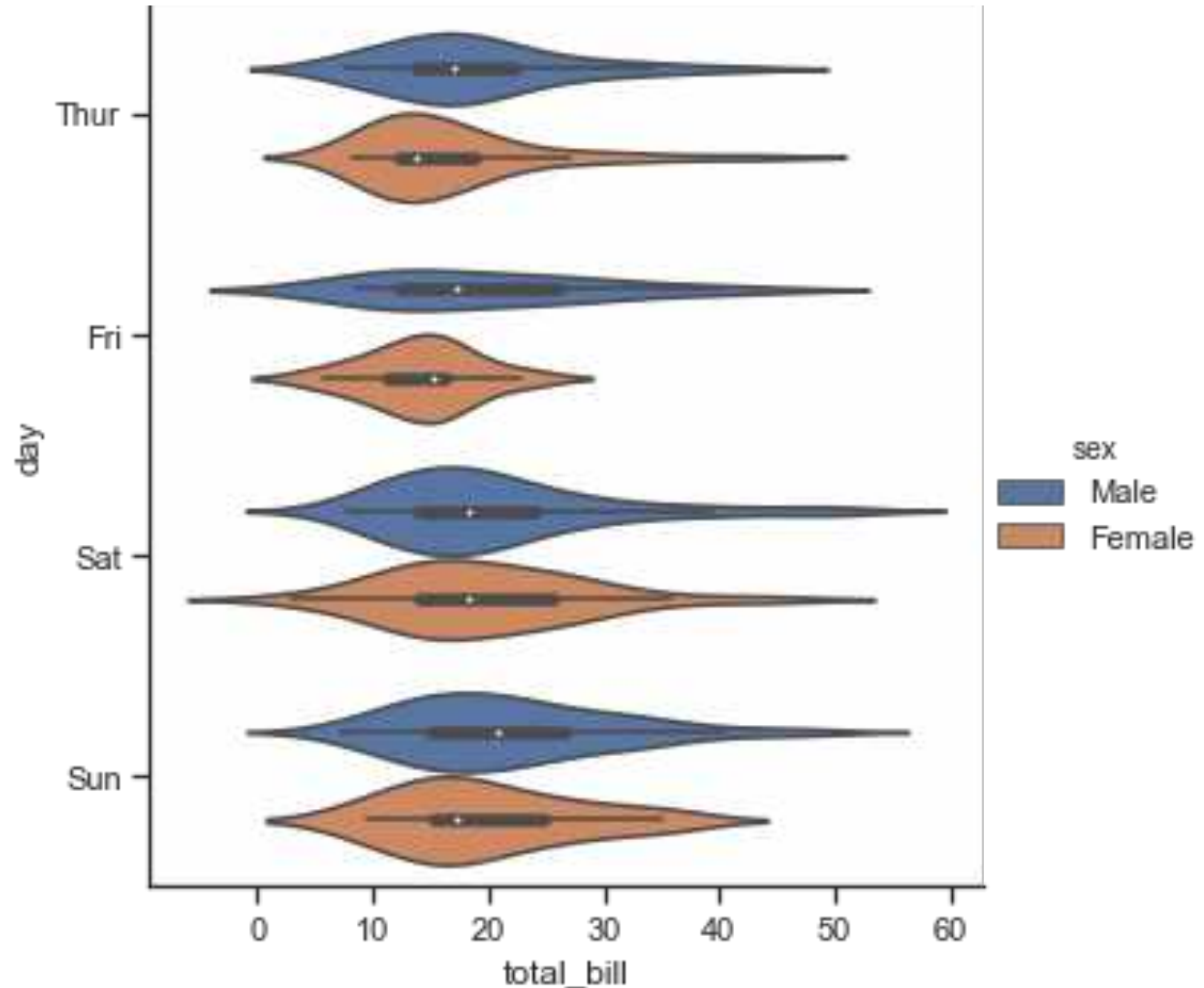
Quartiles  
supplémentaires  
(**boxen plot**)



# Visualisation : Relation entre deux variables dont une catégorielle

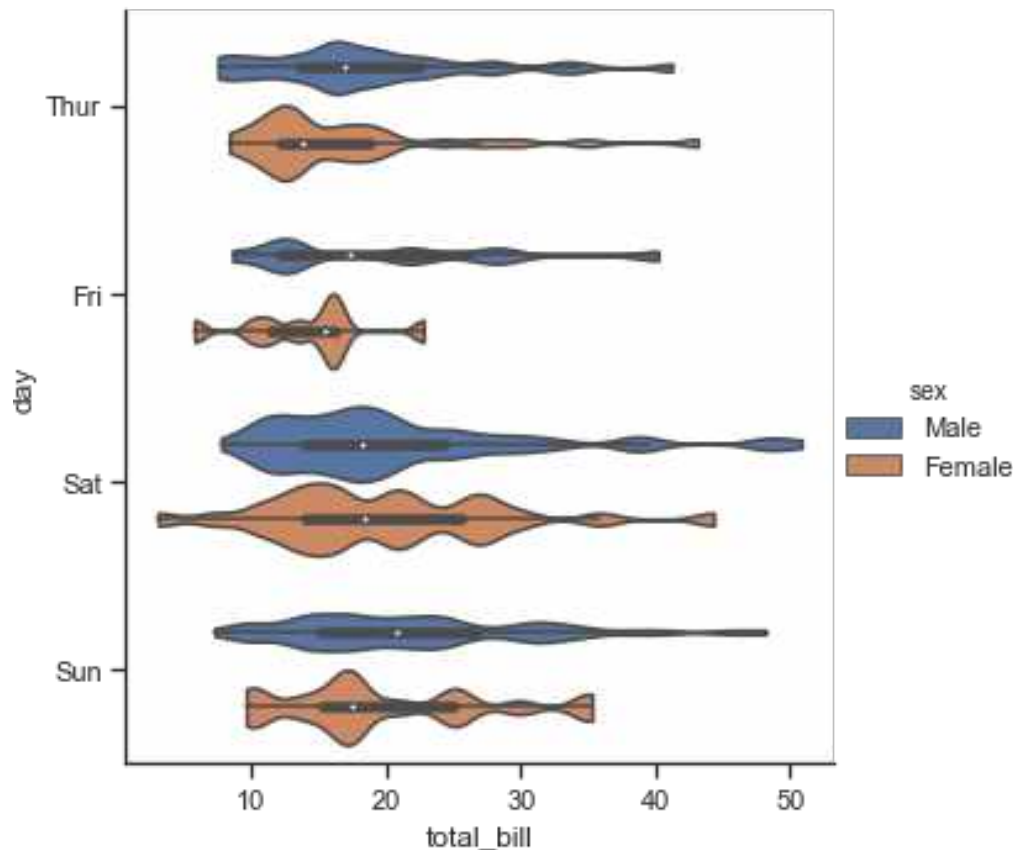
- ❑ Boîtes à moustaches et dérivées : comparer les distributions dans les différentes catégories

Estimation de distributions  
En violon (**violin plot**)

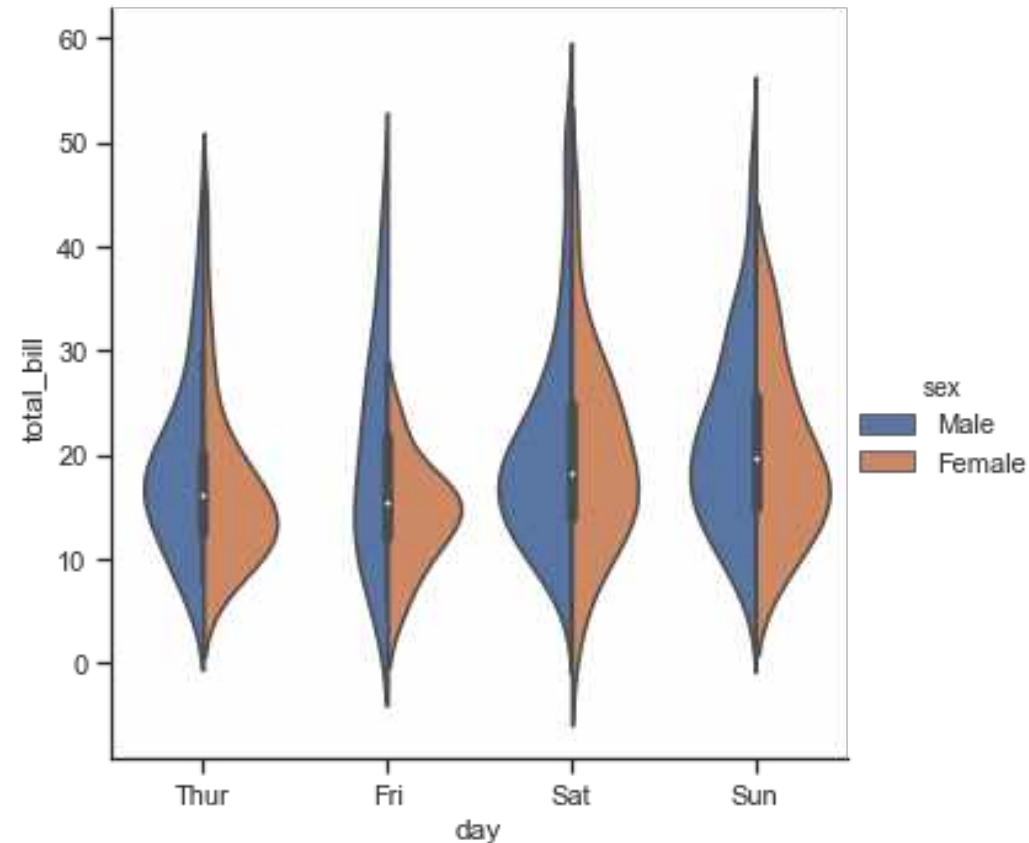


# Visualisation : Relation entre deux variables dont une catégorielle

- ❑ Boîtes à moustaches et dérivées : comparer les distributions dans les différentes catégories



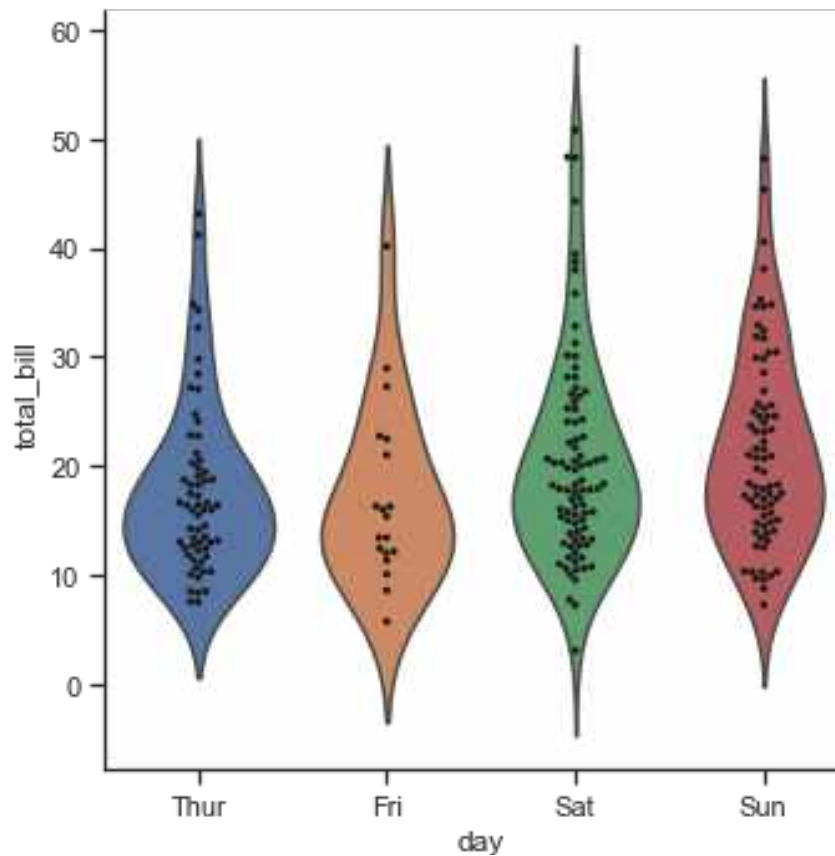
Large de bande (bandwidth bw)  
3<sup>e</sup> variable : couleur (hue)



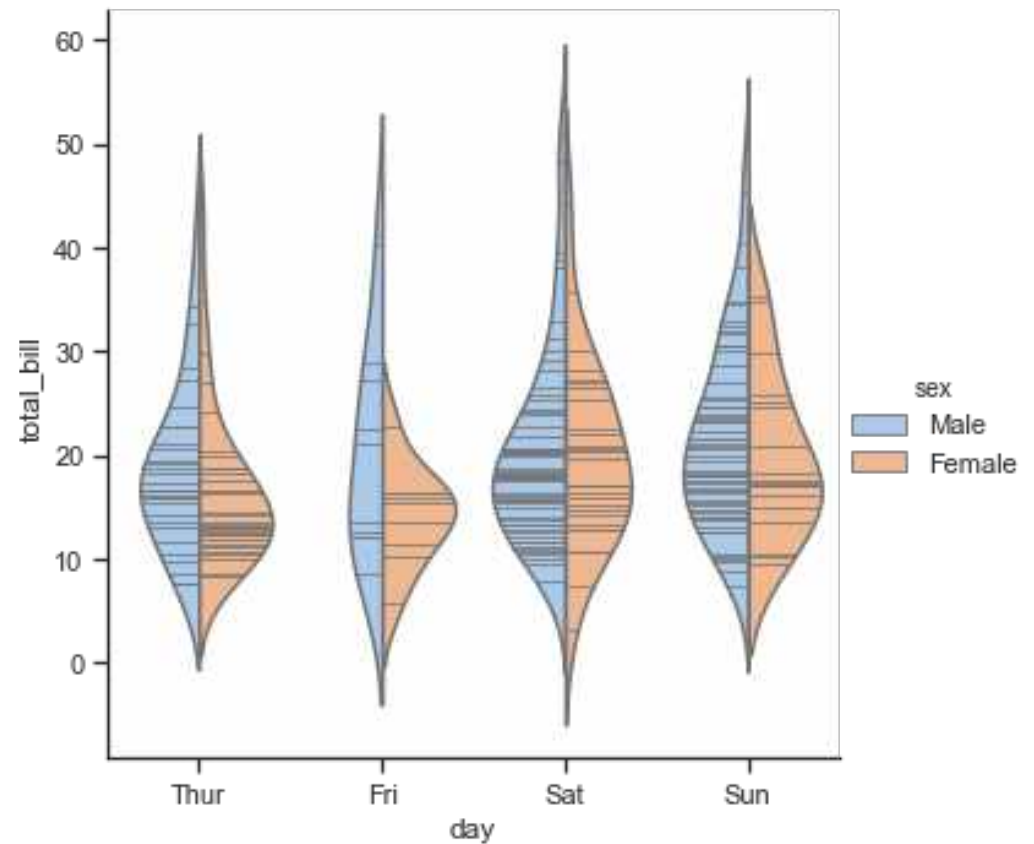
3<sup>e</sup> variable : couleur (hue, split)

# Visualisation : Relation entre deux variables dont une catégorielle

- ❑ Boîtes à moustaches et dérivées : comparer les distributions dans les différentes catégories



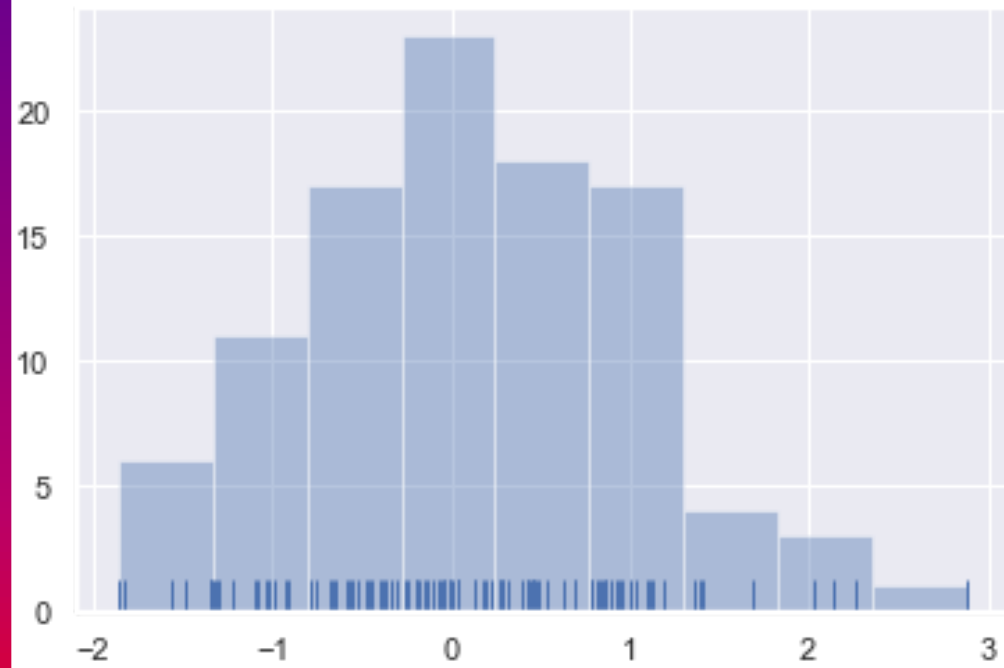
Inner : swarmplot



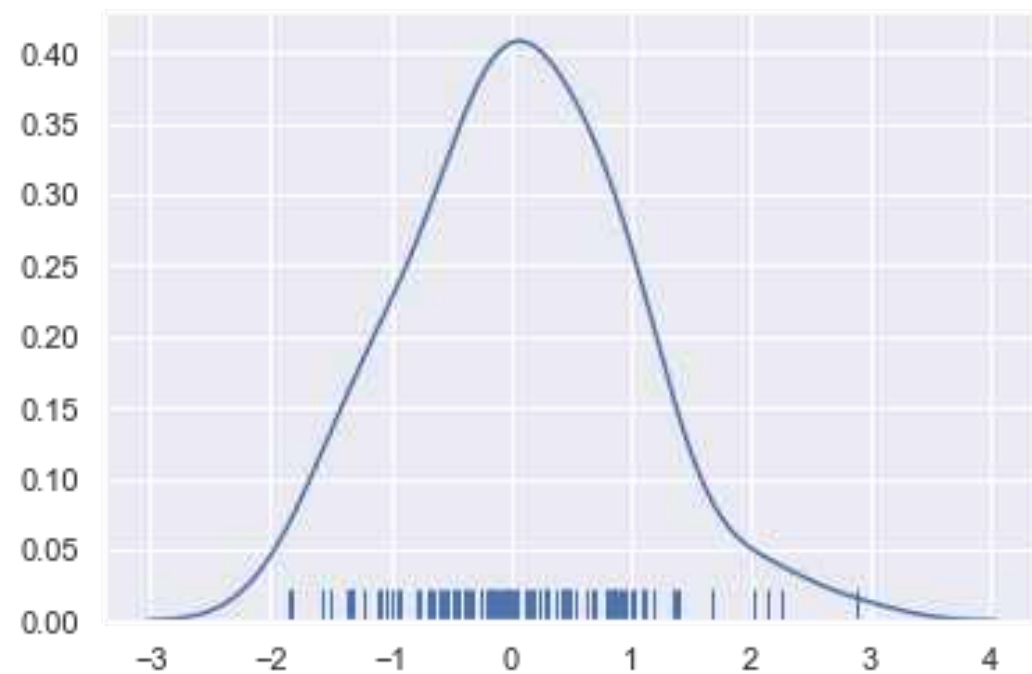
Inner : stick

# Visualisation : Distribution univariée et bivariée

## □ Distribution univariée



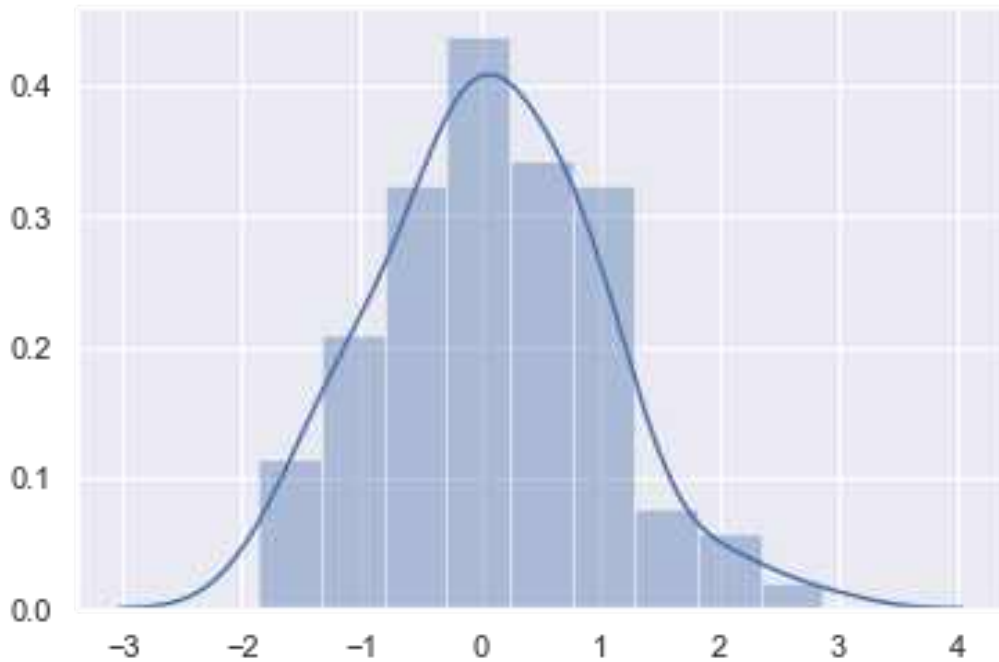
Observation  
En histogramme (**hist plot**)



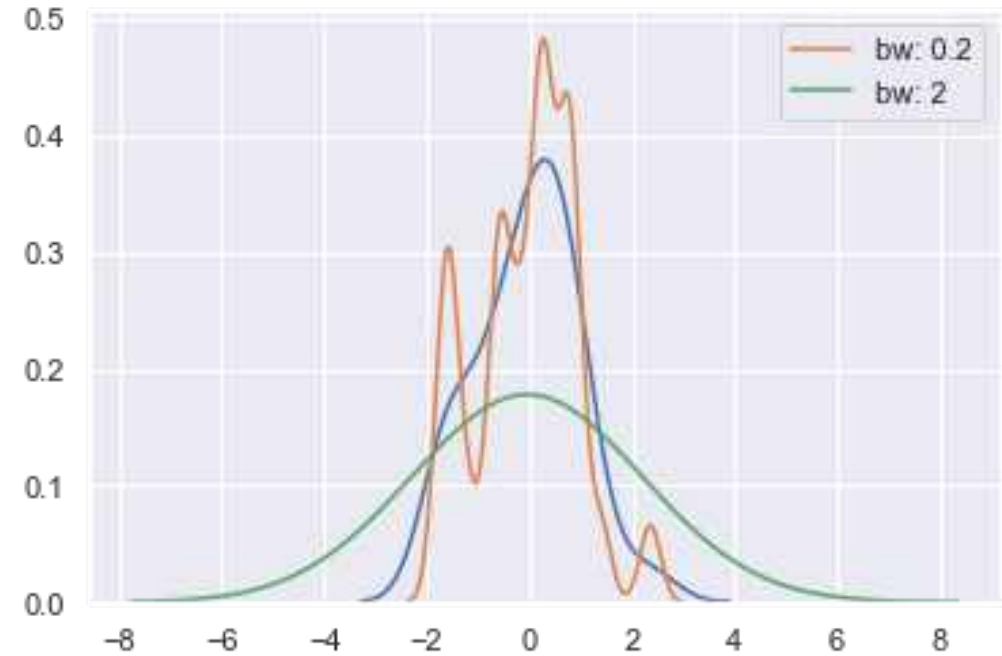
Estimation de distributions  
suivant loi normale  
En courbe (**kde plot**)

# Visualisation : Distribution univariée et bivariable

## □ Distribution univariée



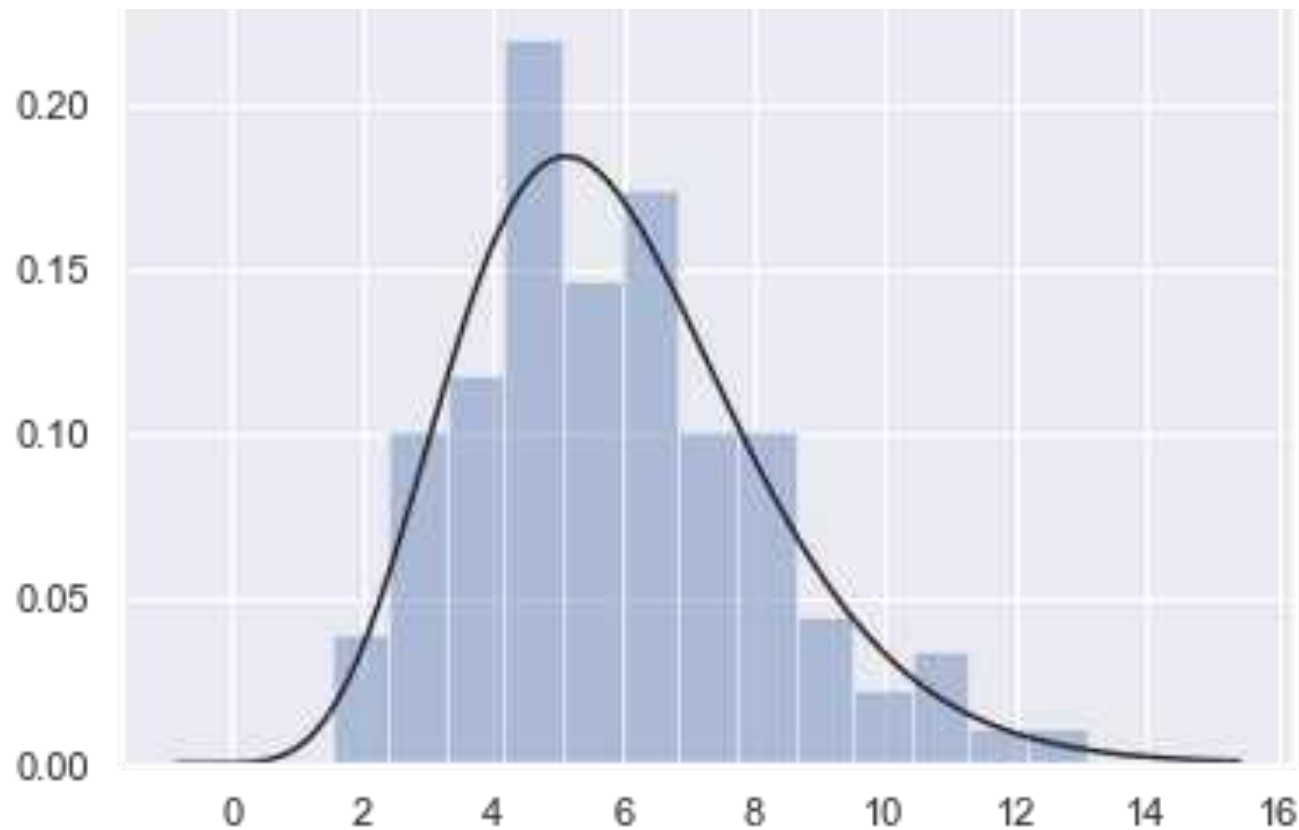
Histogramme + courbe



Large de bande (bandwidth bw)  
3<sup>e</sup> variable : couleur (hue)

# Visualisation : Distribution univariée et bivariable

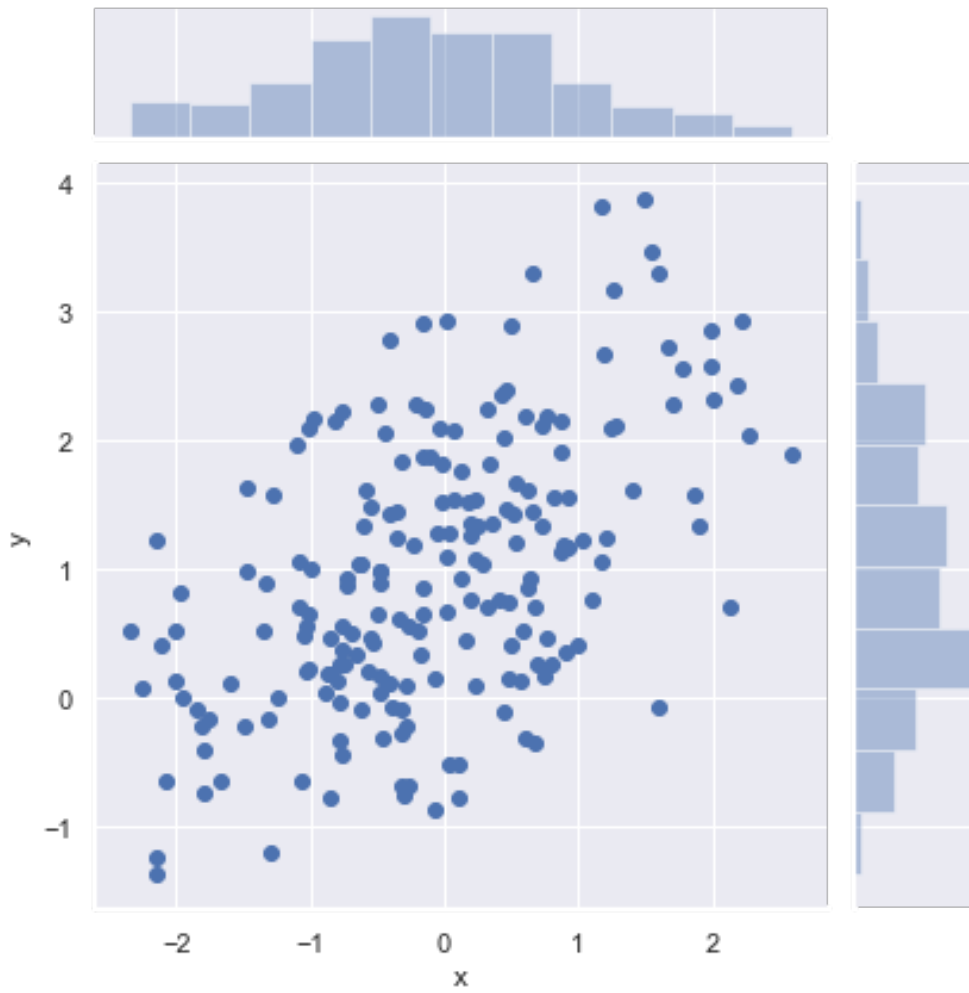
## □ Distribution univariée



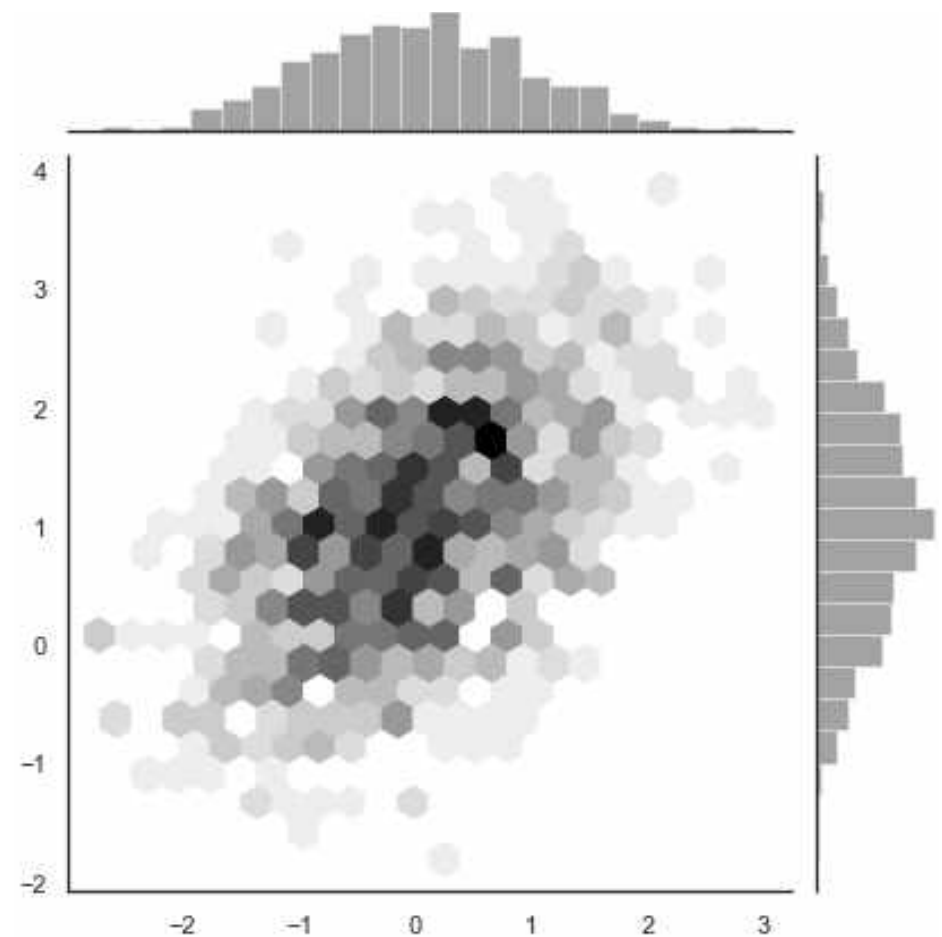
Estimation de distributions  
suivant une loi donnée  
Ex: loi gamma

# Visualisation : Distribution univariée et bivariée

❑ Distribution bivariée : graphique conjoint bivarié + 2 univariés (**joint plot**, **joint grid**)



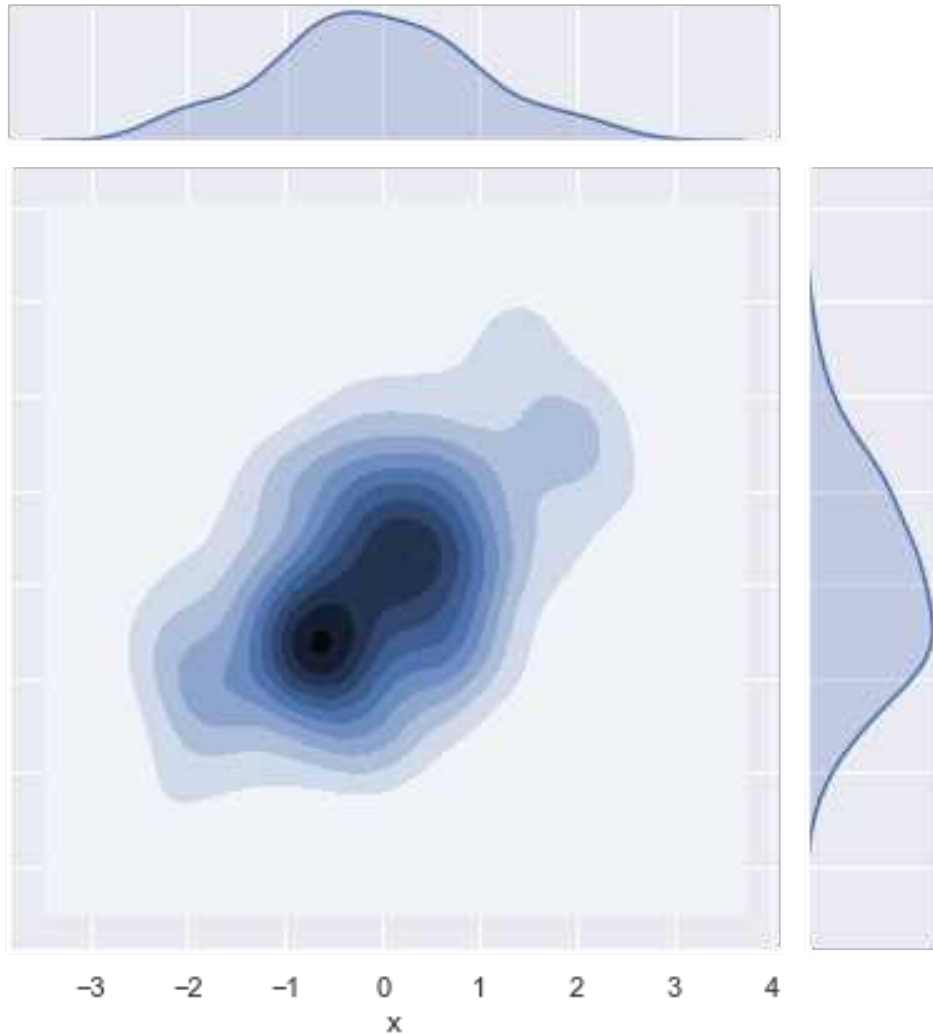
Bivarié (scatter) + univarié (hist)



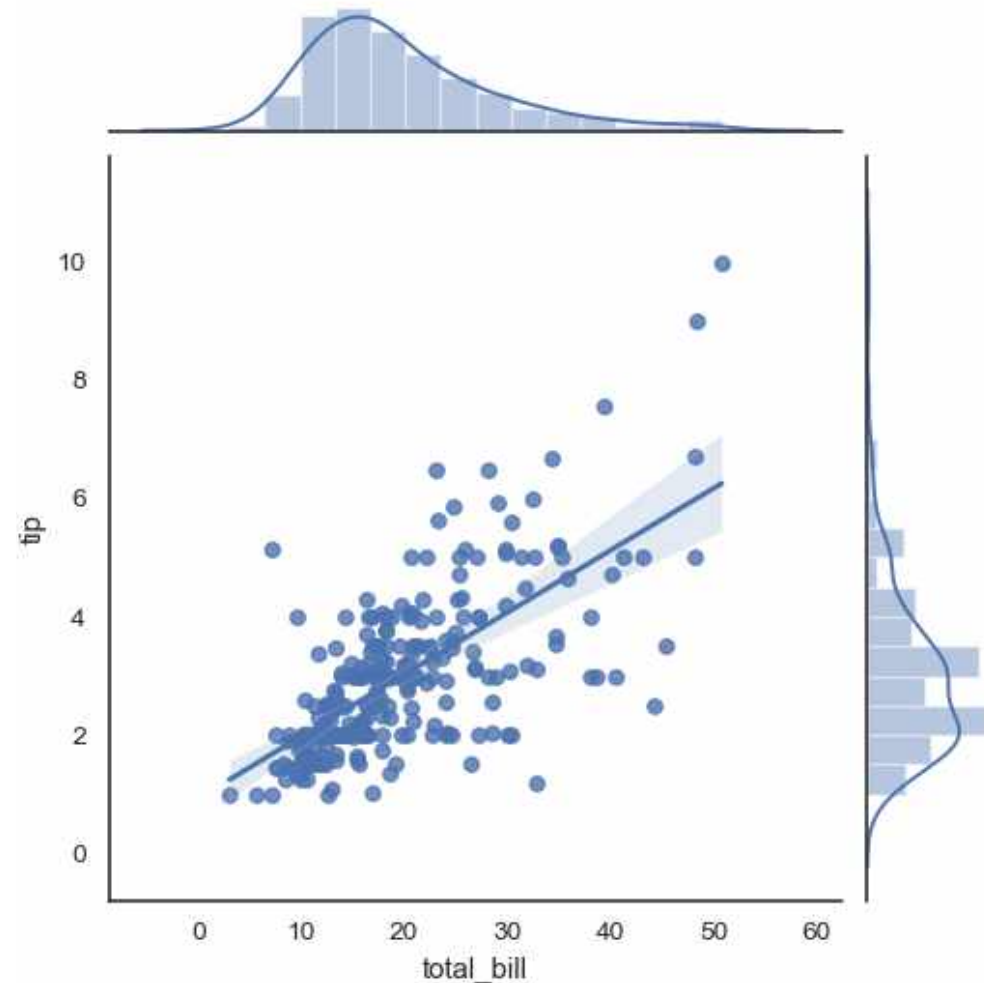
Bivarié hexagone (hex) + univarié (hist)

# Visualisation : Distribution univariée et bivariée

- Distribution bivariée : graphique conjoint bivarié + 2 univariés (**joint plot**, **joint grid**)



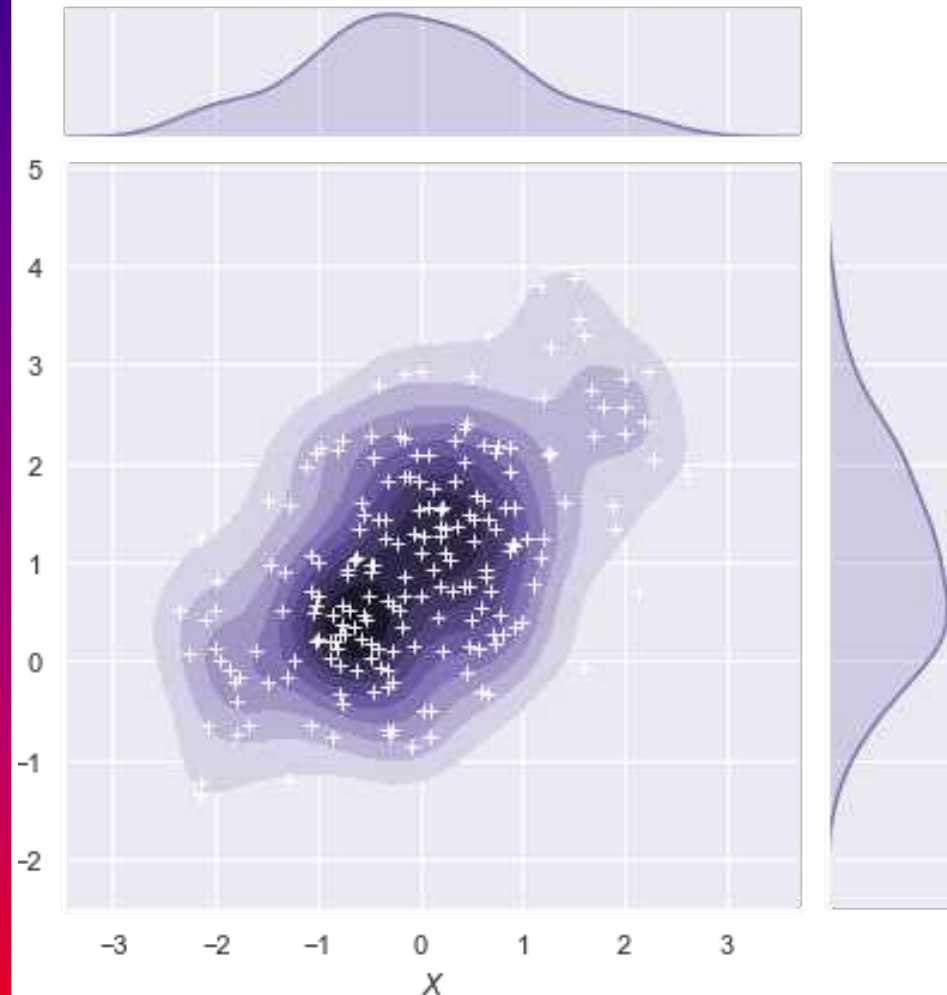
Bivarié (kde) +  
univarié (kde)



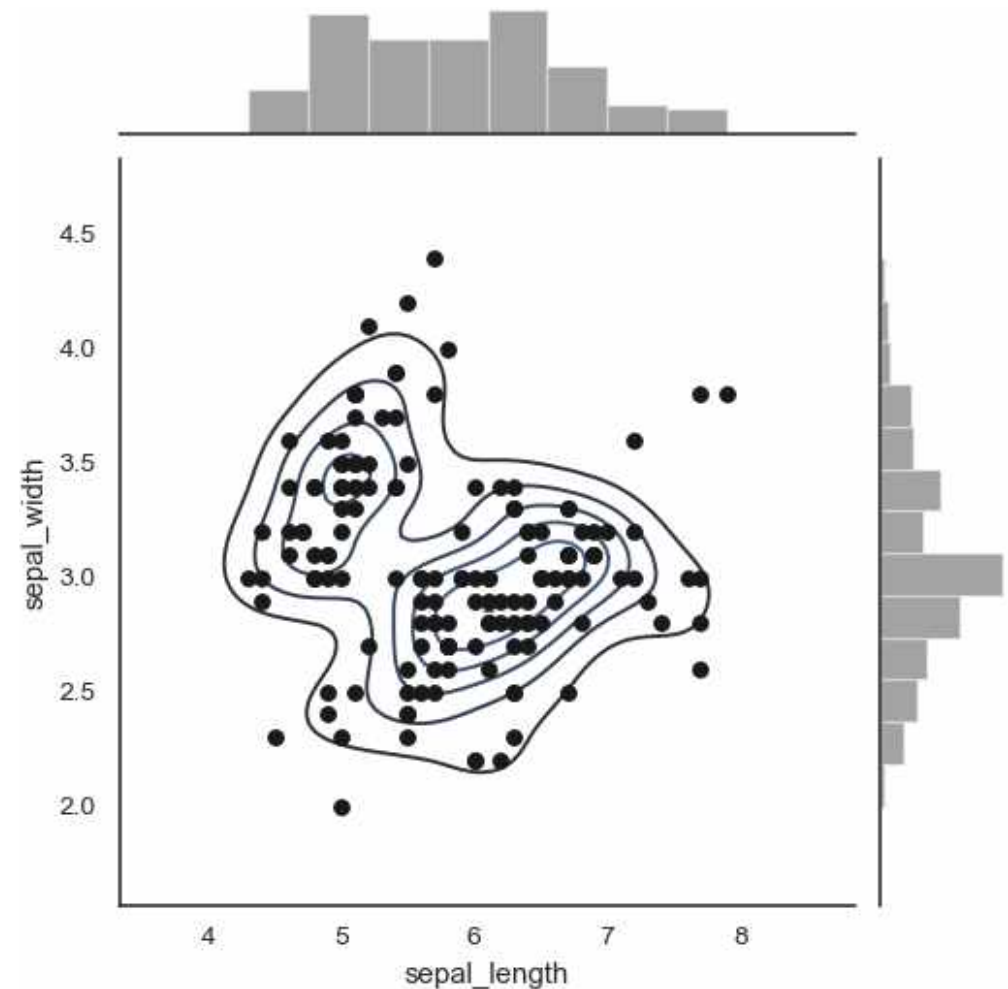
Bivarié : regression (reg) +  
univarié (hist + kde)

# Visualisation : Distribution univariée et bivariée

❑ Distribution bivariée : graphique conjoint bivarié + 2 univariés (**joint plot**, **joint grid**)



Bivarié (kde + scatter) +  
univarié (kde)



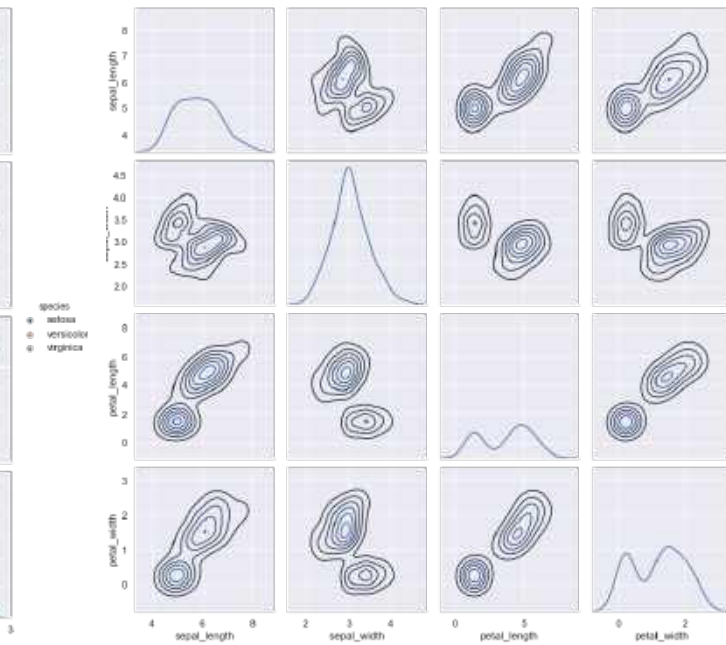
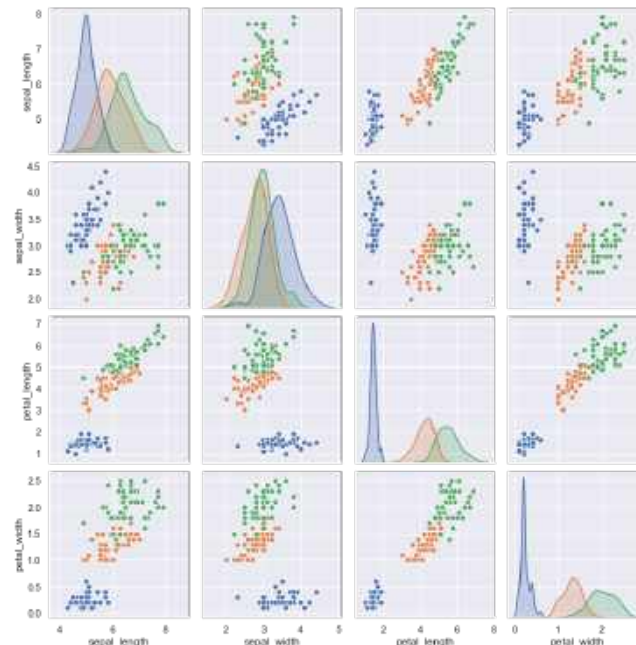
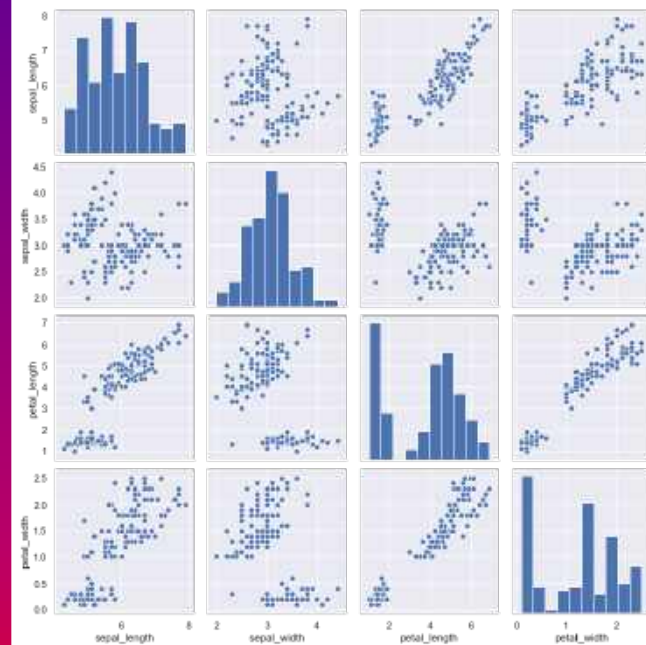
Bivarié : (scatter + kde) +  
univarié (hist)

# Visualisation : Distributions univariée et bivariée

❑ Distribution bivariée : pour toutes les paires d'attributs (**pair plot**, **pair grid**)

Case diagonale univariée  
(diag\_kind : hist, kde)

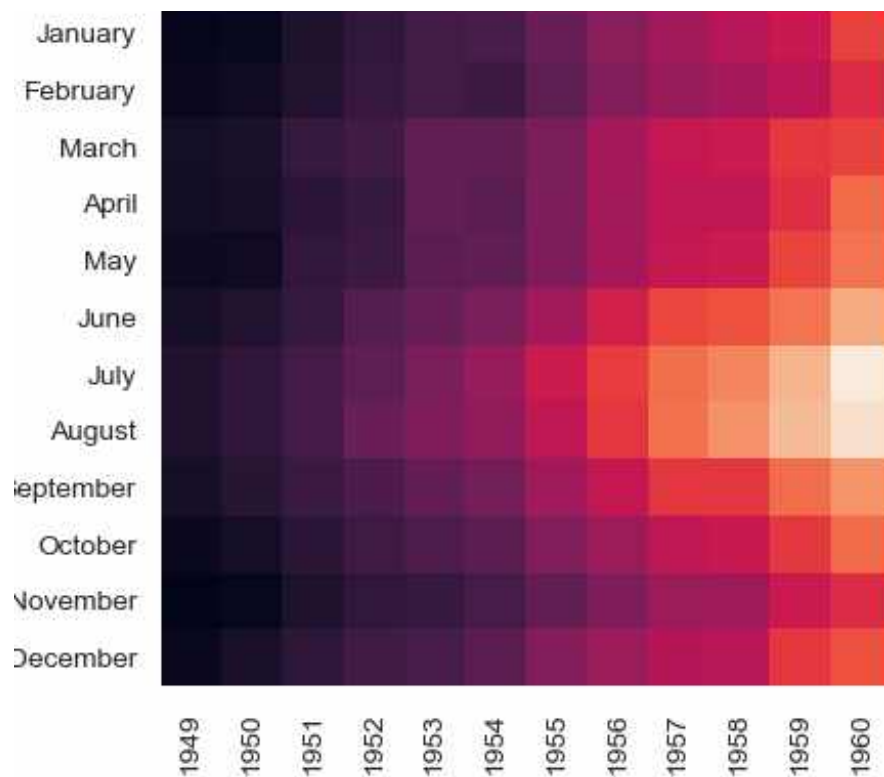
Autre cas bivarié  
(kind : scatter, reg, kde, hex)



# Visualisation : Données rectangulaires

❑ Matrice de couleurs (**heatmap**,**clustermat**): valeurs => couleurs

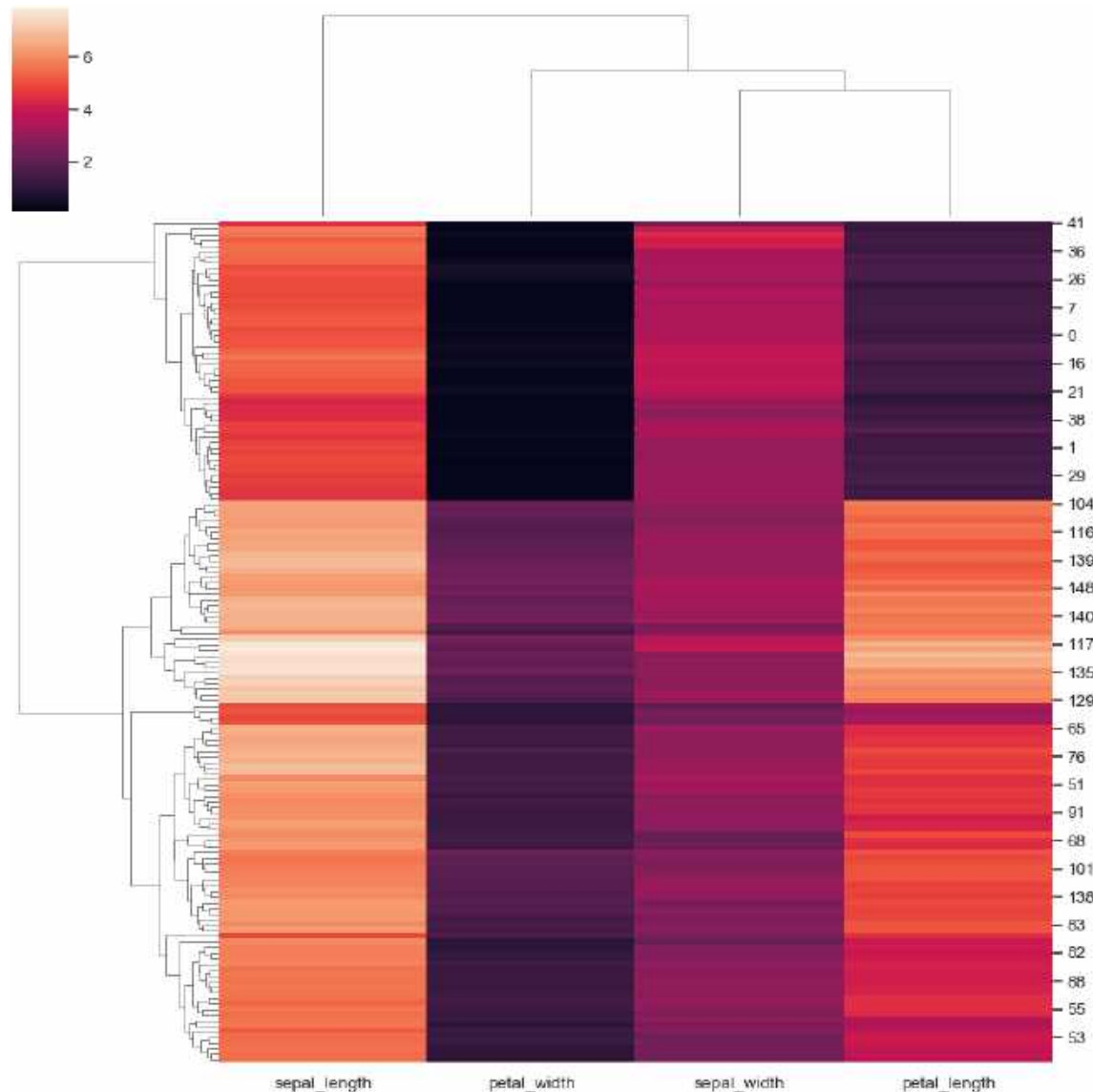
Heatmap



# Visualisation : Données rectangulaires

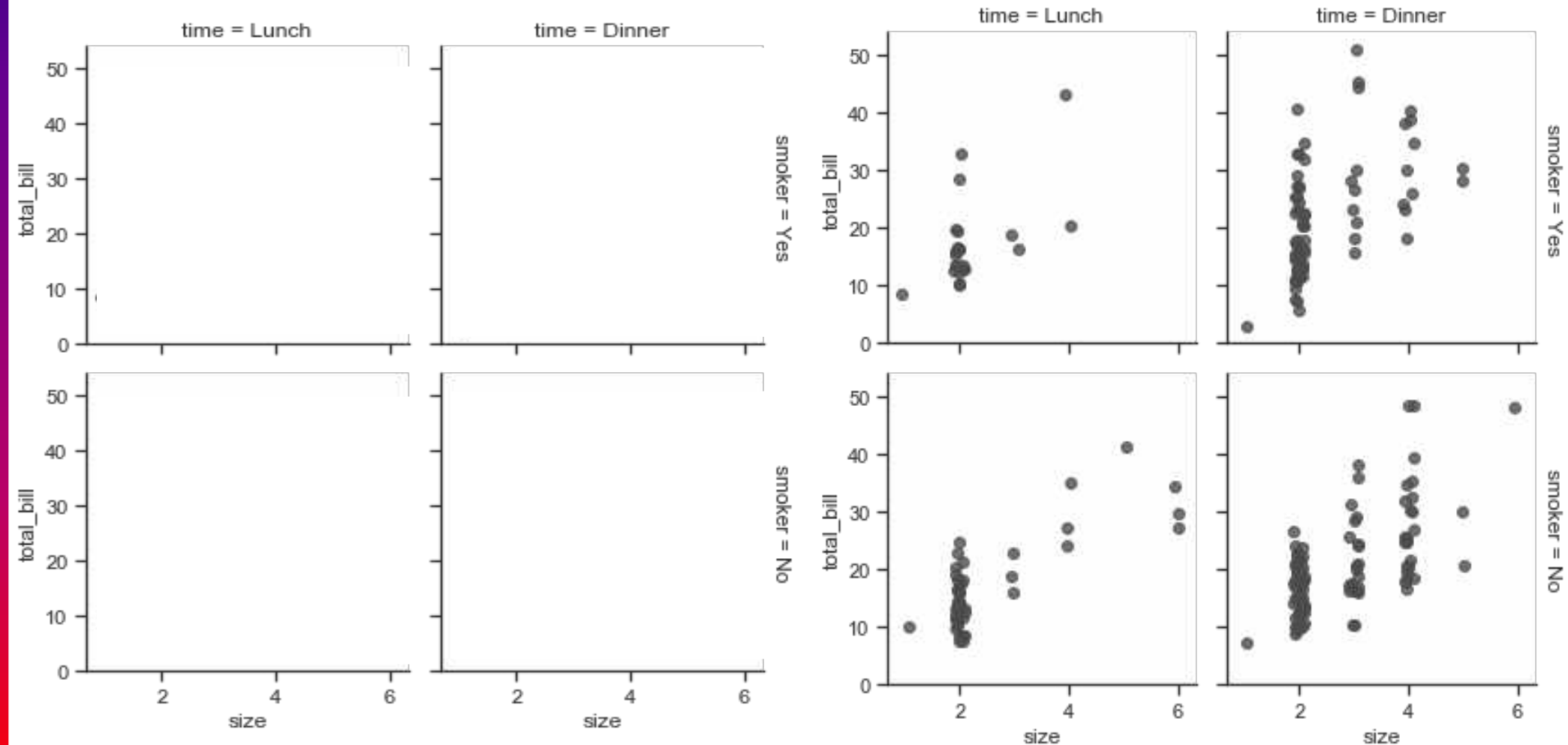
❑ Matrice de couleurs (**heatmap**, **clustermat**): valeurs => couleurs

**Clustermat**: choix de la mesure de distance ou de similarité.



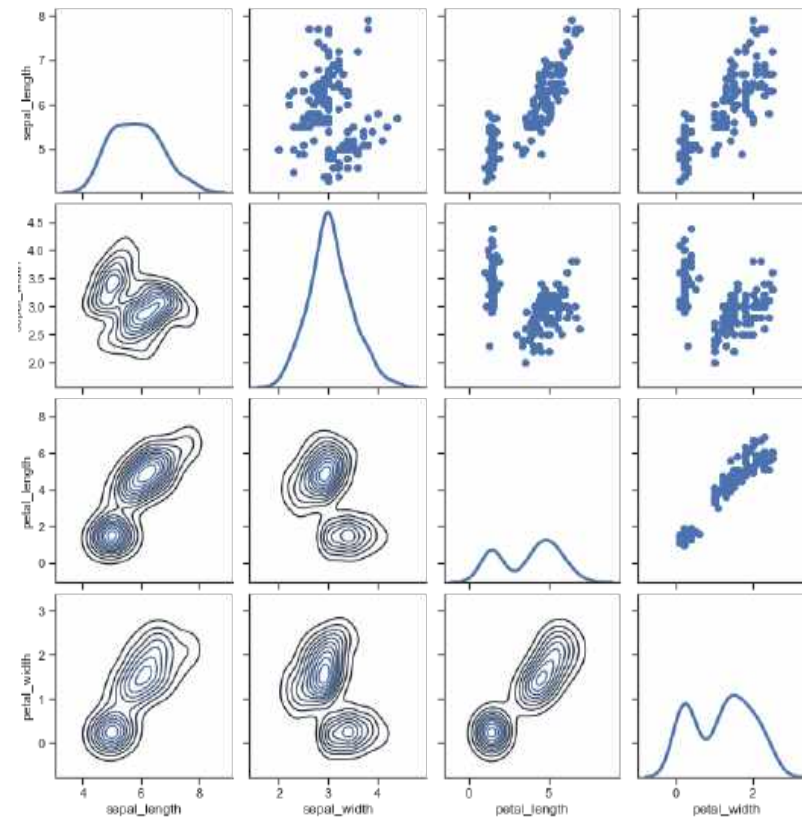
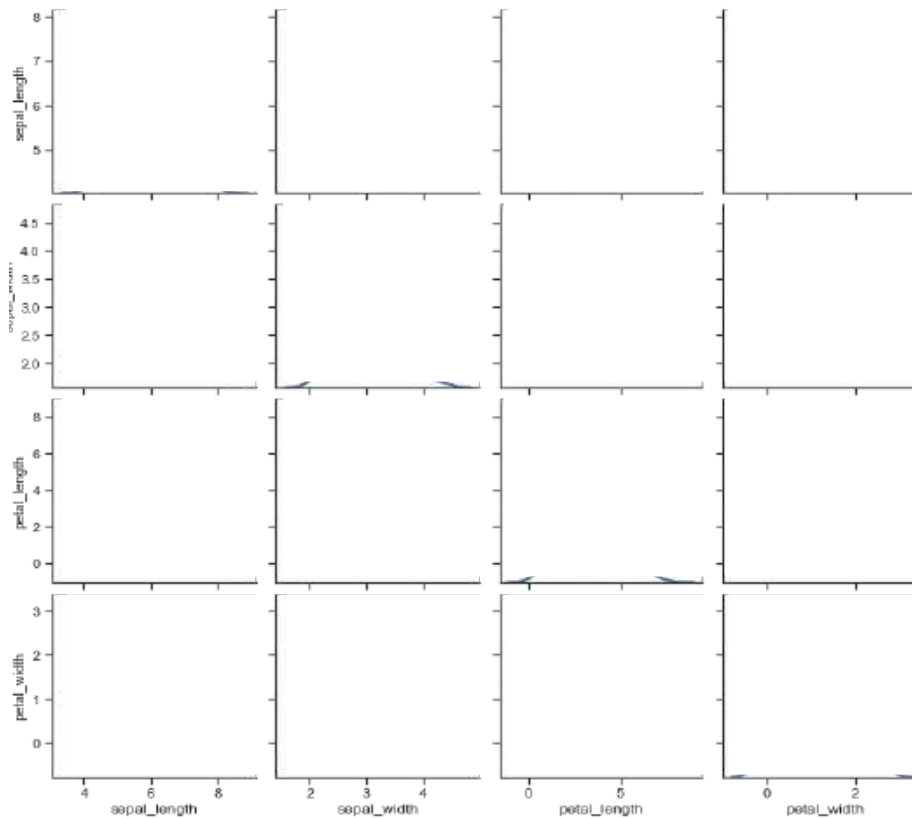
# Visualisation : Grille multi-graphique

- Groupement suivant deux variables (**FacetGrid**) (row, column)



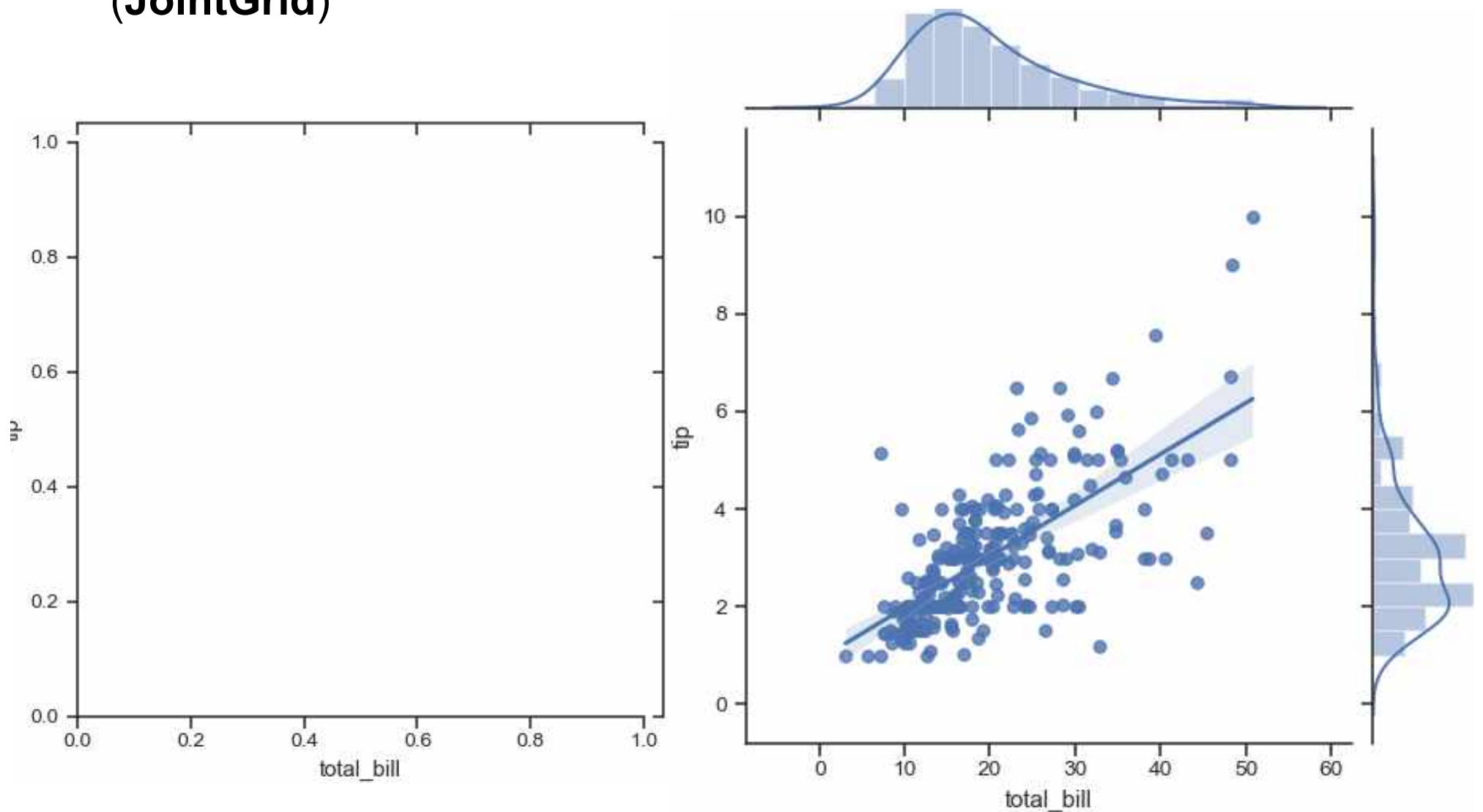
# Visualisation : Grille multi-graphique

- ❑ Grille de relations deux-à-deux entre toutes les variables (**PairGrid**)



# Visualisation : Grille multi-graphique

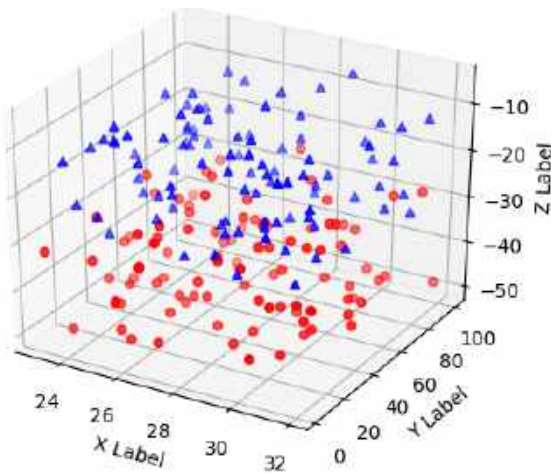
- ❑ Relation ou distribution bivariable couplée aux 2 distributions univariées (**JointGrid**)



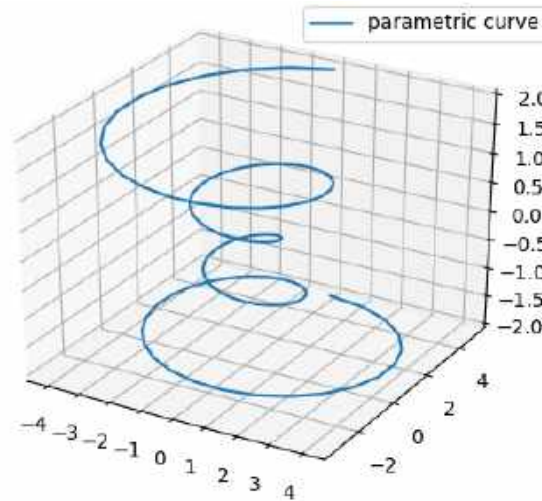
# Visualisation : 3D (Matplotlib)

## ❑ Relation entre 3 variables

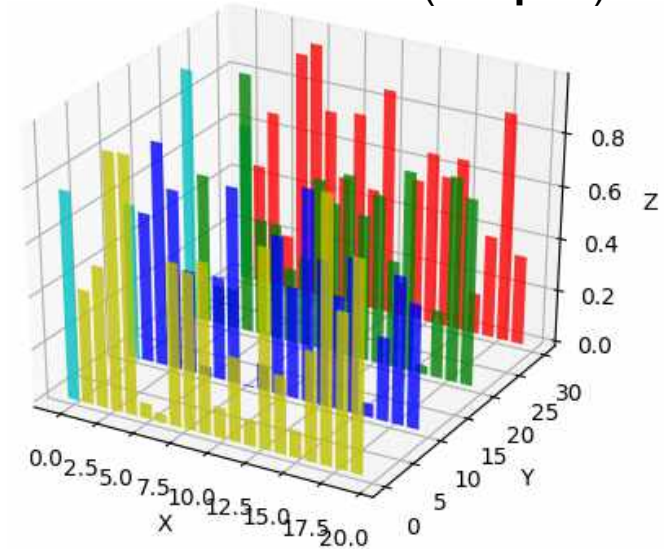
Point3D (**Scatter plot**)



Courbe reliant Point3D (**line plot**)

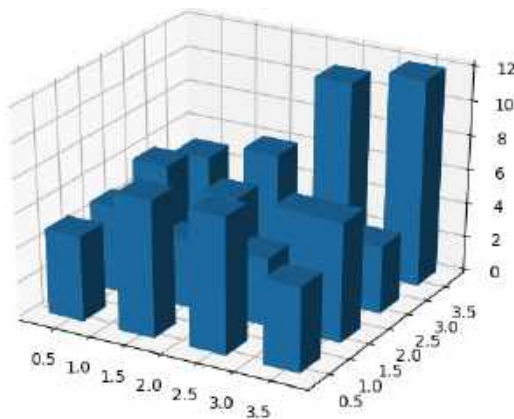


Courbe reliant Point3D (**bar plot**)



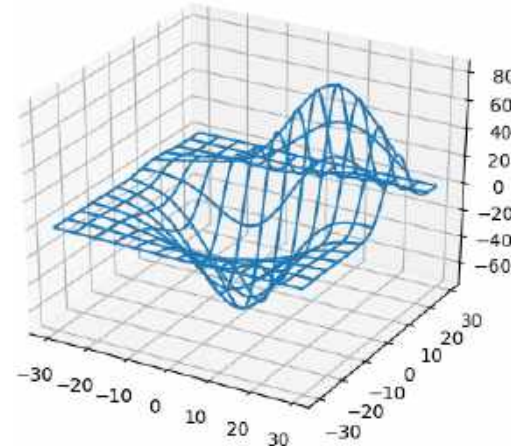
## ❑ Distribution bi-variée

Histogramme (**hist plot**)



## ❑ Distribution tri-variée

wireframe plot, surface plot



# Mesure de similarité / dissimilarité

- ❑ Similarité : mesure de la ressemblance entre deux objets. Plus les objets se ressemblent, plus la mesure est élevée.
- ❑ Dissimilarité : mesure de la différence entre deux objets. Plus les objets sont différents, plus la mesure est élevée.

Exemple : distance

Matrice rectangulaire ( $n \times m$ ) de  $n$  objets représentés sur  $k$  dimensions

$$\begin{bmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1m} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{ij} & \dots & x_{im} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nj} & \dots & x_{nm} \end{bmatrix}$$

Matrice carrée ( $n \times n$ ) symétrique ou triangulaire des distances deux-à-deux

$$\begin{bmatrix} 0 & & & & \\ d_{2,1} & 0 & & & \\ d_{3,1} & d_{3,2} & 0 & & \\ \vdots & \vdots & \vdots & & \\ d_{n,1} & d_{n,2} & \dots & d_{n,n-1} & 0 \end{bmatrix}$$

# Similarité / dissimilarité entre vecteurs de valeurs catégorielles

❑ Proportion d'égalités (similarité):  $s(x_i, x_j) = \frac{\sum_{k=1}^m (x_{ik} \equiv x_{jk})}{m}$

❑ Proportion de différences (dissimilarité):  $d(x_i, x_j) = \frac{\sum_{k=1}^m (x_{ik} \neq x_{jk})}{m}$

- ❑ Autre méthode : transformer chaque attribut catégoriel à  $v$  valeurs en  $v$  attributs binaires, puis appliquer une mesure de similarité ou dissimilarité entre vecteurs binaires.

# Similarité / dissimilarité entre vecteurs de valeurs binaires

binaires

❑ Matrice de contingence:

		$x_j$	
		0	1
$x_i$	0	m00	m01
	1	m10	m11

$$m = m00 + m01 + m10 + m11$$

❑ Distance pour variables binaires symétriques:

$$d(x_i, x_j) = \frac{m01 + m10}{m}$$

❑ Distance pour variables binaires asymétriques:

$$d(x_i, x_j) = \frac{m01 + m10}{m01 + m10 + m11}$$

❑ Coefficient de correspondance simple (SMC) (similarité pour variables binaires symétriques):

$$SMC(x_i, x_j) = \frac{m00 + m11}{m}$$

❑ Coefficient de Jaccard (similarité pour variables binaires non-symétriques):

$$Jaccard(x_i, x_j) = \frac{m11}{m11 + m01 + m10}$$

# Similarité / dissimilarité entre vecteurs de valeurs ordinales

- ❑ Pour chaque variable ordinale, remplacer chaque valeur par son rang, puis appliquer une mesure de similarité ou dissimilarité entre vecteurs numériques.

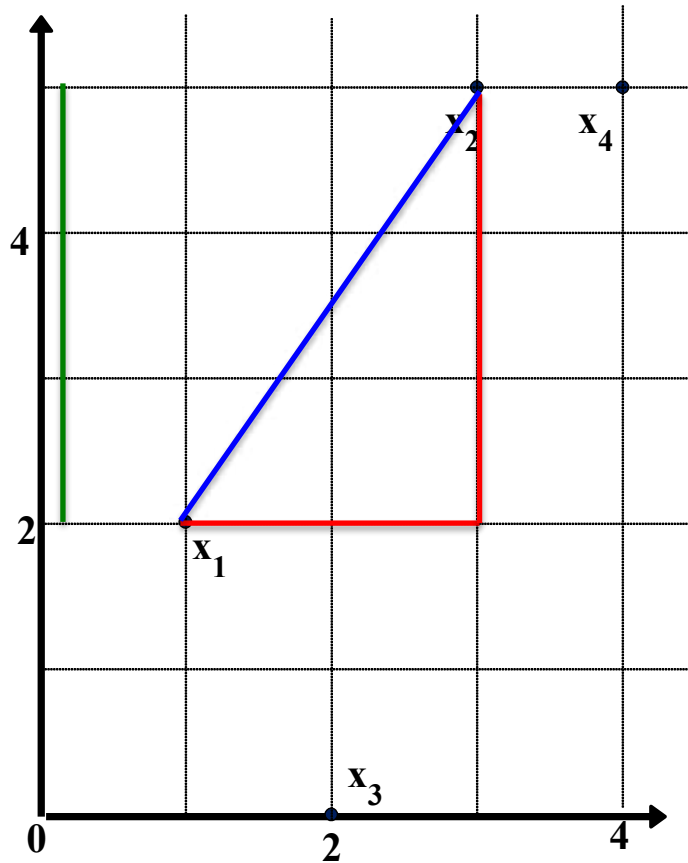
# Similarité / dissimilarité entre vecteurs de valeurs numériques

- ❑ Distance de Minkowski : 
$$d(x_i, x_j) = \sqrt[h]{\sum_{k=1}^m |x_{ik} - x_{jk}|^h}$$
- ❑ Si  $h = 1$  (norme L1) : Distance de Manhattan (Distance de Hamming pour vecteurs binaires)
- ❑ Si  $h = 2$  (norme L2) : Distance euclidienne
- ❑ Si  $h = \text{infini}$  : Supremum (plus grande différence parmi tous les attributs)

# Similarité / dissimilarité entre vecteurs de valeurs numériques

□ Distance de Minkowski :

	attribut 1	attribut 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



**Manhattan ( $L_1$ )**

**Matrices de dissimilarité**

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

**Euclidienne ( $L_2$ )**

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

**Supremum**

$L_\infty$	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

# Similarité / dissimilarité entre vecteurs de valeurs numériques

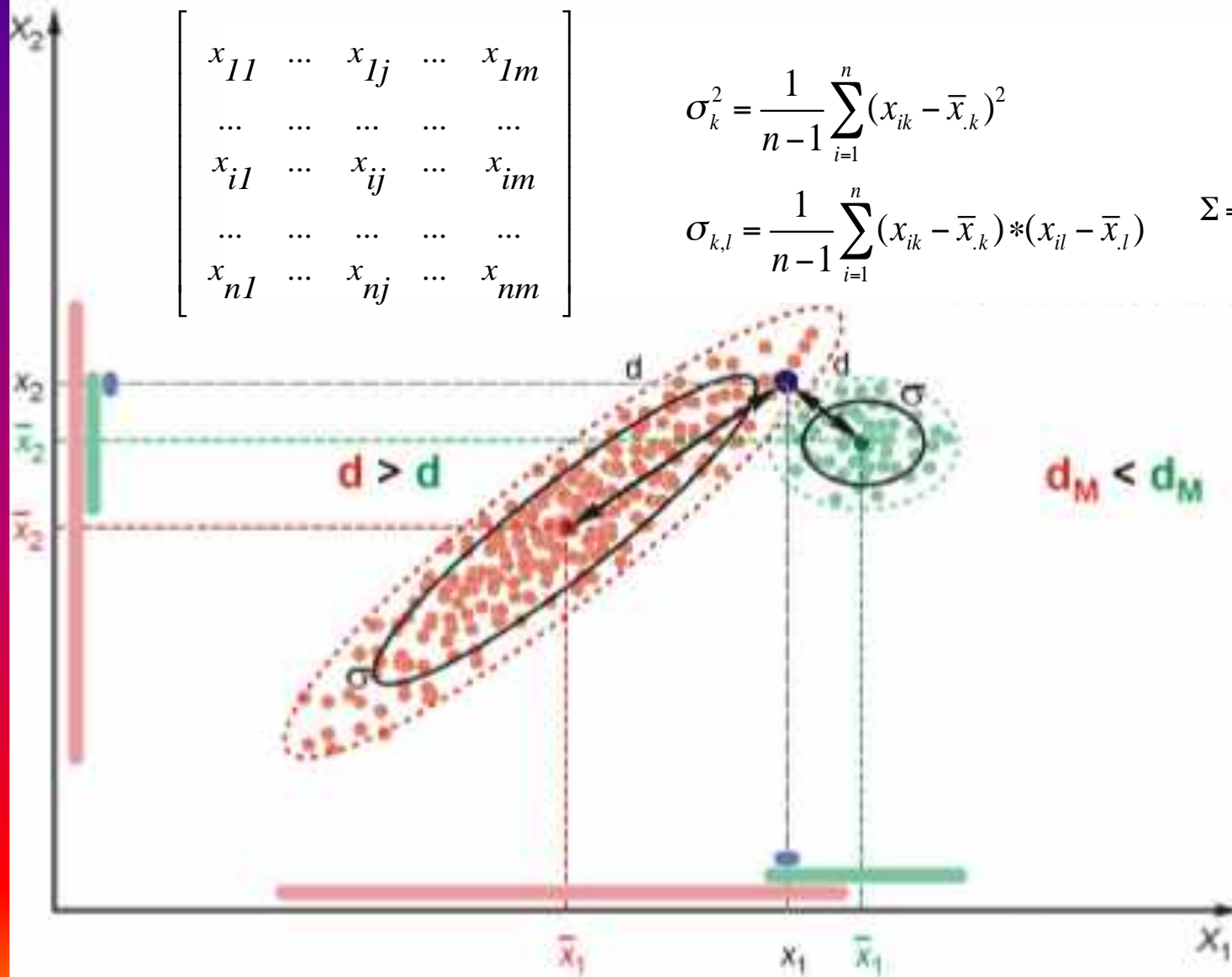
- Distance de Mahalanobis : distance entre deux points en tenant compte de la contribution de différentes variances et des corrélations existant entre elles.

$$\begin{bmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1m} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{ij} & \dots & x_{im} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nj} & \dots & x_{nm} \end{bmatrix}$$

$$\sigma_k^2 = \frac{1}{n-1} \sum_{i=1}^n (x_{ik} - \bar{x}_{.k})^2$$

$$\sigma_{k,l} = \frac{1}{n-1} \sum_{i=1}^n (x_{ik} - \bar{x}_{.k}) * (x_{il} - \bar{x}_{.l})$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{2,1} & \sigma_{3,1} & \dots & \sigma_{m,1} \\ \sigma_{2,1} & \sigma_2^2 & \sigma_{3,2} & \dots & \sigma_{m,2} \\ \sigma_{3,1} & \sigma_{3,2} & \sigma_3^2 & \dots & \sigma_{m,3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sigma_{m,1} & \sigma_{m,2} & \dots & \sigma_{m,m-1} & \sigma_m^2 \end{bmatrix}$$



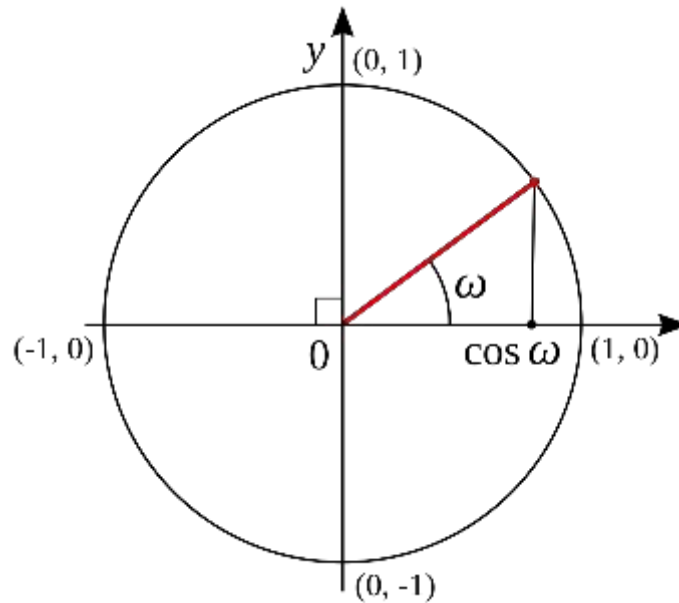
$$D = x_i - x_j$$

$$\text{Euclidienne : } d(x_i, x_j) = \sqrt{D^t \cdot D}$$

$$\text{Mahalanobis : } d_m(x_i, x_j) = \sqrt{D^t \cdot \Sigma^{-1} \cdot D}$$

# Similarité / dissimilarité entre vecteurs de valeurs numériques

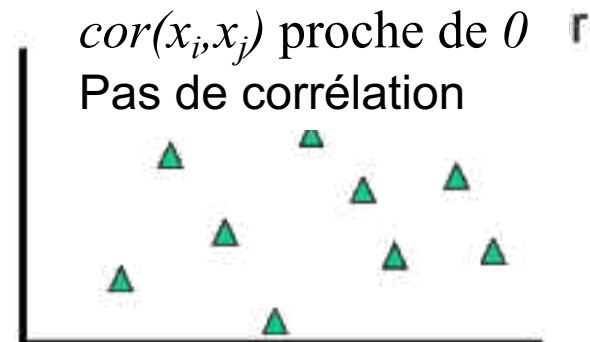
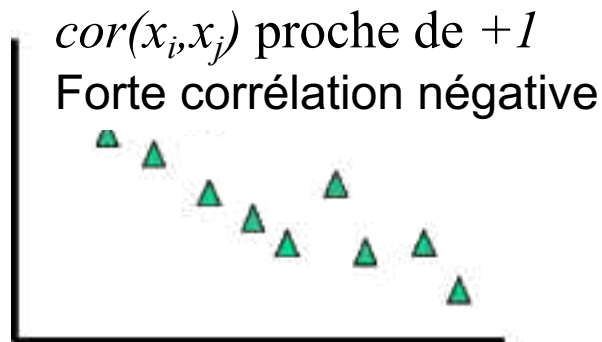
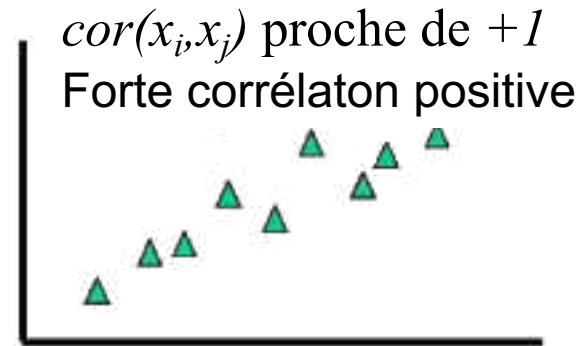
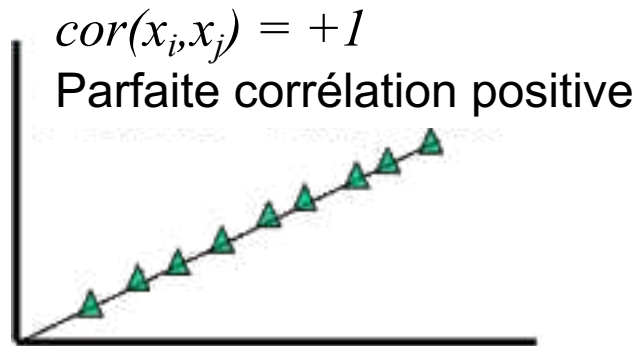
- ❑ Similarité de Cosinus : cosinus de l'angle entre les vecteurs  $x_i$  et  $x_j$   
Compare uniquement l'orientation des deux vecteurs



$$\cos(x_i, x_j) = \frac{x_i \bullet x_j}{\|x_i\| * \|x_j\|} = \frac{\sum_{k=1}^m x_{ik} * x_{jk}}{\sqrt{\sum_{k=1}^m x_{ik}^2} * \sqrt{\sum_{k=1}^m x_{jk}^2}}$$

# Similarité / dissimilarité entre vecteurs de valeurs numériques

□ Coefficient de corrélation :



$$cor(x_i, x_j) = \frac{\sum_{k=1}^m (x_{ik} - \bar{x}_{i.}) * (x_{jk} - \bar{x}_{j.})}{\sqrt{\sum_{k=1}^m (x_{ik} - \bar{x}_{i.})^2} * \sqrt{\sum_{k=1}^m (x_{jk} - \bar{x}_{j.})^2}}$$

# Références

- [1] PEDREGOSA et al. : *Scikit-learn : Machine Learning in Python*. JMLR 12, pp. 2825-2830. (User guide and API : [https ://scikit-learn.org/stable/](https://scikit-learn.org/stable/)), 2011.
  
- [2] Jiawei HAN, Micheline KAMBER, Jian PEI. *DataMining: Concepts and Techniques (Third edition)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2011.