SY09 P2025
TD/TP 2 — Visualisation de données

`numpy`=2.2.3; `seaborn`=0.13.2; `matplotlib`=3.10.1; `pandas`=2.2.3

La visualisation est une part importante de l'analyse de données que ce soit lors de travaux préliminaires d'exploration pour mettre en évidence des informations cachées ou pour la présentation de résultats.

1 Travaux pratiques

Il existe de nombreux outils permettant de faire de la visualisation. Parmi les solutions existantes en Python, nous utiliserons la bibliothèque `seaborn` qui a l'avantage d'être assez exhaustive tout en proposant une interface consistante. Elle s'appuie sur la bibliothèque `matplotlib` que nous utiliserons également pour adapter certaines figures.

Un alias couramment utilisé est le suivant :

```
import seaborn as sns
```

Nous utiliserons les jeux de données `titanic` et `iris`, présents dans `seaborn`, qu'on peut charger avec les instructions suivantes :

```
titanic = sns.load_dataset("titanic")
iris = sns.load_dataset("iris")
```

1.1 Visualisation univariée

La visualisation univariée consiste à étudier un seul prédicteur à la fois sans tenir compte des liens avec les autres prédicteurs.

Prédicteur quantitatif

Pour les prédicteurs quantitatifs, on peut utiliser les boites à moustaches. Avec `seaborn`, on a la syntaxe suivante :

```
sns.boxplot(x=<Série Pandas>)
```

1 Tracer la boite à moustaches des âges des passagers du Titanic. Quel est l'âge médian ?

L'autre visualisation classique d'une variable quantitative est l'histogramme qu'on peut obtenir avec l'instruction suivante :

```
sns.histplot(<Série Pandas>)
```

2 Tracer l'histogramme des âges des passagers. Que font les arguments optionnels suivants ?

1. `bins`
2. `kde`

Prédicteur qualitatif

Pour les prédicteurs qualitatifs, on utilise un diagramme en barres avec la syntaxe suivante :

```
sns.countplot(x=<Colonne>, data=<DataFrame Pandas>)
```

3 Représenter la donnée des classes des passagers.

1.2 Visualisation multivariée

La visualisation multivariée consiste à étudier les relations pouvant exister entre plusieurs prédicteurs. Les techniques utilisées varient principalement en fonction de la nature des prédicteurs : qualitatifs ou quantitatifs.

Quantitatif *vs* quantitatif

Diagramme de dispersion Le diagramme de dispersion permet de visualiser les relations existantes entre deux variables qualitatives. On trace un diagramme de dispersion avec l'instruction suivante :

```
sns.scatterplot(
  x=<Colonne des abscisses>,
  y=<Colonne des ordonnées>,
  data=<DataFrame Pandas>
)
```

4 Tracer le diagramme de dispersion de la largeur du sépale en fonction de sa longueur.

Il est possible de visualiser d'autres variables en jouant sur la couleur, la forme ou la taille des points avec les arguments `hue`, `style` et `size`.

5 Rajouter la donnée de classe avec la couleur.

6 En plus de la donnée de classe, ajouter la longueur du pétale avec la taille du point.

Diagramme de corrélation Le diagramme de corrélation est utile pour avoir une idée des liens de type linéaire entre des variables quantitatives. Le diagramme de corrélation n'est pas directement supporté par `seaborn` : il faut d'abord créer ce tableau à l'aide de la fonction `Pandas corr` et utiliser ensuite la visualisation `heatmap`.

```
corr = <DataFrame Pandas>.corr()
sns.heatmap(corr)
```

7 Étudier les corrélations linéaires des variables quantitatives du jeu de données iris. Peut-on dire que les largeurs du sépale et du pétale sont corrélés négativement ? Étudier l'influence de l'espèce.

Qualitatif *vs* quantitatif

Diagramme en barres avec dispersion La fonction `barplot` permet de visualiser une variable quantitative avec une indication sur la dispersion des données sous forme d'un intervalle de confiance en fonction d'une colonne qualitative. La syntaxe est la suivante :

```
sns.barplot(
  x=<Colonne qualitative>,
  y=<Colonne quantitative>,
  data=<DataFrame Pandas>
)
```

8 Avec le jeu de données `iris`, visualiser la longueur des sépales en fonction de l'espèce.

Boites à moustaches multiples Au lieu d'afficher un diagramme en bâtons, on peut afficher des boites à moustaches. La syntaxe est similaire.

```
sns.boxplot(
  x=<Colonne qualitative>,
  y=<Colonne quantitative>,
  data=<DataFrame Pandas>
)
```

9 Avec le jeu de données `iris`, visualiser les boites à moustaches des longueur des sépales en fonction de l'espèce.

Qualitatif *vs* qualitatif

Lorsque les deux variables sont qualitatives, les représentations possibles sont centrées autour du tableau de contingence. On peut utiliser la fonction `countplot` en fournissant la deuxième variable qualitative en argument.

```
sns.countplot(
  x=<Colonne qualitative>,
  hue=<Colonne qualitative>,
  data=<DataFrame Pandas>
)
```

10 Tracer les lieux d'embarcation en fonction de la classe.

11 Visualiser le descripteur « who » en fonction de la classe.

Représentation multiple

Dans certains cas, il est utile de pouvoir appliquer une même visualisation en parallèle sur des sous-jeux de données identifiés par les modalités d'un ou plusieurs descripteurs. Chacune des représentations est appelée une facette.

Pour réaliser cela avec `seaborn`, suivant le type de la facette, il faut utiliser les fonctions suivantes :
- `catplot` pour des facettes de type *qualitatif* vs *quantitatif* ou *qualitatif*
- `relplot` pour des facettes de type *quantitatif* vs *quantitatif*
- `displot` pour des facettes de type histogramme

Pour chacune des fonctions, on spécifie comment partitionner le jeu de données avec les arguments `row` et `col` qui sont des noms de colonnes dans le jeu de données.

```
sns.relplot(
  x=<Colonne des abscisses>,
  y=<Colonne des ordonnées>,
  kind=<Type de facette>,
  row=<Colonne des lignes>,
  col=<Colonne des colonnes>,
  data=<DataFrame Pandas>
)
```

12 Visualiser « age » en fonction de « class » avec trois histogrammes.

13 Visualiser « age » en fonction de « class » et « who » avec des histogrammes.

1.3 Visualisation du jeu de données `sy02-p2019.csv`

14 Charger le jeu de données contenu dans le fichier `data/sy02-p2019.csv`. Quel est le type de la colonne « Note médian » ? Pourquoi ? On pourra utiliser l'argument `na_values`.

15 Visualiser les effectifs des différentes branches. Rajouter

1. l'information de « niveau »,
2. la note ECTS.

Les modalités de cette seconde variable sont-elles dans l'ordre ?

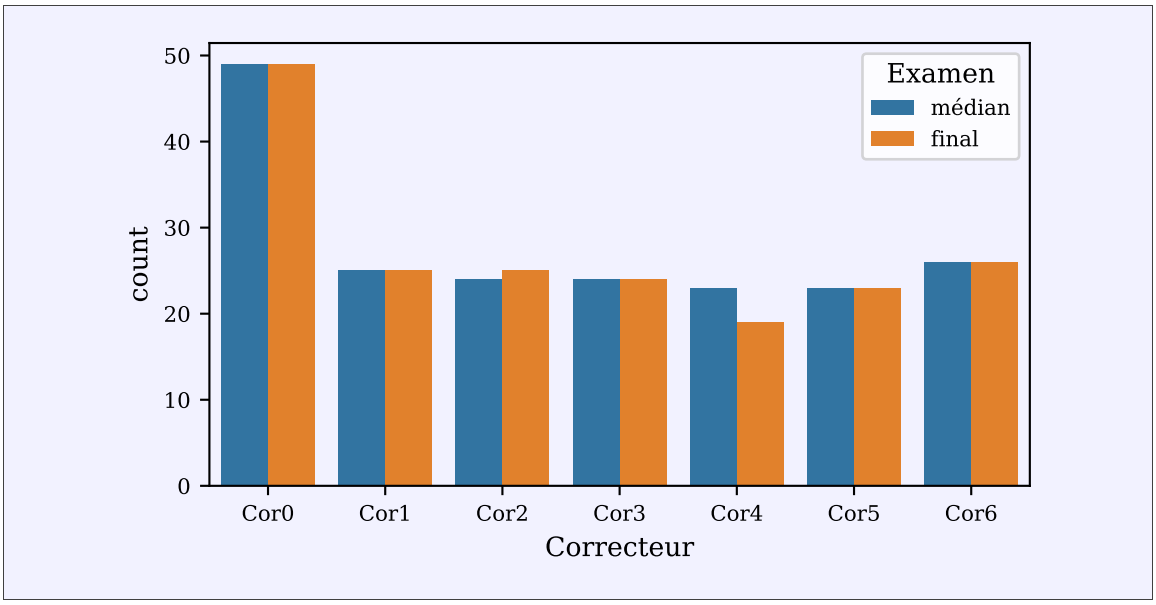
16 Visualiser les distributions des notes de médian et des notes de final. Afficher le diagramme de dispersion.

Synthétiser les trois visualisations précédentes en utilisant `sns.jointplot`.

17 Tracer en parallèle les boites à moustaches des notes de médian et de final. On pourra utiliser la fonction `melt`.

18 Trouver une visualisation du nombre de copies corrigées sur tout le semestre (médian et final confondu) par correcteur en indiquant pour chaque correcteur la part des copies correspondant à un même étudiant (médian et final corrigés par ce même correcteur) et la part des copies qui ne sont pas corrigées par le même correcteur.

19 Comment obtenir la figure suivante ?



2 Exercices

2.1 Histogramme et estimation de densité

On considère une population $\mathcal{P}$ d'individus, sur lesquels on observe un caractère $X \in \mathbb{R}$. On souhaite montrer formellement qu'un histogramme est un estimateur de la densité de X.

20 Montrer que [1]

$$f(x) = \lim_{h \to 0} \frac{1}{h} \mathbb{P}\left(x \leq X \leq x + h\right). \tag{1}$$

21 On considère maintenant un histogramme construit à partir d'un échantillon de réalisations $x_1, x_2, \ldots, x_n$ de X. Justifier l'emploi de l'histogramme comme estimation de la densité de X telle que définie par l'équation (1).

2.2 Méthode des noyaux et apprentissage

On reprend l'estimateur de densité par la méthode des noyaux, en utilisant le noyau gaussien :

$$\widehat{f}(x; h) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \quad K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right). \tag{2}$$

On veut apprendre le paramètre $h^\star$ optimal par la méthode du maximum de vraisemblance. Notons que cette stratégie n'est pas la seule possible, et qu'elle n'est pas la plus utilisée en pratique.

On considère pour cela un échantillon de valeurs y_j ($j = 1, \ldots, m$), *qui peut être distinct de l'échantillon de valeurs* x_i *(*$i = 1, \ldots, n$*)* [2] : l'échantillon des x_i est utilisé pour estimer la densité en tout point x via l'équation (2), et l'échantillon des y_j est utilisé pour choisir le paramètre h. En pratique, si on dispose d'un ensemble de données de taille suffisante, on peut sélectionner au hasard une partie des données pour constituer l'échantillon $x_1, \ldots, x_n$, et le reste pour l'échantillon $y_1, \ldots, y_m$.

22 Donner l'expression développée de la densité estimée en un point x quelconque. En déduire l'expression de la vraisemblance $L(h; y_1, \ldots, y_m)$.

23 Est-il raisonnable d'utiliser l'échantillon $x_1, \ldots, x_n$ pour estimer le paramètre h par maximum de vraisemblance (c'est-à-dire de le substituer à l'échantillon $y_1, \ldots, y_m$) ? Pourquoi ?

24 Calculer la log-vraisemblance, puis sa dérivée première.

25 Montrer qu'annuler la dérivée première revient à exprimer le paramètre h comme

$$h = \left(\frac{1}{m} \sum_{j=1}^{m} \frac{\sum_{i=1}^{n} w_{ji}(h)(y_j - x_i)^2}{\sum_{i=1}^{n} w_{ji}(h)} \right)^{1/2}, \tag{3}$$

en précisant l'expression des poids $w_{ji}(h)$ qui dépendent de h.

26 L'annulation de la log-vraisemblance ne permet pas de trouver une forme analytique d'un EMV. On propose donc de calculer une estimation $h^\star$ qui maximise $\ln L(h; \cdots)$ de manière numérique.

On remarquera que l'équation (3) est du type $h = g(h)$: tout point fixe $h^\star$ de g est donc un point critique de $\ln L(h; \cdots)$. On propose donc de choisir une valeur $h_{(0)}$, puis de calculer itérativement

$$h_{(q+1)} = \left(\frac{1}{m} \sum_{j=1}^{m} \frac{\sum_{i=1}^{n} w_{ji}(h_{(q)})(y_j - x_i)^2}{\sum_{i=1}^{n} w_{ji}(h_{(q)})} \right)^{1/2},$$

jusqu'à ce que $|h_{(q+1)} - h_{(q)}| < \varepsilon$ (avec ε une valeur choisie). On admettra que les conditions d'application du théorème du point fixe sont satisfaites, et que la procédure ci-dessus mène bien à un maximum de la log-vraisemblance.

On pourra utiliser la fonction `pairwise_distances` de la librairie `scikit-learn` pour calculer les distances euclidiennes (au carré) entre les points de `X` et `Y` :

1. On remarquera que la propriété (1) n'a pas tout-à-fait la même forme que dans le polycopié de cours.
2. Les valeurs x_i et y_j vivent dans le même espace, mais on utilise volontairement une notation spécifique pour les différencier les unes des autres, étant donné les rôles bien distincts des deux échantillons.

```
from sklearn.metrics import pairwise_distances as pedist
```

27 En pratique, on ne dispose souvent pas d'un échantillon $y_1, \ldots, y_m$ distinct pour calculer $h^\star$ (par exemple parce qu'on a trop peu de données disponibles). On « ré-utilise » donc l'échantillon $x_1, \ldots, x_n$ à la place de $y_1, \ldots, y_m$, mais *on prend soin d'écarter la valeur x_i de l'échantillon X lorsqu'on calcule la densité en x_i* : on obtient ainsi une *version leave-one-out* de la log-vraisemblance :

$$\ln L_{loo}(h; x_1, \ldots; x_m) = \sum_{i=1}^{n} \ln \widehat{f}_{loo}(x_i; h), \quad \widehat{f}_{loo}(x_i; h) = \frac{1}{n-1} \sum_{j \neq i} (2\pi)^{-1/2} h^{-1} \exp\left(-\frac{1}{2} \frac{(x_i - x_j)^2}{h^2}\right).$$

Modifier le code précédent pour implémenter la stratégie d'estimation par maximisation de la log-vraisemblance leave-one-out.