Cours 2 - Visualisation des données

Neila Mezghani

2 février 2021

COLLÈGE DE
BOIS-DE-BOULOGNE

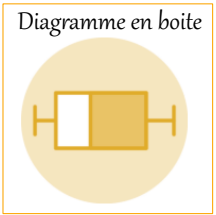
Plan du cours

Introduction

- Tout le monde connait l'expression « une image vaut mille mots » $\Longrightarrow$ L'objectif de ce cours
- Aujourd'hui'hui, à l'ère des données numériques des méga données, ...nous sommes inondées d'informations provenant de différentes modalités $\Longrightarrow$ ce vieux proverbe est devenu encore plus pertinent.
- La visualisation des données rend l'analyse beaucoup plus facile et plus rapide d'où le besoin de choisir les outils appropriés de visualisation

Table of Contents

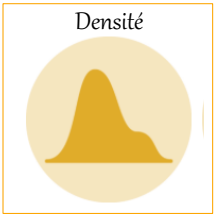
Visualisation de la distribution



BoxPlot

VilonPlot

Histogramme

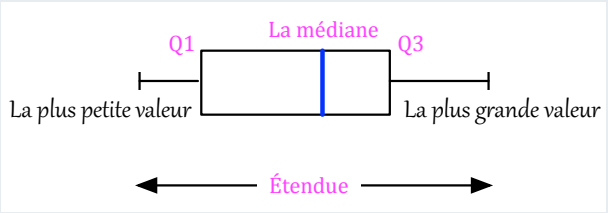


Densité

Diagramme en boite

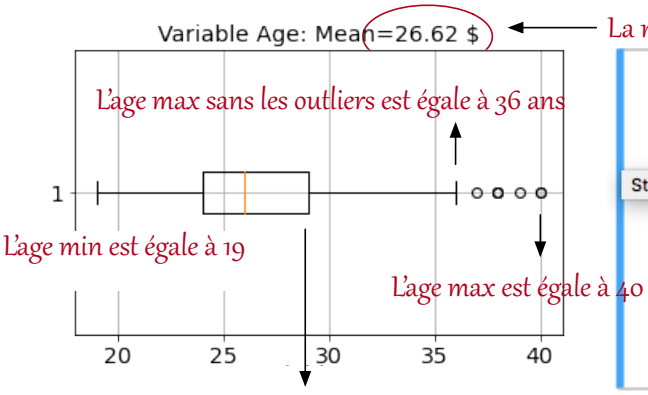
- Diagramme en boite (*boxplot*) est l'un des graphiques les plus couramment utilisés.
- Il résume cinq paramètres importants de la variable : la valeur minimale, le premier quartile (Q1), la médiane (ou deuxième quartile Q2), le troisième quartile (Q3) et la valeur maximale.
- La ligne qui divise la boîte en deux parties représente la médiane.

Diagramme en boite



- Les limites de la boite montrent les quartiles supérieur (Q3) et inférieur (Q1).
- Les lignes extrêmes montrent la valeur la plus élevée et la plus basse de la variable tout en excluant les valeurs aberrantes.

Exemple : Diagramme en boite



La moyenne d'âge arrondie

```
In [16]: df1['Age'].describe()

Out[16]: count    364.000000
         mean      26.615385
         std        4.233591
         min       19.000000
         25%       24.000000
         50%       26.000000
         75%       29.000000
         max       40.000000
         Name: Age, dtype: float64
```

StackAbuse

Q3 = 29 ce qui veut dire que l'âge de 75% des joueurs est <29 ans

Histogramme

- Un histogramme représente une estimation de la densité d'une variable quantitative.
- La forme de l'histogramme est obtenue suite à la répartition des données selon un ensemble d'intervalles. De ce fait celle-ci peut être différente selon le nombre d'intervalles défini.

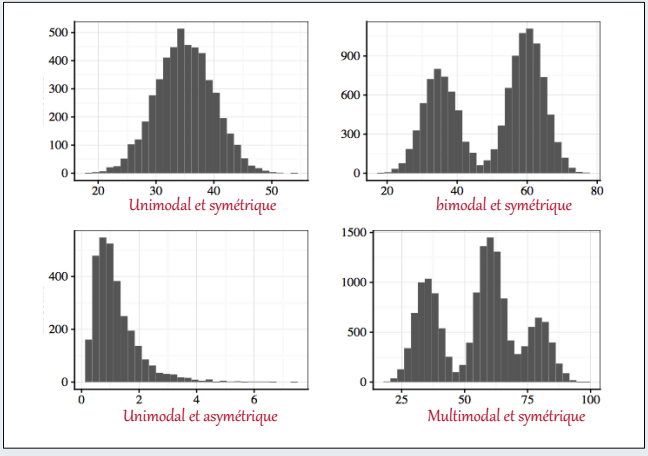
Histogramme : Principe de construction

- Identification de la valeur minimale *min* et maximale *max* de la variable à explorer
- Division de l'intervalle [*min*; *max*] en *I* sous-intervalles
- Dénombrement des observations pour lesquels la valeur de la variable tombe dans chacun des intervalles (appelés aussi classes)
- Représentation du nombre d'observations par intervalle par une barre dont la surface est proportionnelle aux décomptes.

Histogramme : Principe de construction

- Un histogramme est dit symétrique lorsque son profil à gauche est identique ou très similaire à son profil à droite autour d'un mode.
- Les modes d'un histogramme correspondent aux classes (intervalles) les plus abondantes localement.
- Un histogramme peut avoir un ou plusieurs modes.

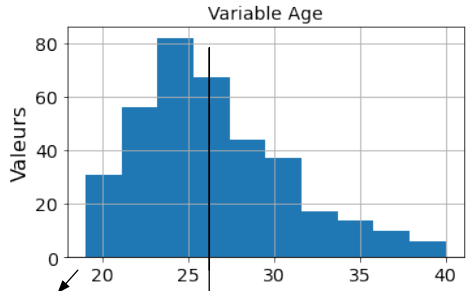
Histogramme



Exemple : Histogramme (1)

L'histogramme est unimodal et asymétrique

La médiane est < à la moyenne —> La majorité des joueurs ont un âge < à la moyenne d'âge

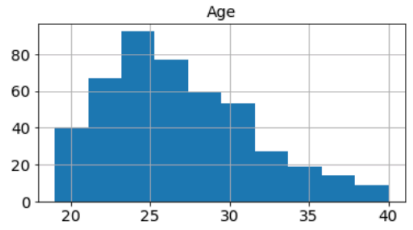



```
In [27]: df1['Age'].describe()

Out[27]: count    364.000000
         mean      26.615385
         std        4.233591
         min       19.000000
         25%       24.000000
         50%       26.000000
         75%       29.000000
         max       40.000000
         Name: Age, dtype: float64
```

Exemple : Histogramme (2)

In [23]:
```
hist,bin_edges = np.histogram(df['Age'].dropna(), bins=10)
display(hist)
display(bin_edges)
df.hist(column='Age', bins =10, figsize=(6,3))
None
```

```
array([40, 67, 92, 77, 59, 53, 27, 19, 14,  9])

array([19. , 21.1, 23.2, 25.3, 27.4, 29.5, 31.6, 33.7, 35.8, 37.9, 40. ])
```



Les valeurs des limites des intervalles

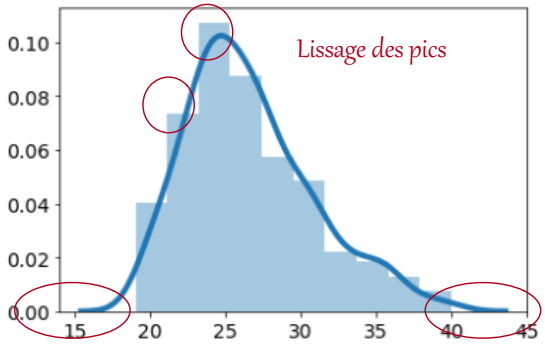
La médiane est < à la moyenne —> La majorité des observations ont un âge < à la moyenne d'âge

Densité

- La densité est aussi une représentation graphique de la distribution d'une variable numérique.
- Elle se présente comme un histogramme lissé.
- Le passage d'un histogramme à une courbe de densité consiste à lisser les pics plus ou moins fort dans l'histogramme. Ceci se fait souvent par des techniques d'estimation

Exemple : Densité

```
In [29]: sns.distplot(df1["Age"], hist=True, bins=10, kde=True,
                      kde_kws={'linewidth': 4});None
```

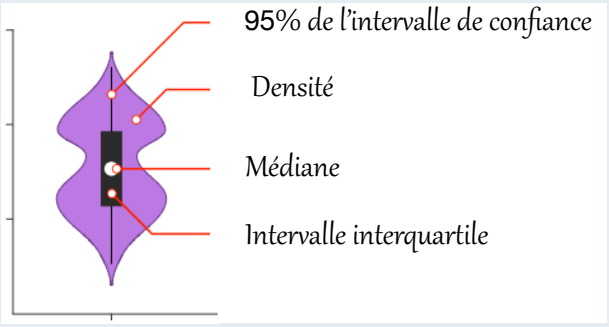


Extensions des limites minimales et maximales —> Lissage

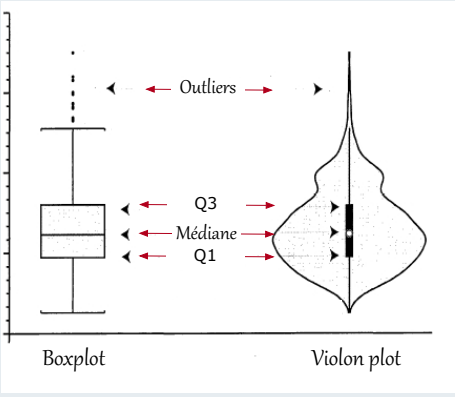
Diagramme en violon

- Le diagramme en violon (*violin plot*) permet de visualiser la distribution d'une variable numérique.
- Il est constitué en même temps de deux graphiques de densité en miroir et d'un boxplot $\Longrightarrow$ permet une compréhension plus approfondie de la densité.

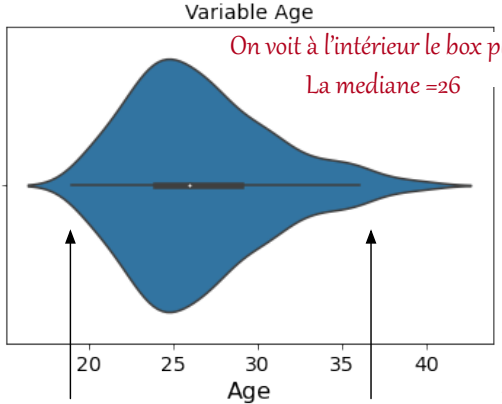
Diagramme en violon



Comparaison des diagrammes en boite et en violon



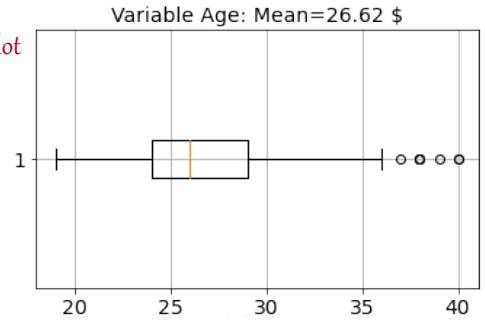
Exemple : Digramme en violon



On voit à l'intérieur le box plot

La mediane =26

Les valeurs minimales et maximales

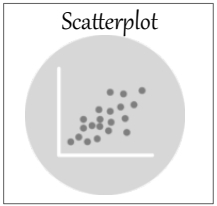


Le diagramme en violon ajoute de l'information sur la distribution de données

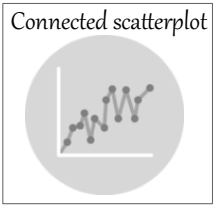
—> Unimodale et asymatrique

Table of Contents

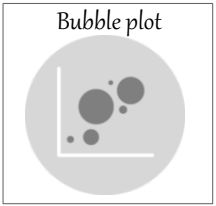
Visualisation des corrélation



ScatterPlot

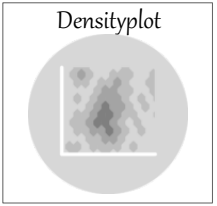


Connected

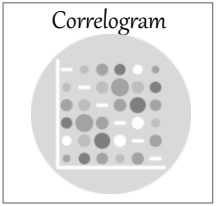


Bubble

HeatMap



DensityPlot



Correlogram

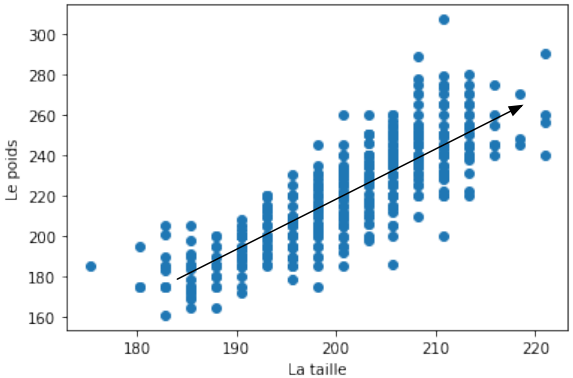
Diagramme de dispersion

- Un nuage de point (*Scatterplot*) permet de représenter une variable numérique en fonction d'une autre variable numérique :

$$Y \sim X$$

- Chaque point représente une observation.
- Les positions sur l'axe X (horizontal) et Y (vertical) représentent les valeurs des 2 variables.

Exemple : Digramme de dispersion



Le diagramme représente la variation de
Poids ~ Taille

Lorsque la taille augmente, le poids augmente

Le poids est une fonction linéaire de la taille

Diagramme de dispersion connecté

- Un nuage de points connectés (*Connected Scatterplot*) est très proche d'un nuage de points, sauf que les points sont reliés les uns aux autres par des lignes.
- Cela signifie que les valeurs des observations sur l'axe des X sont ordonnées pour que ce type de représentation soit utile.
- Les diagrammes de dispersions connectés sont souvent utilisés pour les séries chronologiques où l'axe X représente le temps.

Exemple : Diagramme de dispersion connecté (1)

```
flights = sns.load_dataset("flights")
flights.tail()
```

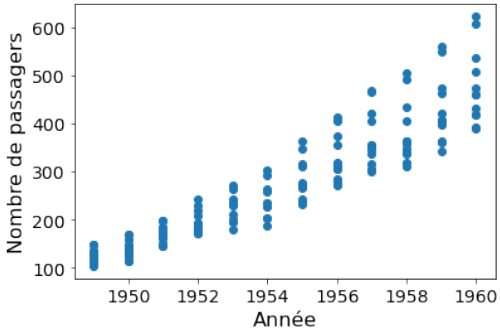
	year	month	passengers
139	1960	Aug	606
140	1960	Sep	508
141	1960	Oct	461
142	1960	Nov	390
143	1960	Dec	432

La base de données comprend 144 observations

Chacune est décrite par l'année, le mois et le nombre de passagers

Variation du nombre de passagers / année

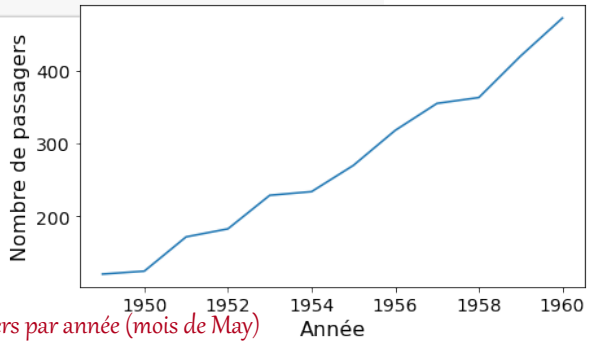
Chaque année comprend plusieurs mois



Exemple : Diagramme de dispersion connecté (2)

```
may_flights = flights.query("month == 'May'")
may_flights.head()
```

	year	month	passengers
4	1949	May	121
16	1950	May	125
28	1951	May	172
40	1952	May	183
52	1953	May	229



Variation du nombre de passagers par année (mois de May)

Diagramme de dispersion connecté possible parce que les valeurs de X sont ordonnées

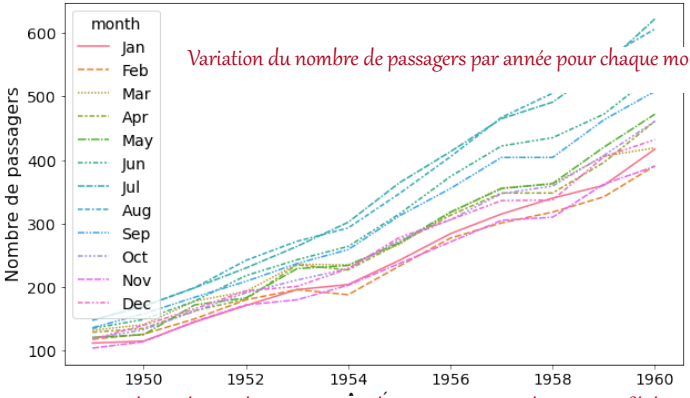
Exemple : Diagramme de dispersion connecté (3)

	year	month	passengers
0	1949	January	112
1	1949	February	118
2	1949	March	132
3	1949	April	129

Transformation de la matrice des données
Pour une meilleure exploration des données

month	January	February	March	April	May	June	July	August	September	October	November	December
year												
1949	112	118	132	129	121	135	148	148	136	119	104	118
1950	115	126	141	135	125	149	170	170	158	133	114	140
1951	145	150	178	163	172	178	199	199	184	162	146	166
1952	171	180	193	181	183	218	230	242	209	191	172	194
1953	196	196	236	235	229	243	264	272	237	211	180	201

Exemple : Diagramme de dispersion connecté (4)



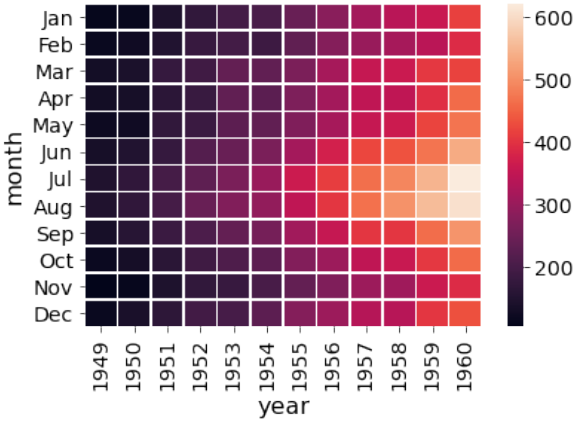
Variation du nombre de passagers par année pour chaque mois

On remarque que la tendance de croissance est la même pour tout les mois au fil des années

Carte de chaleur

- Une carte de chaleur (*heatmap*) est une représentation graphique des données où les valeurs individuelles contenues dans une matrice sont représentées sous forme de couleurs.
- Permet d'afficher une vue générale des données numériques.

Exemple : Carte de chaleur



Variation du nombre de passagers en fonction de l'année et du mois

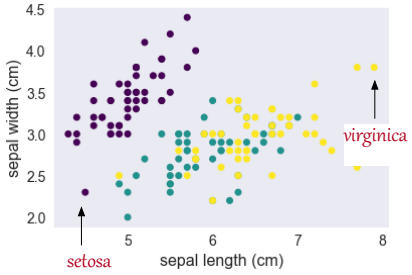
Le max de nombre de passagers est d'environ 600 pour les mois de juillet et aout 1960

Diagramme à bulles

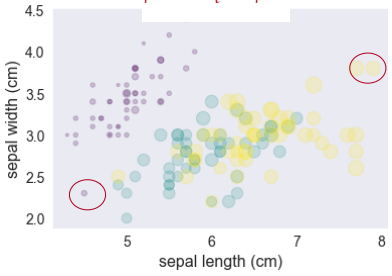
- Un diagramme à bulles (*bubbleplot*) est un nuage de points où d'autres dimensions sont ajouter pour avoir plus d'informations.
- Besoin de 3 variables numériques en entrée : une est représentée par l'axe X, une par l'axe Y, et une par la taille des bulles.
- La surface des bulles doit être proportionnelle à la valeur des données.

Exemple : Diagramme à bulles

largeur des sépales ~ longueur des sépales
pour chaque espèce



largeur des sépales ~ longueur des sépales
~ longueur des pétales
pour chaque espèce

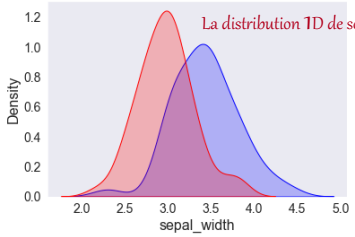


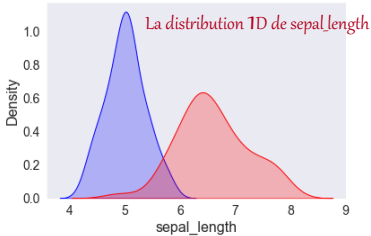
Id	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
42	4.5	2.3	1.3	0.3	setosa
132	7.9	3.8	6.4	2.0	virginica

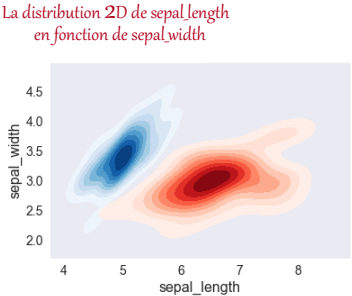
Diagramme de densité 2D

- Un diagramme de densité 2D ou histogramme 2D est une extension de l'histogramme (1D).
- Il montre la distribution des valeurs d'un ensemble de données sur la plage de deux variables quantitatives.

Exemple : Diagramme de densité 2D







Corrélogramme

- Un corrélogramme permet d'analyser la relation entre chaque paire de variables numériques d'une matrice.
- La corrélation entre chaque paire de variables est visualisée par un nuage de points, ou un symbole qui représente la corrélation (bulle, ligne, nombre..).
- La diagonale représente la distribution de chaque variable, en utilisant un histogramme ou un diagramme de densité.

Exemple : Diagramme de densité 2D

La corrélation entre chaque paire de variables par espèce à l'aide d'une droite de regréssion

La diagonale contient les distribution (ou l'histogramme)

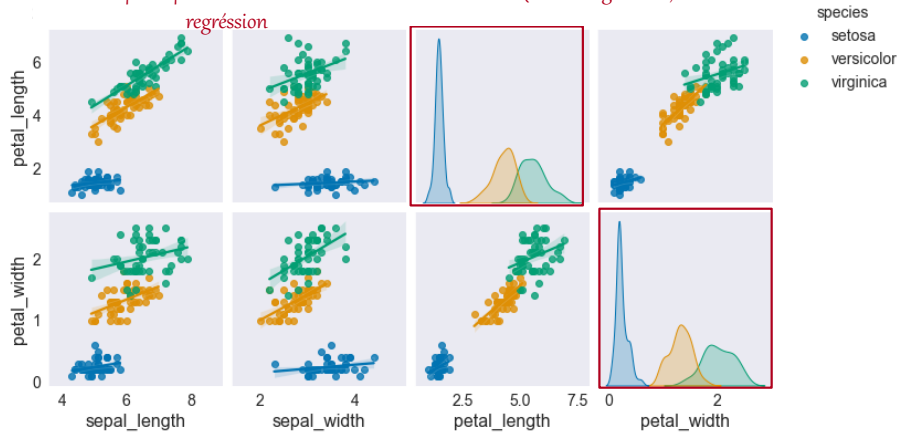
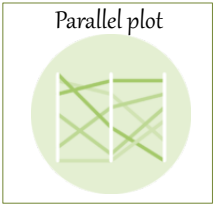


Table of Contents

Visualisation du rang

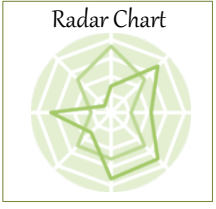
Bar plot



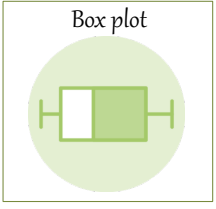
Parrallel plot

Lollipop plot

Word cloud



Radar Chart



Box Plot

Table of Contents

Table of Contents