

Cours 2 - Visualisation des données

Neila Mezghani

2 & 9 février 2021

Plan du cours

- 1 Visualisation de la distribution
- 2 Visualisation de la corrélation
- 3 Visualisation du rang
- 4 Représentation des proportions
- 5 Visualisation de l'évolution

Introduction

- Tout le monde connaît l'expression « une image vaut mille mots » \implies
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 \implies 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

Introduction à l'AED
(cours 1)
«Étapes de l'AED»
La collecte de données



Transformation
Analyse en composante principale (cours 5)
Analyse factorielle de correspondance (cours 6)
Analyse factorielle mixte (cours 7)



Visualisation et interprétation
(cours 2)

Connaissances

Données

1

2

3

4

5

6

Prétraitement

Préparation des données
(cours 3)
Statistiques uni- et
multidimensionnelles
(cours 4)



Analyse

Regroupement (cours 8)
Classification (cours 9)
Analyse d'association (cours 10)



Librairies à utiliser

- Outre la librairie Matplotlib, la librairie Seaborn est très utile et efficace pour la visualisation des données
- Les fonctions Seaborn ont presque toutes la même structure :

Sns.fonction (x,y,data,hue,size,style)

Options

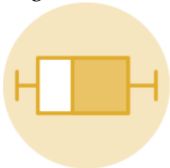
Les données à visualiser

Table of Contents

- 1 Visualisation de la distribution
- 2 Visualisation de la corrélation
- 3 Visualisation du rang
- 4 Représentation des proportions
- 5 Visualisation de l'évolution

Visualisation de la distribution

Diagramme en boîte



BoxPlot

Diagramme en violon



VilonPlot

Histogramme



Histogramme

Densité

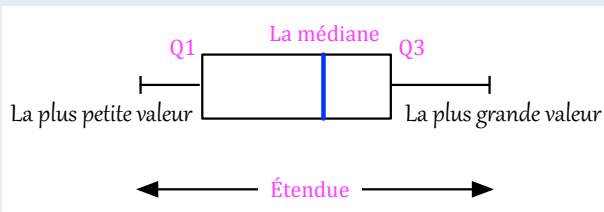


Densité

Diagramme en boîte

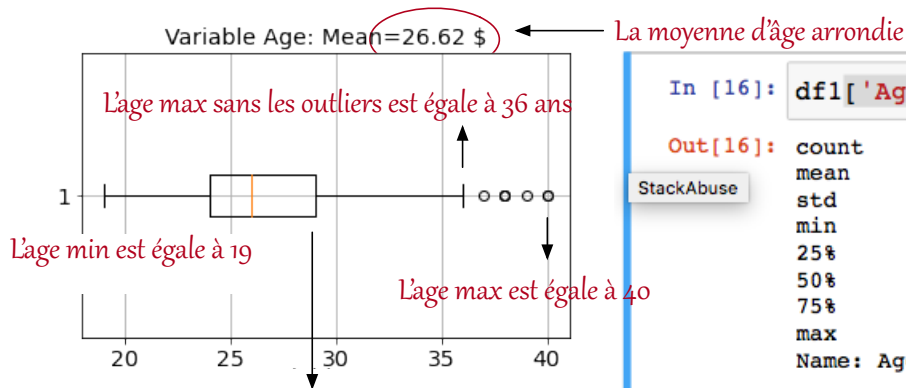
- Diagramme en boîte (*boxplot*) est l'un des graphiques les plus couramment utilisés.
- Il résume cinq paramètres importants de la variable : la valeur minimale, le premier quartile (Q1), la médiane (ou deuxième quartile Q2), le troisième quartile (Q3) et la valeur maximale.
- La ligne qui divise la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

Diagramme en boîte



- Les limites de la boîte 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.
- `seaborn.boxplot`

Exemple : Diagramme en boîte



```
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	

$Q_3 = 29$ 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.

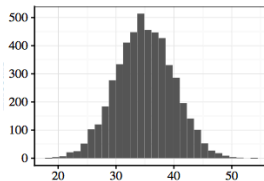
Histogramme : Principe de construction

- 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.

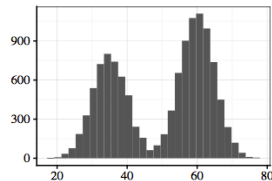
Histogramme : Principe de construction

- 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.
- `seaborn.histplot`

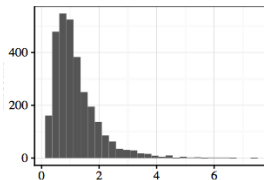
Histogramme



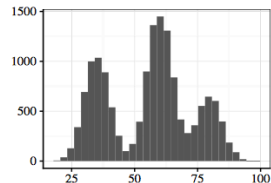
Unimodal et symétrique



bimodal et symétrique



Unimodal et asymétrique

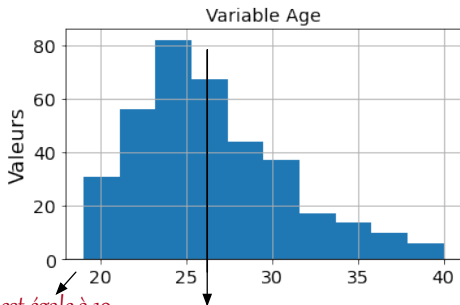


Multimodal et symétrique

Exemple : Histogramme (1)

L'histogramme est unimodal et asymétrique

La médiane est < à la moyenne \rightarrow 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
```

L'age min est égale à 19

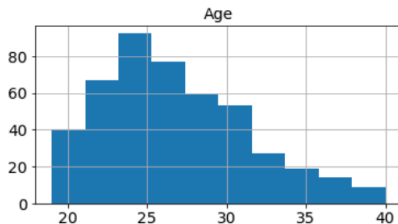
La moyenne d'âge = 26.61

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])
```

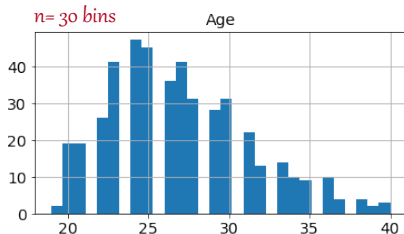
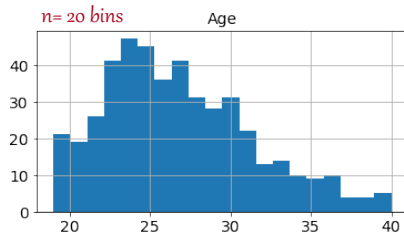
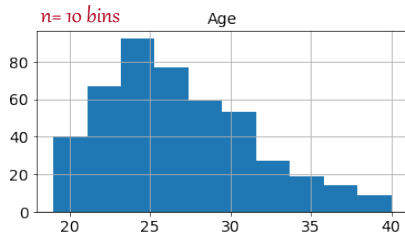
```
array([19. , 21.1, 23.2, 25.3, 27.4, 29.5, 31.6, 33.7, 35.8, 37.9, 40. ])
```



Les valeurs des limites des intervalles

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

Exemple : Histogramme (3)

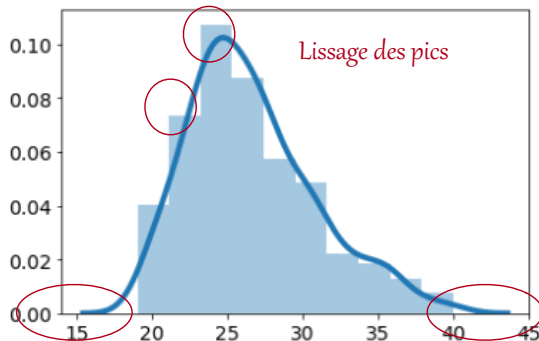


Densité

- 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
- `seaborn.kdeplot`

Exemple : Densité

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



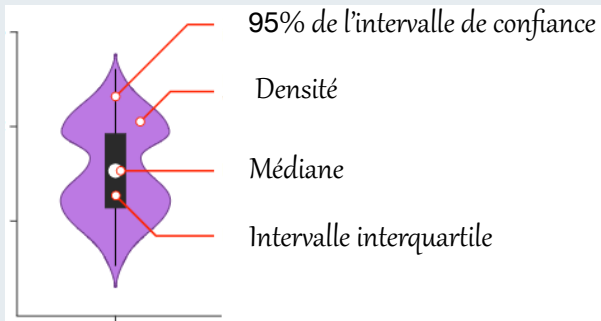
Lissage des pics

Extensions des limites minimales et maximales —> Lissage

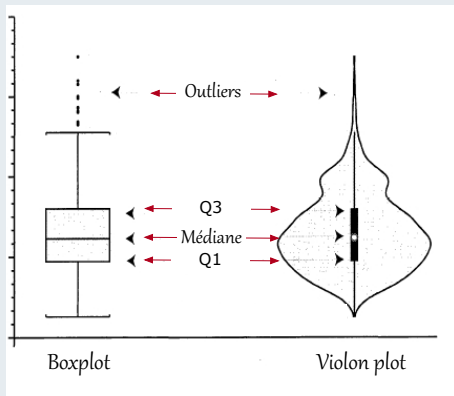
Diagramme en violon

- 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 \Rightarrow permet une compréhension plus approfondie de la densité.
- `seaborn.violinplot`

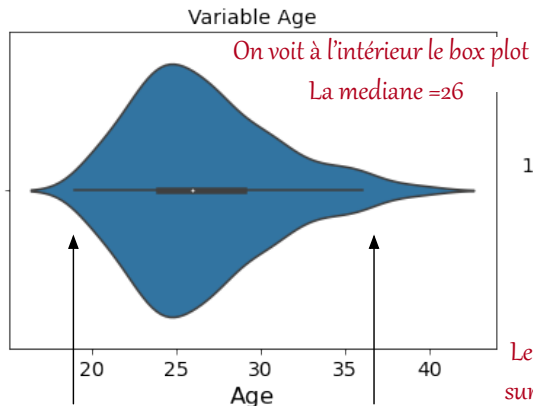
Diagramme en violon



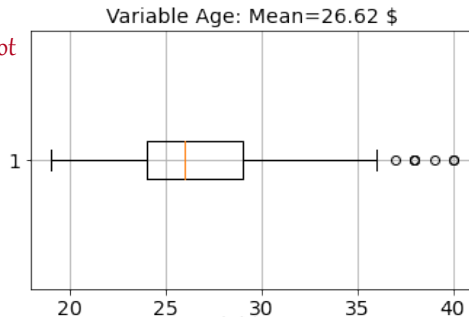
Comparaison des diagrammes en boîte et en violon



Exemple : Digramme en violon



Les valeurs minimales et maximales



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

—> Unimodale et asymétrique

Table of Contents

- 1 Visualisation de la distribution
- 2 Visualisation de la corrélation**
- 3 Visualisation du rang
- 4 Représentation des proportions
- 5 Visualisation de l'évolution

Visualisation de la corrélation

Scatterplot



ScatterPlot

Connected scatterplot



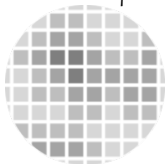
Connected

Bubble plot



Bubble

Heatmap



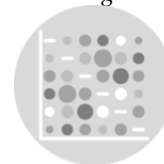
HeatMap

Densityplot



DensityPlot

Correlogram



Correlogram

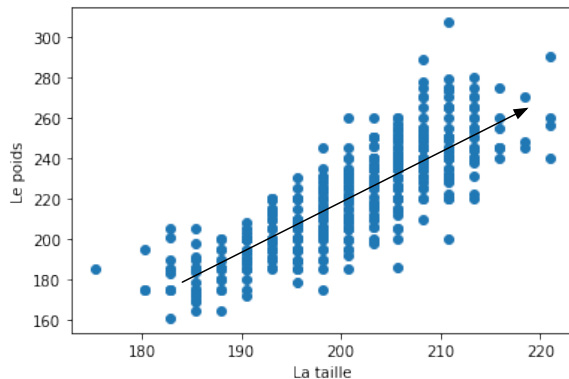
Diagramme de dispersion

- 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.
- `seaborn.scatterplot`

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.
- `seaborn.scatterplot`

Exemple : Diagramme de dispersion connecté (1)

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

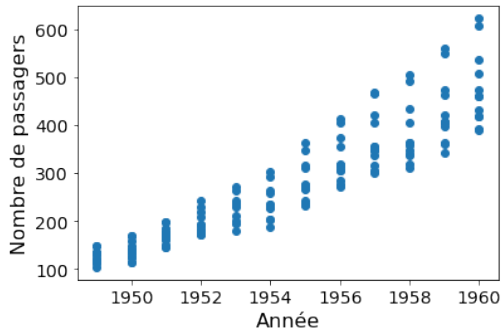
	year	month	passengers
139	1960	Aug	606
140	1960	Sep	508
141	1960	Oct	461
142	1960	Nov	390
143	1960	Dec	432

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

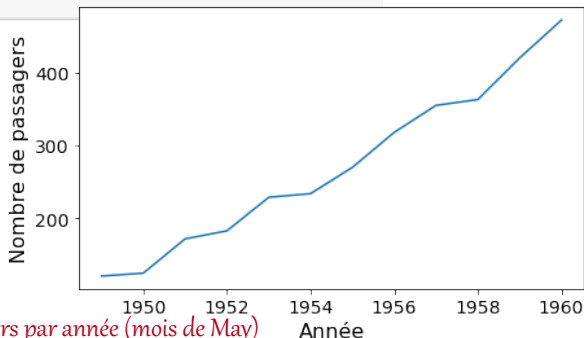
Chaque année comprend plusieurs mois



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

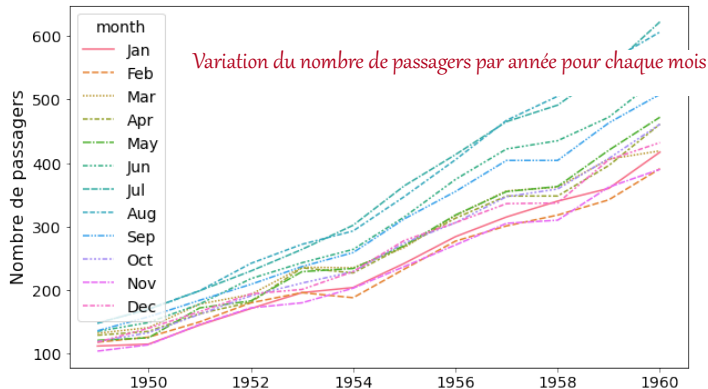
Exemple : Diagramme de dispersion connecté (3)

	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*

	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)

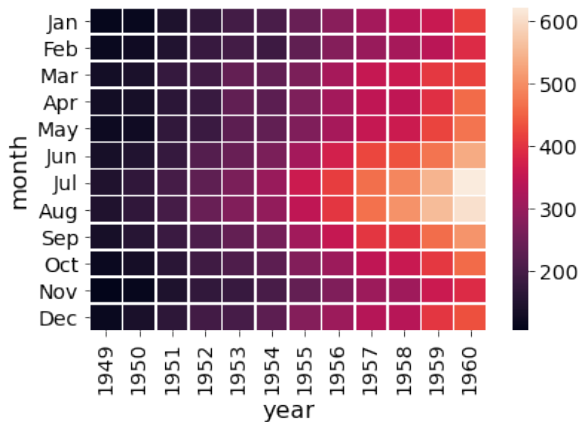


On remarque que la tendance de croissance est la même pour tout les mois au fil des années

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.
- `seaborn.heatmap`

Exemple : Carte de chaleur



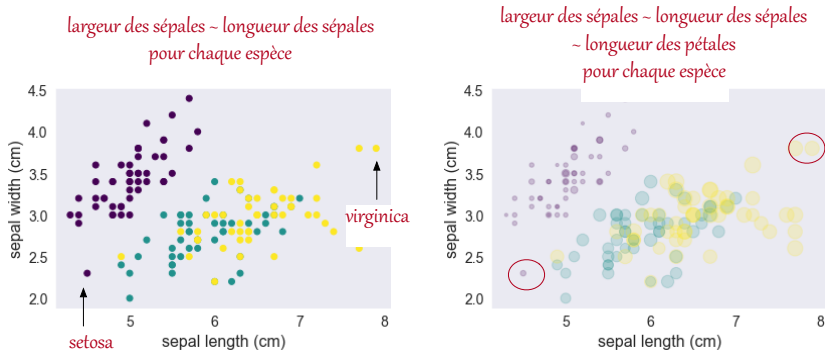
Variation du nombre de passagers en fonction de l'année et du mois

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

Diagramme à bulles

- Un diagramme à bulles (*bubbleplot*) est un nuage de points où d'autres dimensions sont ajoutées 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.
- `seaborn.scatterplot`

Exemple : Diagramme à bulles

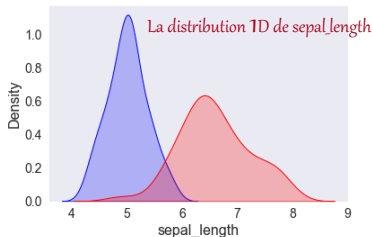
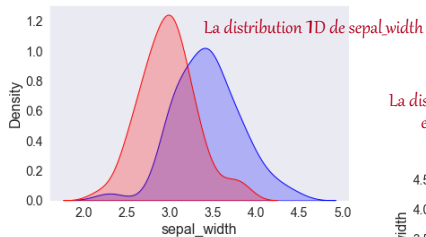


	Id	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
	42	4.5	2.3	1.3	0.3	setosa
	132	7.9	3.8	6.4	2.0	virginica

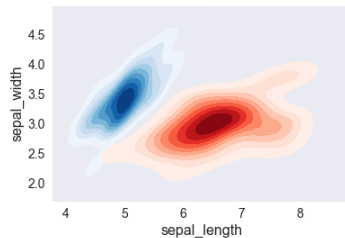
Diagramme de densité 2D

- 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.
- `seaborn.kdeplot`

Exemple : Diagramme de densité 2D



La distribution 2D de sepal_length
en fonction de sepal_width



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.
- `seaborn.pairplot`

Exemple : Corrélogramme

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

La diagonale contient les distributions (ou histogrammes)

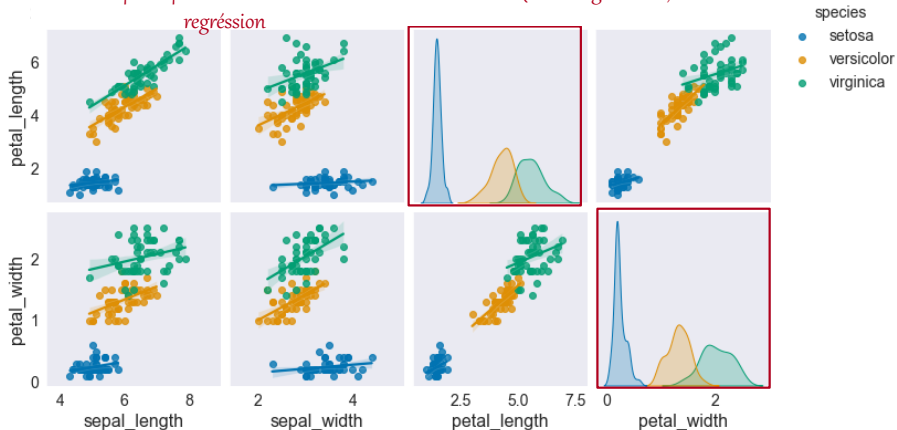


Table of Contents

- 1 Visualisation de la distribution
- 2 Visualisation de la corrélation
- 3 Visualisation du rang**
- 4 Représentation des proportions
- 5 Visualisation de l'évolution

Visualisation du rang

Bar plot



Bar plot

Parallel plot



Parallel plot

Lollipop plot



Lollipop plot

Wordcloud



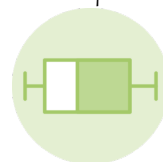
Word cloud

Radar Chart



Radar Chart

Box plot



Box Plot

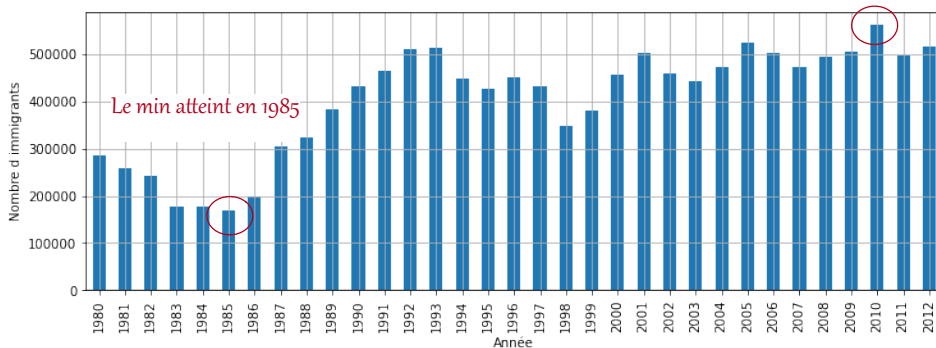
Diagramme en bâtons

- Le diagramme en bâtons ou graphique en barres permet de représenter les effectifs des différentes modalités d'une variable qualitative.
- Le diagramme en bâton peut être représenté horizontalement ou verticalement
- Dans certains cas, le diagramme en batons peut être aussi employé pour résumer des données numériques via la moyenne.
- `seaborn.barplot`

Exemple : Diagramme en bâtons

Diagramme de variation de l'effectif du nombre d'immigrants entre 1980 et 2012

Le max atteint en 2010



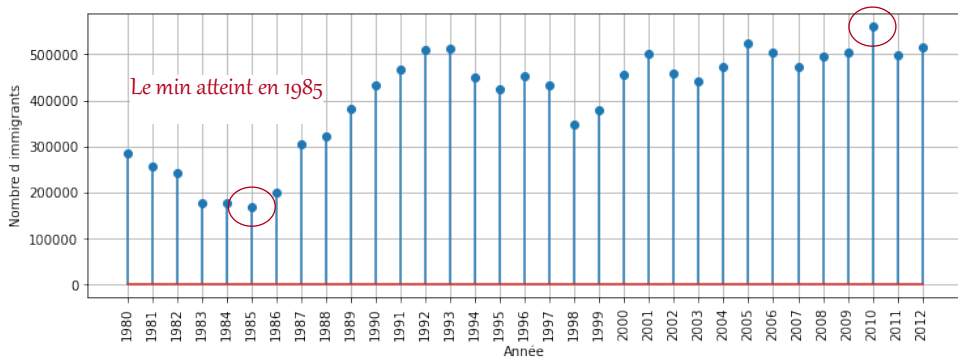
le graphique « Lollipop »

- Le graphique « Lollipop » (en forme de bâtons de sucette) est un diagramme de plus en plus attrayant grâce à sa simplicité
- Comme le diagramme en bâton, il permet de représenter les effectifs des différentes modalités d'une variable qualitative.
- `matplotlib.pyplot.stem`

Exemple : Graphique « Lollipop »

Diagramme de variation de l'effectif du nombre d'immigrants entre 1980 et 2012

Le max atteint en 2010

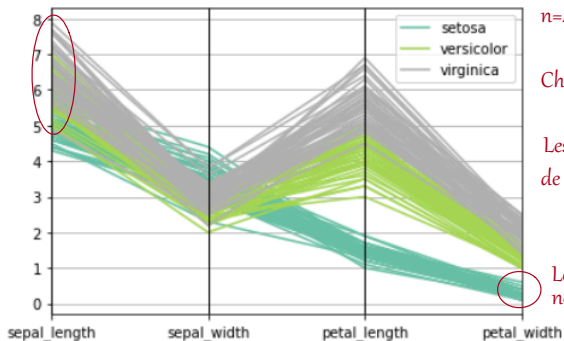


Les coordonnées parallèles

- Coordonnées parallèles permettent de représenter une observation de dimension n par n axes parallèles sur un plan, chaque axe représentant une dimension.
- L'intérêt des coordonnées parallèles se trouve dans le fait que certaines propriétés géométriques des variables se transforment et s'interprètent facilement en 2D.
- `pandas.plotting.parallel_coordinates`

Exemple : Les coordonnées parallèles

```
In [2]: data = sns.load_dataset('iris')  
parallel_coordinates(data, 'species', colormap=plt.get_cmap("Set2"))  
plt.show()
```



$n=4$: les 4 features

Chaque ligne est une observation

Les extrémités correspondent à la valeurs
de deux variables successives

Les valeurs de la variable largeur des pétales
ne sont pas très différentes pour l'espèce Setosa

Nuage de mots-clés

- Le Nuage de mots-clés est une représentation visuelle des mots-clés les plus utilisés dans un texte
- Les mots-clés s'affichent dans des tailles et épaisseurs de caractères proportionnelles à leurs visibilitées
- [wordcloud](#)

Exemple : Nuage de mots-clés



Diagramme en radar

- Plusieurs appellations : le diagramme en radar = Diagramme de Kiviat = diagramme en étoile = toile d'araignée
- Un diagramme bidimensionnelle conçue pour tracer une ou plusieurs séries de valeurs sur plusieurs variables quantitatives.
- Formé par autant d'axes que de variables. Les valeurs (ou des séries) sont affichées à l'intérieur de la toile
- Les valeurs doivent être normalisées pour qu'on puisse les superposer

Exemple : Diagramme en radar

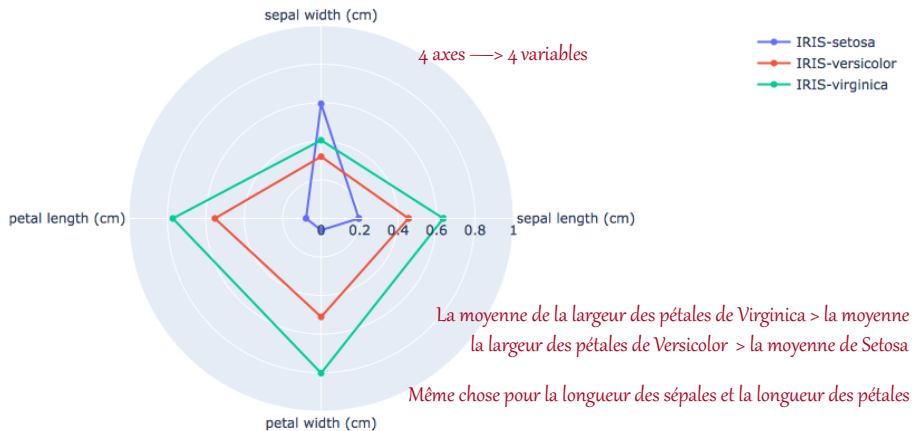


Table of Contents

- 1 Visualisation de la distribution
- 2 Visualisation de la corrélation
- 3 Visualisation du rang
- 4 Représentation des proportions**
- 5 Visualisation de l'évolution

Représentation des proportions

Stacked Barplot

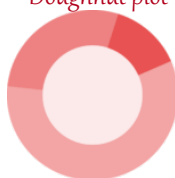


Stacked Bar

Pie Plot



Doughnut plot

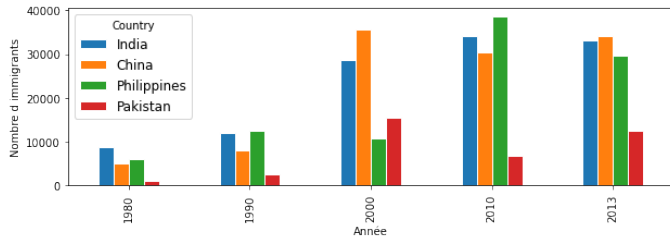


Pie plot

Graphique à barres groupées ou empilées

- Les diagrammes à barres groupées ou empilées affichent la taille relative (sous la forme d'un nombre total, d'un pourcentage, ou d'une autre métrique) d'une variable catégorielle, subdivisée par couleur en fonction d'un sous-groupe.
- `matplotlib.pyplot.bar`

Exemple : Graphique à barres groupées ou empilées



Graphique du nombre d'immigrants
dont le pays d'origine est l'un de ces 4 pays

Permet de voir comment le nombre
d'immigrants est distribué selon
le pays d'origine

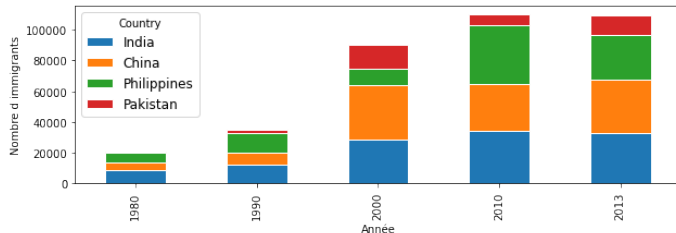


Diagramme à tarte/Diagramme à anneau

- Diagramme circulaire = Camembert = Diagramme à tarte
- Permettent de présenter les proportions de données qualitatives (catégorielles)
- La taille de chaque partie représentant la proportion de chaque catégorie.
- Le diagramme en anneau (Doughnut) a le même le principe que le diagramme à tarte
- `matplotlib.pyplot.pie`

Exemple : Diagramme à tarte

Distribution (de la proportion) des 4 premiers pays de provenances des immigrants au Canada

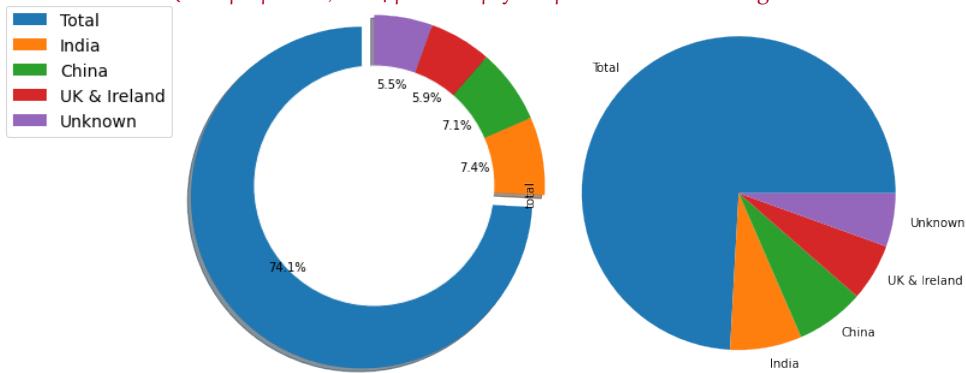
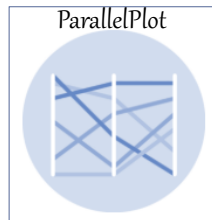
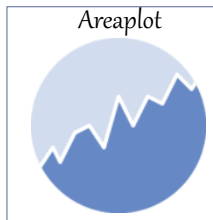
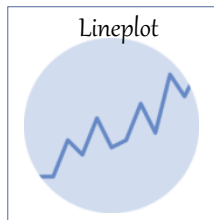


Table of Contents

- 1 Visualisation de la distribution
- 2 Visualisation de la corrélation
- 3 Visualisation du rang
- 4 Représentation des proportions
- 5 Visualisation de l'évolution**

Visualisation de l'évolution



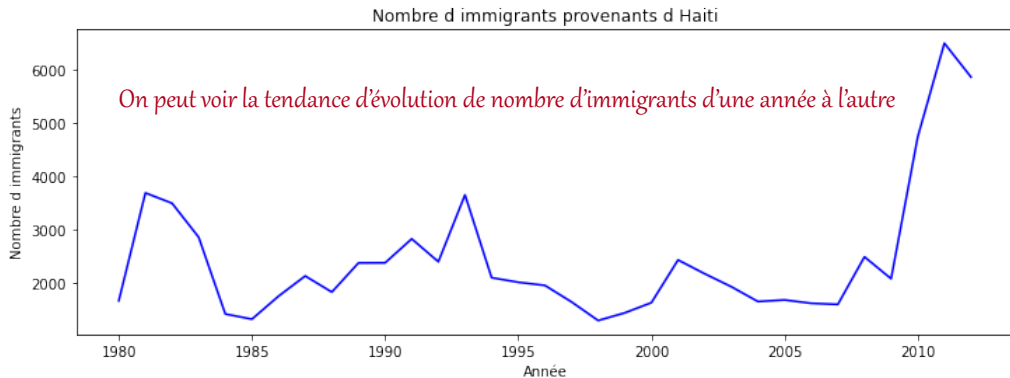
Diagrammes linéaires

- Les diagrammes linéaires présentent les informations sous forme d'une série de points de données reliés par des lignes droites.
- Les catégories apparaissent le long de l'axe des x et les statistiques le long de l'axe des y.
- Les diagrammes chronologiques peuvent être considérés comme un cas particulier de diagrammes linéaires.
- `matplotlib.pyplot.plot`

Diagramme chronologique

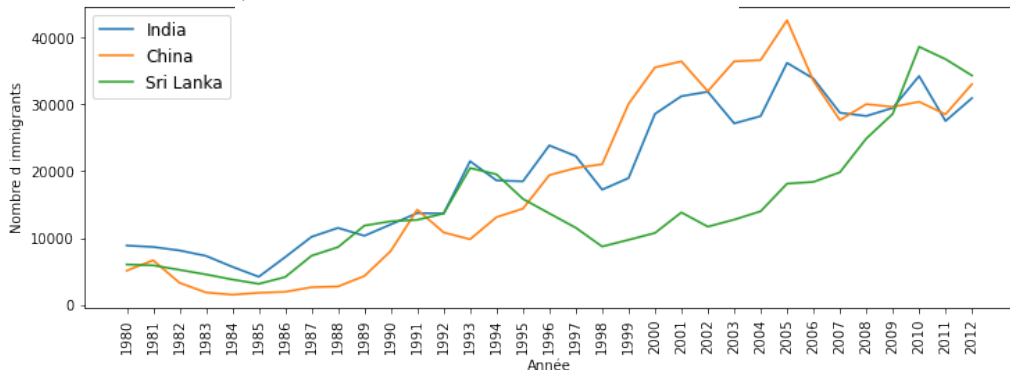
- Un diagramme chronologique permet de visualiser des tendances des totaux ou valeurs numériques dans le temps.
- Les diagrammes chronologiques permettent d'extraire des connaissances sur la tendance d'évolution d'une variable dans le temps.
- `matplotlib.pyplot.plot`

Exemple : Diagramme chronologique



Exemple : Diagrammes chronologiques superposés

On peut voir la tendance d'évolution de nombre d'immigrants d'une année à l'autre pour les trois provenances suivantes : l'Inde, la Chine, le Sri Lanka

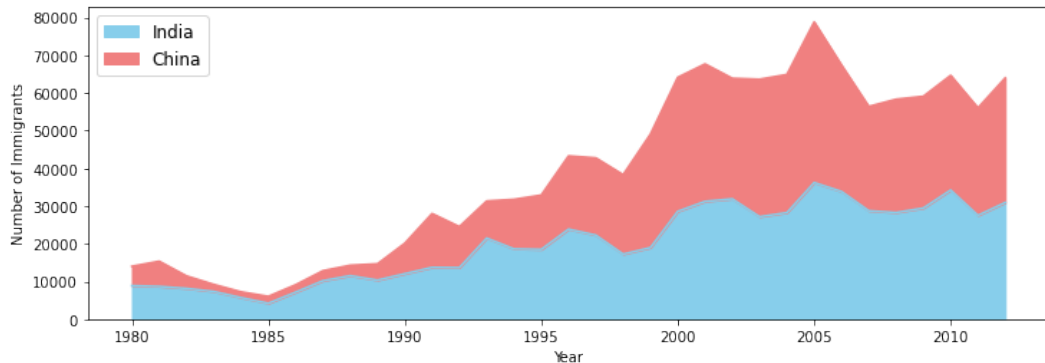


Graphique en aires

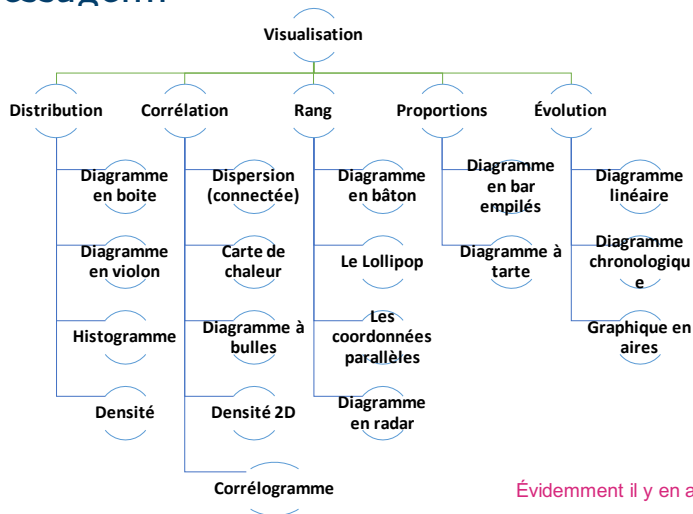
- Le graphique en aires représente l'évolution d'une variable quantitative en fonction d'une autre variable quantitative.
- La zone entre l'axe et la ligne est généralement soulignée par des couleurs, des textures ou des hachures.
- Généralement avec un graphique en aires, on compare au moins deux quantités.
- `matplotlib.pyplot.plot`

Exemple : Graphique en aires

Graphique représentant une comparaison de la variation du nombre d'immigrants provenant de l'inde par rapport au nombre d'immigrants provenant de la chine



Take-home message....



Évidemment il y en a d'autres