USTHB Faculté d'Informatique Département d'IA et SD 05 Mars 2023
Laiche Nacéra Master 1, Informatique Visuelle MIV)

Analyse des données

TP 1

Exercice 1

Considérons la matrice des données X suivante, résultat de 4 observations sur 10 individus.

J / I	Moyenne	Age	Taille	Poids
Enfant 1	14	13	1.50	45
Enfant 2	16	13	1.60	50
Enfant 3	15	13	1.65	50
Enfant 4	9	15	1.75	60
Enfant 5	10	14	1.70	60
Enfant 6	7	14	1.70	60
Enfant 7	8	14	1.60	70
Enfant 8	13	13	1.60	65
Enfant 9	17	15	1.55	60
Enfant 10	11	14	1.70	65

Alors :

$$X(J) =$$

	Moyenne arithmétique	Ecart-type	Variances
Moyenne	**12**	**3.316**	**11**
Age	**13.8**	**0.748**	**0.56**
Taille	**1.635**	**0.074**	**0.005**
Poids	**58.5**	**7.433**	**55.25**

1. Calculez le centre de gravité du nuage des données. Afficher le résultat.

$$g = (12, 13.8, 1.635, 58.5).$$

2. Calculer toutes les covariances possibles entre les variables données. Afficher les résultats obtenus dans une matrice notée V.

$$V =$$

	Moyenne	Age	Taille	Poids
Moyenne	11	- 0.8	-0.155	-14
Age		0.56	0.748	2.7
Taille			0.005	0.202
Poids				55.25

Analyse des données

3. Calculer les coefficients de corrélations des couples : (Moyenne, Taille), (Age, Poids). Que remarquez-vous ? Commenter les résultats obtenus.

R =

	Moyenne	**Age**	**Taille**	**Poids**
Moyenne	**1**		**-0.631**	
Age		**1**		**0.48**
Taille			**1**	
Poids				**1**

4. Déterminer les valeurs propres de la matrice V ainsi que les vecteurs propres associés aux valeurs propres déterminées.

Détermination des valeurs propres de A

>>> **np.linalg.eigvals** (A)

Array ([,……,])

Détermination des valeurs propres et vecteurs propres de X

>>> **valp, vecp = np.linalg.eig** (A)
>>> valp
Array ([…….])

>>> vecp
array ([[ ],
[ ],
[ ]])

Remarque

Les colonnes de la matrice obtenue sont les vecteurs propres.

>>> B = np.transpose (vecp)

Ou bien

>>> B = vecp.T

B

5. Proposer une instruction qui vous permettra de vérifier si un vecteur donné Y est le vecteur propre associé à une valeur propre λ d'une matrice A. Appliquer à ce qui précède (aux résultats de la question 9).

Une valeur propre u d'une matrice A vérifie : il existe un vecteur propre v tel que :

A*v = u*v.

Analyse des données

Le calcul de valeurs propres donne les valeurs suivantes triées dans un ordre décroissant :

$$\lambda_1 = 2.391, \quad \lambda_2 = 0.750, \quad \lambda_3 = 0.584, \quad \lambda_4 = 0.274.$$

Les vecteurs propres associés sont :

$$u_1 = (0.508, \quad 0.503, \quad 0.445, -0.538),$$

$$u_2 = (0.306, \quad -0.464, \quad 0.705, \quad 0.438),$$

.

.

.

Pour le graphisme

6. Représenter graphiquement le **nuage des individus** dans le plan des couples :

   (Moyenne, Taille), et (Age, Poids).

   Présenter les deux graphes dans une même fenêtre. C'est-à-dire **partitionner** la fenêtre de visualisation en **deux sous fenêtres** (horizontales ou verticales à vous de choisir). Dans la première sous fenêtre, présenter le graphe des individus dans le premier plan et le deuxième graphe dans la deuxième sous fenêtre. N'oublier pas de **nommer** (légender) les axes pour les deux graphes et de donner un **titre** pour chaque graphe.

7. Interprétez les graphes obtenus.

Pour les mesures de liaison

1) *Mesure de proximités entre les individus en utilisant le mètre comme unité :*

   - **D (Enfant 4, Enfant 5)** = (2.0025)**(0.5) = **1.415**
   - **D (Enfant 4, Enfant 6)** = 5**(0.5) = **2.236**
   - **D (Enfant 5, Enfant 6)** = 9**(0.5) = **3.**

   Nous en déduisons que l'individu 6 est plus proche (ressemblance) de l'individu 4

2) Mesure de proximités entre les individus en utilisant le *centi*mètre comme unité :

   - **D (Enfant 4, Enfant 5)** = (27)**(0.5) = **5.196**
   - **D (Enfant 4, Enfant 6)** = (30)**(0.5) = **5.477**
   - **D (Enfant 5, Enfant 6)** = (9)**(0.5) = **3**

   Nous en déduisons que l'individu 6 est plus proche (ressemblance) de l'individu *5*

Analyse des données

3) ***C'est clair que les résultats obtenus sont contradictoires et nous ne pouvons pas en prendre de décision.***

Conclusion

- **L'impact de l'unité des données sur l'analyse des résultats obtenus.**
- **La mesure de proximité dépend d'une manière directe de l'unité des données.**

Solution

D'où la nécessité de Centrer et de réduire les données.

4) **Matrice centrée réduite des données**

Pour tout $j = 1, 2, \ldots, n$, nous avons :

$$\mathbf{X^j} = (\mathbf{x_1^j}, \mathbf{x_2^j}, \ldots, \mathbf{x_m^j})^{\mathbf{t}} \in \Re^{\mathrm{m}}.$$

C'est-à-dire :

$$\mathbf{X^j} = \begin{pmatrix} \mathbf{x_1^j} \\ \mathbf{x_2^j} \\ \vdots \\ \mathbf{x_m^j} \end{pmatrix}.$$

Donc, la variable centrée - réduite est définie par :

$$\mathbf{Z^j} = \begin{pmatrix} \dfrac{\mathbf{x_1^j} - \bar{\mathbf{X}}^{\mathbf{j}}}{\boldsymbol{\sigma}_{\mathbf{X^j}}} \\ \dfrac{\mathbf{x_2^j} - \bar{\mathbf{X}}^{\mathbf{j}}}{\boldsymbol{\sigma}_{\mathbf{X^j}}} \\ \vdots \\ \dfrac{\mathbf{x_m^j} - \bar{\mathbf{X}}^{\mathbf{j}}}{\boldsymbol{\sigma}_{\mathbf{X^j}}} \end{pmatrix}.$$

Analyse des données

Dans notre cas : **m = 10** *et* **n = 4.**

J(C-R) / I	Moyenne	Age	Taille	Poids
Enfant 1	$\frac{14-12}{3.316}$	$\frac{13-13.8}{0.748}$	$\frac{1.50-1.635}{0.074}$	$\frac{45-58.5}{7.433}$
Enfant 2	$\frac{16-12}{3.316}$	$\frac{13-13.8}{0.748}$	$\frac{1.60-1.635}{0.074}$	$\frac{50-58.5}{7.433}$
Enfant 3	$\frac{15-12}{3.316}$	$\frac{13-13.8}{0.748}$	$\frac{1.65-1.635}{0.074}$	$\frac{50-58.5}{7.433}$
Enfant 4	$\frac{9-12}{3.316}$	$\frac{15-13.8}{0.748}$	$\frac{1.75-1.635}{0.074}$	$\frac{60-58.5}{7.433}$
Enfant 5	$\frac{10-12}{3.316}$	$\frac{14-13.8}{0.748}$	$\frac{1.70-1.635}{0.074}$	$\frac{60-58.5}{7.433}$
Enfant 6	$\frac{7-12}{3.316}$	$\frac{14-13.8}{0.748}$	$\frac{1.70-1.635}{0.074}$	$\frac{60-58.5}{7.433}$
Enfant 7	$\frac{8-12}{3.316}$	$\frac{14-13.8}{0.748}$	$\frac{1.60-1.635}{0.074}$	$\frac{70-58.5}{7.433}$
Enfant 8	$\frac{13-12}{3.316}$	$\frac{13-13.8}{0.748}$	$\frac{1.60-1.635}{0.074}$	$\frac{65-58.5}{7.433}$
Enfant 9	$\frac{17-12}{3.316}$	$\frac{15-13.8}{0.748}$	$\frac{1.55-1.635}{0.074}$	$\frac{60-58.5}{7.433}$
Enfant 10	$\frac{11-12}{3.316}$	$\frac{14-13.8}{0.748}$	$\frac{1.70-1.635}{0.074}$	$\frac{65-58.5}{7.433}$

Question
Recalculer les mesures ci-dessus.

Pour le graphisme,

On va voir deux types de présentation

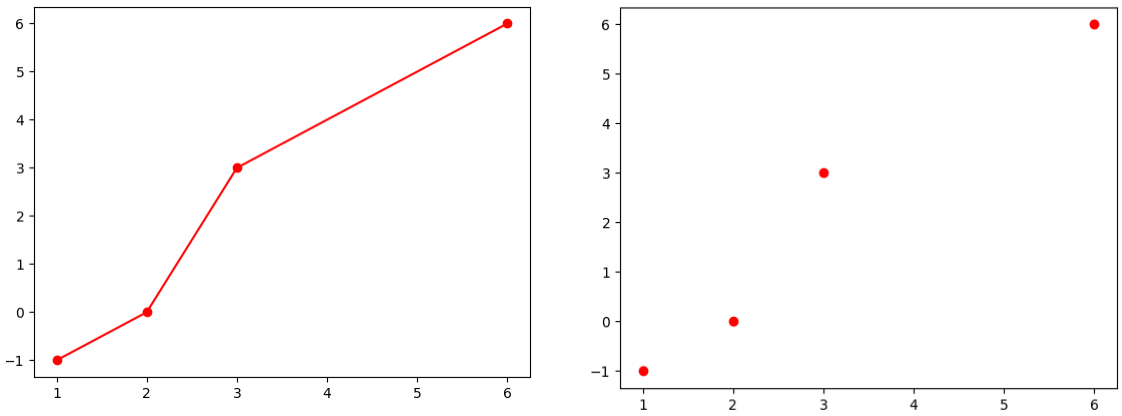
```
>>> plt.plot ([1,2,3,6],[-1,0,3,6], linestyle = 'none', marker = '+', c = 'y')
```

Ou bien

Scatter plot = Nuage de points (points de dispersion)

```
>>> plt.scatter (x, y, c = 'r', marker = 'o')
```

Analyse des données



Présentation de plusieurs courbes (graphes) dans la même fenêtre

Pour ce faire, nous utilisons *l'instruction* : **plt.subplot (L C i)**

L *:* le nombre de lignes, **C** : le nombre de colonnes, **i** : la position du graphe.

Dans notre cas :

2 lignes, *1*colonnes et *2* graphes.

Ou bien

***1* ligne, *2* colonnes et *2* graphes.**

>>> plt.subplot (*211*)

>>> plt.subplot (*212*)

Sinon

>>> plt.subplot (*121*)

>>> plt.subplot (*122*)

Exercice 2

1) Ecrire une fonction qui calcule la moyenne arithmétique des 8 variables de la matrice donnée. Prenez 6 chiffres décimaux.

>>> stat.mean(X.T[i]) for i = 0,1, 2, ..., 9

>>> round(stat.mean(X.T[i]), 6)

9.0 9.0 9.0 9.0 7.500909 7.500909 7.5 7.5009009

Analyse des données

2) g = [les moyennes]

g = [9.000000, 9.000000, 9.000000, 9.000000, 7.500909, 7.500909, 7.500000, 7.500909]

3) Ecrire une fonction qui calcule la variance des 8 variables de la matrice donnée. Prenez 6 chiffres décimaux.

Var (xi) = (1/11)*somme des carrés (xij-mi). Pour tout i = 0 a 7.

Ou bien

Exploiter les variances prédéfinies et les corriger.

```
>>> stat.variance(X.T[i])*(10/11)
```

Variances = vecteur

Variances [10., 10., 10., 10., 3.752062, 3.752390, 3.747836, 3.748408]

4) Calculer les coefficients de corrélation des couples de variables suivantes :

$$(X^1, X^5), \quad (X^2, X^6), \quad (X^3, X^7), \quad (X^4, X^8).$$

Corrélation de (x1, x5) = Correlation (X.T[0], X.T[4]) = cov(,)/les écarts types

Les corrélations :

0.817756 0.8162365 0.8162867 0.8165214

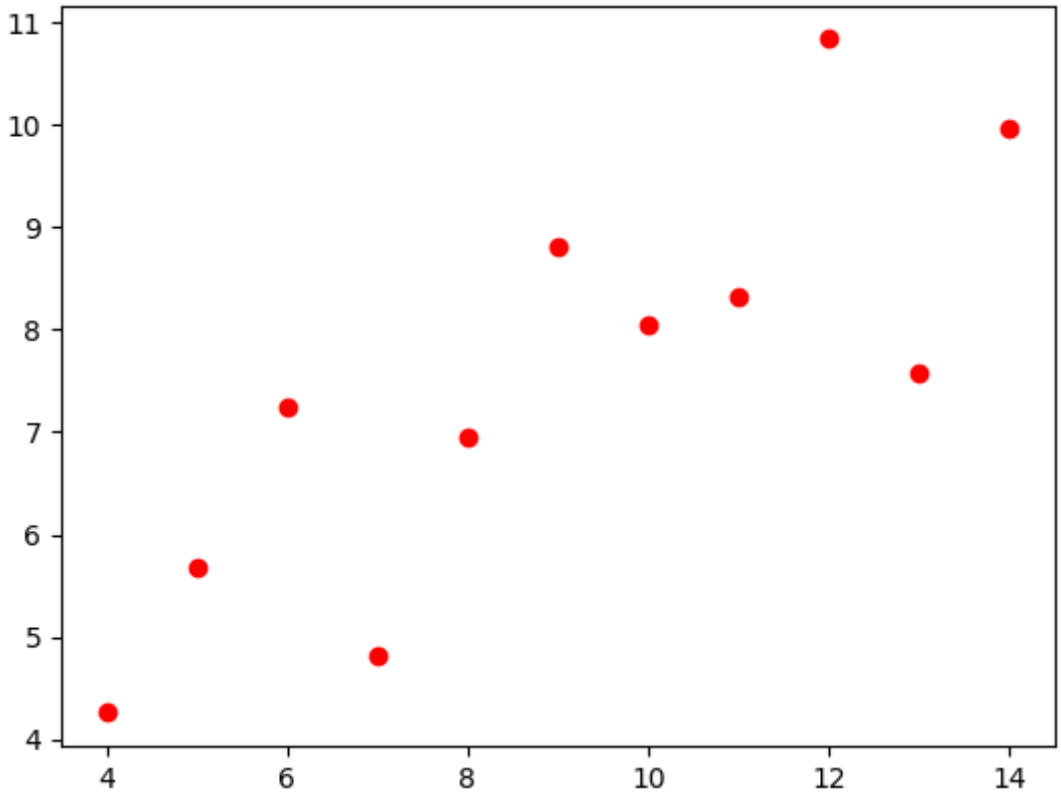
Nous remarquons que :

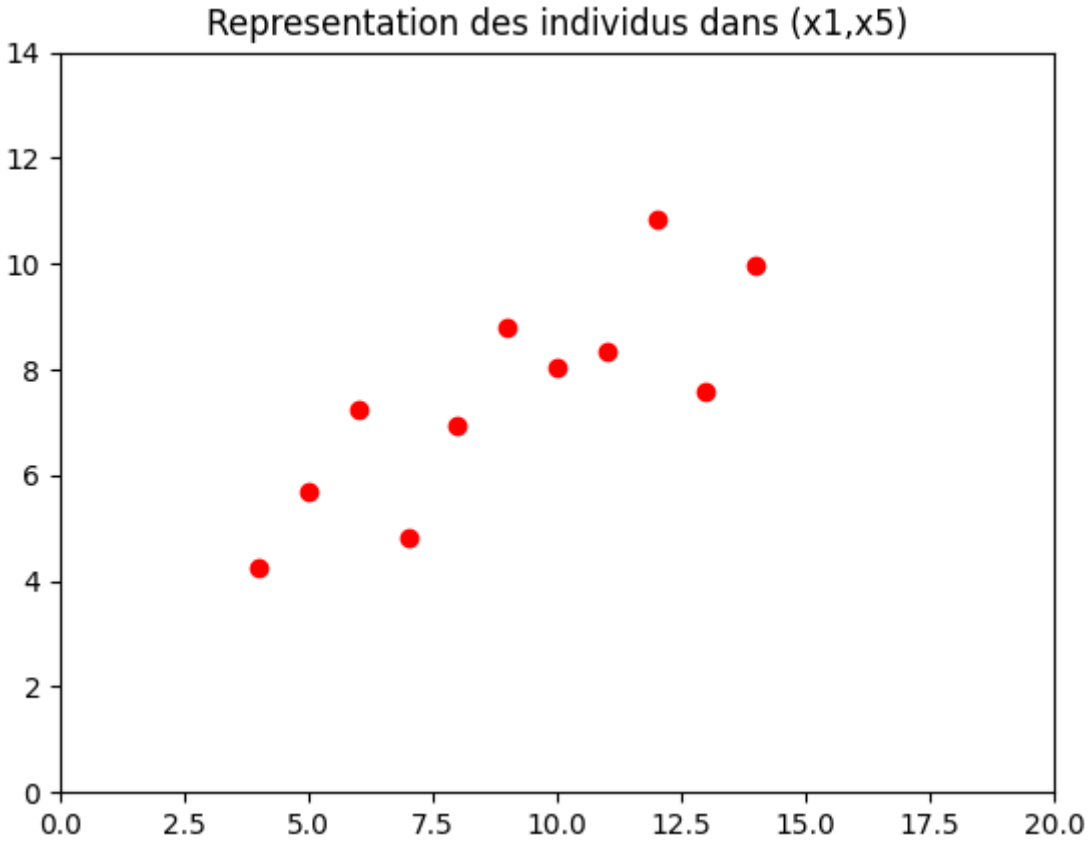
Les couples de variables ont la même corrélation.

```
>>> plt.scatter (x1, x5, c= 'r', marker = 'o')
    <matplotlib.collections. PathCollection object at 0x0000023EA1958F10>
>>> plt.xlim (0, 20)
    (0.0, 20.0)
>>> plt.ylim (0,14)
    (0.0, 14.0)
>>> plt.title ('Representation des individus dans (x1,x5)')
    Text (0.5, 1.0, 'Representation des individus dans (x1, x5)')
>>> plt.show ()
```

Ci-joint les courbes demandées séparément et rassembles dans une même fenêtre.

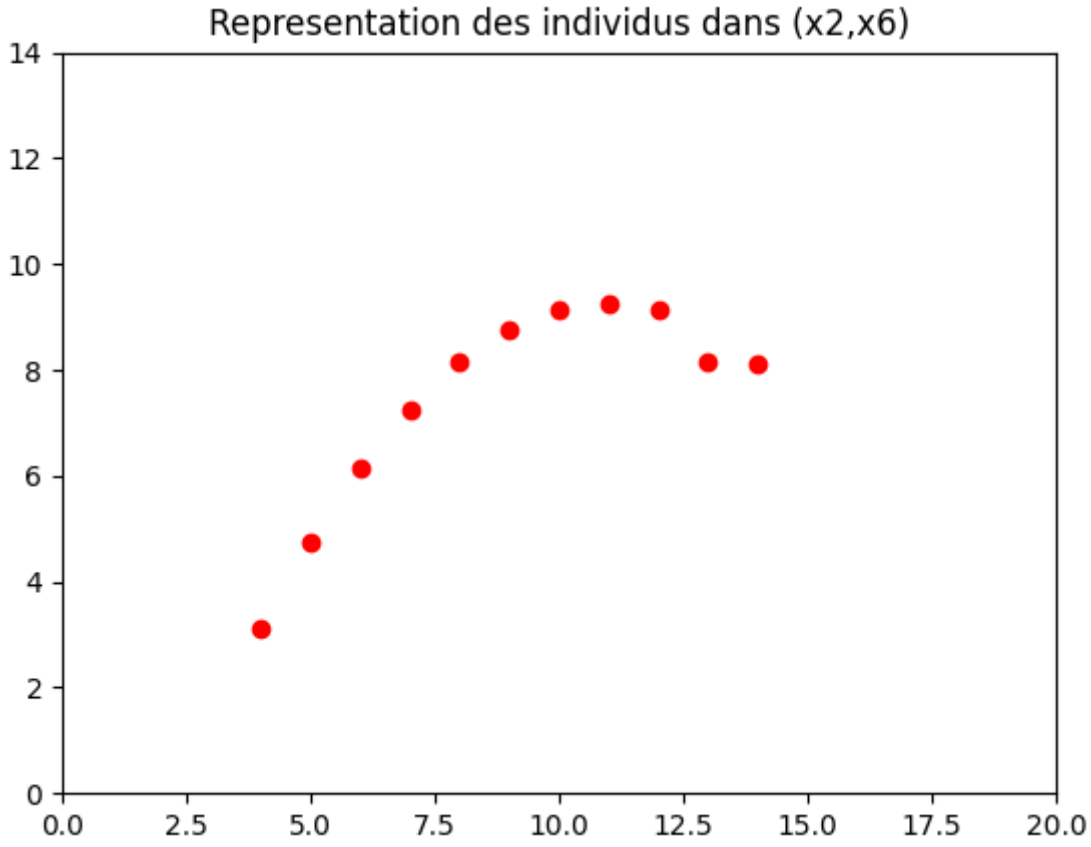
Analyse des données





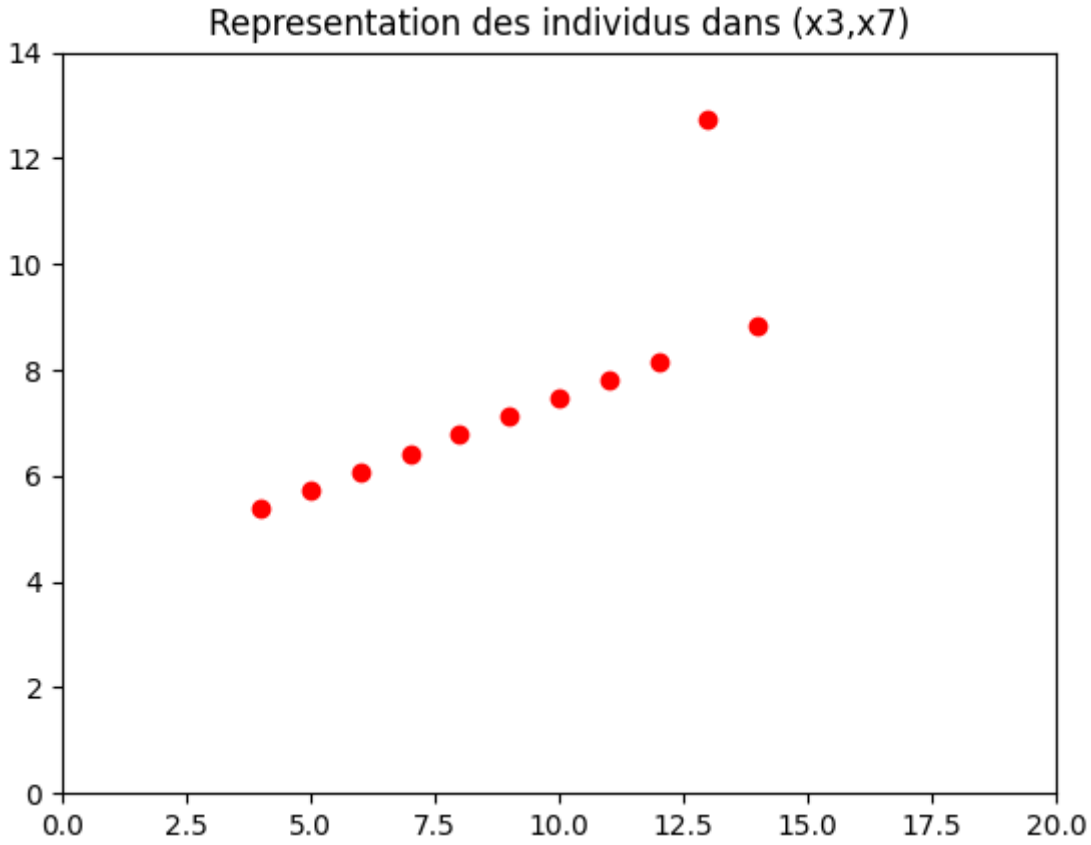
Analyse des données

```
>>> x2= [10,8,13,9,11,14,6,4,12,7,5]
>>> x6= [9.14,8.14,8.14,8.77,9.26,8.10,6.13,3.10,9.13,7.26,4.74]
>>> plt.scatter (x2, x6, c='r', marker = 'o')
    <matplotlib.collections. PathCollection object at 0x0000023EA7408E80>
>>> plt.xlim(0, 20)
    (0.0, 20.0)
>>> plt.ylim(0,14)
    (0.0, 14.0)
>>> plt.title ('Representation des individus dans (x2,  x6)')
    Text (0.5, 1.0, 'Representation des individus dans (x2, x6)')
>>> plt.show ()
```



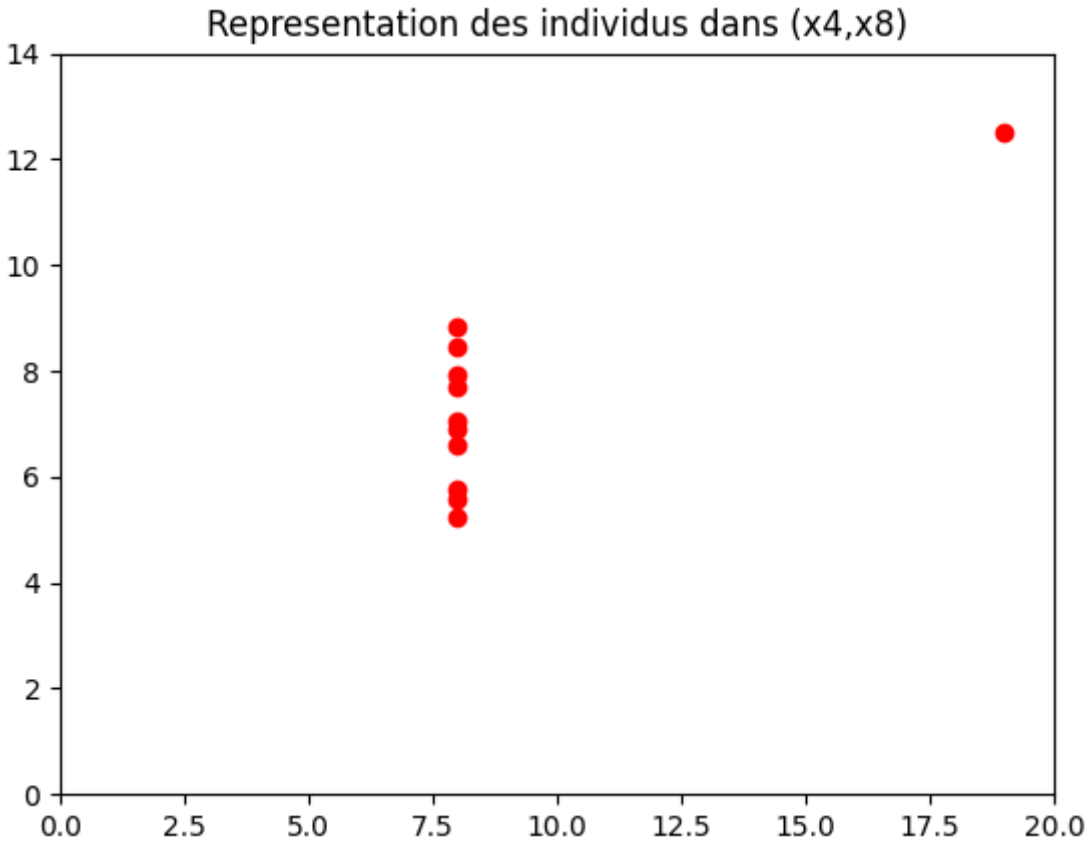

```
>>> x3= [10,8,13,9,11,14,6,4,12,7,5]
>>> x7= [7.46,6.77,12.74,7.11,7.81,8.84,6.08,5.39,8.15,6.42,5.73]
>>> plt.scatter (x3, x7, c='r', marker = 'o')
    <matplotlib.collections. PathCollection object at 0x0000023EA77B51B0>
>>> plt.xlim(0, 20)
    (0.0, 20.0)
>>> plt.ylim(0,14)
    (0.0, 14.0)
>>> plt.title ('Representation des individus dans (x3, x7)')
    Text (0.5, 1.0, 'Representation des individus dans (x3, x7)')
>>> plt.show ()
```

Analyse des données




```
>>> x4= [8,8,8,8,8,8,8,19,8,8,8]
>>> x8= [6.58,5.76,7.71,8.84,8.47,7.04,5.25,12.50,5.56,7.91,6.89]
>>> plt.scatter (x4, x8, c='r', marker = 'o')
    <matplotlib.collections.PathCollection object at 0x0000023EA781D600>
>>> plt.xlim(0, 20)
    (0.0, 20.0)
>>> plt.ylim (0,14)
    (0.0, 14.0)
>>> plt.title('Representation des individus dans (x4, x8)')
    Text (0.5, 1.0, 'Representation des individus dans (x4, x8)')
>>> plt.show ()
```

Analyse des données



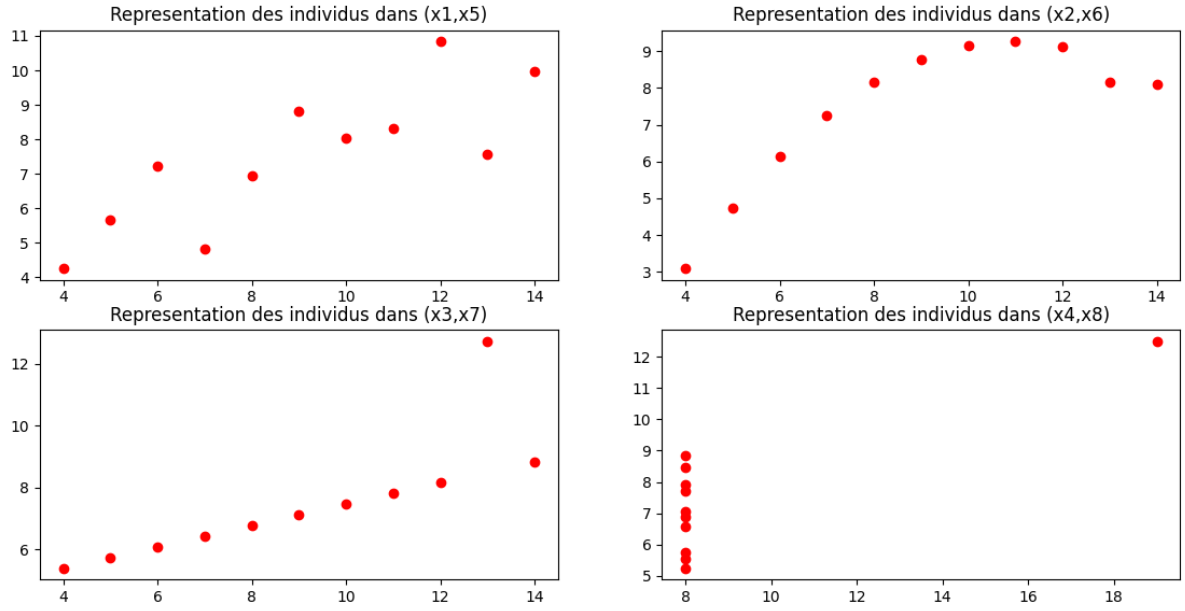
Dans notre cas : **2 lignes, 2colonnes et 4 graphes.**

221	222
223	224

```
>>> plt.subplot(221)
    <AxesSubplot : >
>>> plt.scatter(x1, x5, c='r', marker = 'o')
    <matplotlib.collections.PathCollection object at 0x0000023EA7C11F90>
>>> plt.title('Representation des individus dans (x1, x5)')
    Text (0.5, 1.0, 'Representation des individus dans (x1, x5)')
>>> plt.subplot (222)
    <AxesSubplot : >
>>> plt.scatter (x2, x6, c='r', marker = 'o')
    <matplotlib.collections.PathCollection object at 0x0000023EA7CAEE60>
>>> plt.title('Representation des individus dans (x2, x6)')
    Text (0.5, 1.0, 'Representation des individus dans (x2, x6)')
>>>> plt.subplot (223)
     <AxesSubplot : >
>>> plt.scatter (x3, x7, c='r', marker = 'o')
```

Analyse des données

```
    <matplotlib.collections.PathCollection object at 0x0000023EA7CD0E50>
>>> plt.title ('Representation des individus dans (x3, x7)')
    Text (0.5, 1.0, 'Representation des individus dans (x3, x7)')
>>> plt.subplot(224)
    <AxesSubplot : >
>>> plt.scatter (x4, x8, c='r', marker = 'o')
    <matplotlib.collections.PathCollection object at 0x0000023EA7D45870>
>>> plt.title('Representation des individus dans (x4, x8)')
    Text (0.5, 1.0, 'Representation des individus dans (x4, x8)')
>>> plt.show()
```



Commentaires

Première constatation : Les variables sont différentes entres elles.

- Le **premier graphe** montre que les **données sont bien dispersées, corrélation positive.**
- *Dans le* **deuxième graphe**, **pas de liaison linéaire**, *relation* **parabolique.**
- *Dans le* **troisième graphe**, les données **sont dispersées** d'une manière **linéaire** sauf **1 point**
- *Dans le* **dernier graphe**, **pas de relations (indépendance)** sauf un point et donc **absence de corrélation**.

Attention

Les variables sont différentes entres elles malgré que les résultats de calcul de corrélations montrent carrément autre chose.

Donc, il faut faire attention à l'interprétation des variances, des corrélations.