TD N° 1

Exercice 1

Les Statistiques

Les statistiques en data mining nous aident à **mieux connaître la base de données** et interviennent dans le **prétraitement**.

On a comme données sur l'attribut *âge* :
13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

Remarque Sur Les Données

- Un seul attribut *âge*.
- Les données sont croissantes ce qui facilite les calcules.

Données à plusieurs attributs

- On a tendance à comparer les attributs d'une même base de données afin de découvrir des relations entre eux (des connaissances).
- Pour visualiser une base de données comportant plusieurs attributs (dimensions), on peut utiliser des graphiques comme les *scatter plots* (entre deux attributs) ou
des cartes thermiques *heatmaps* (plusieurs attributs) pour représenter les relations entre les variables.

1. Formule de la moyenne :

$$\boxed{\overline{X} = \frac{\sum x_i}{N}} \qquad \text{où } N \text{ est l'effectif, et } x_i \text{ la } i^{\text{ème}} \text{ valeur.}$$

$$\begin{aligned}
\overline{x} &= \frac{13+15+16\cdot 2+19+20\cdot 2+21+22\cdot 2+25\cdot 4+30+33\cdot 2+35\cdot 4+36+40+45+46+52+70}{27} \\
&= \frac{869}{27} \\
&= \boxed{29.96 \ ans}
\end{aligned}$$

Formule de la mediane

$$\text{Med}(X) = \begin{cases} X\left[\dfrac{N+1}{2}\right] & \text{si } N \text{ est impair,} \\ \dfrac{X\left[\dfrac{N}{2}\right] + X\left[\dfrac{N}{2}+1\right]}{2} & \text{si } N \text{ est pair.} \end{cases}$$

Puisque N est impair $= 27$:

$$\begin{aligned}
\text{Med}(X) &= X\left[\frac{27+1}{2}\right] \\
&= X[14] \\
&= \boxed{25 \ ans}
\end{aligned}$$

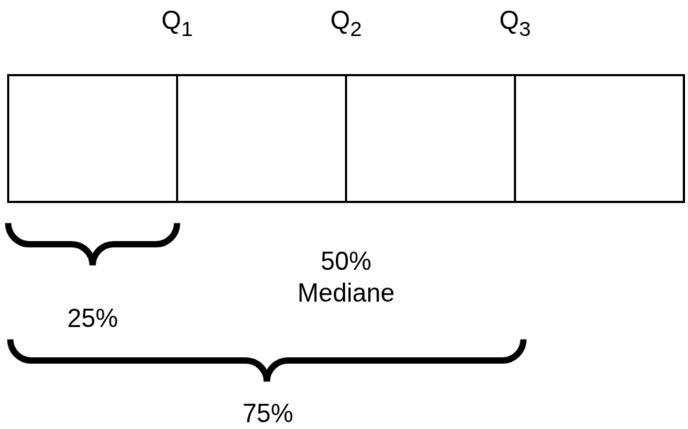
2. Le *mode* représente la ou les valeurs dont la fréquence est la plus élevée (le nombre de répétitions).
   Le *type de modalité* fait référence au nombre de valeurs distinctes correspondant au mode (*bimodale* : 2 valeurs, *trimodale* : 3 valeurs, etc.).

Valeur	Frequence
13	1
16	2
19	1
20	2
21	1
22	2
25	**4**
30	1
33	2
35	**4**
36	1
40	1
45	1
46	1
52	1
70	1

On remarque que les valeurs 25 et 35 ont la plus haute fréquence, répétées 4 fois :

Donc, le *mode* est 25 et 35, et la modalité est dite *bimodale*.

3. Les quartiles permettent de diviser nos données en quatre parties égales :



Les formules :

$$Q_1 = \begin{cases} X\left[\dfrac{N+1}{4}\right] & \text{si } N \text{ est impair,} \\ X\left[\dfrac{N}{4}\right] & \text{si } N \text{ est pair} \end{cases}$$

$$Q_2 = \mathrm{Med}(X)$$

$$Q_3 = \begin{cases} X\left[\dfrac{3\cdot(N+1)}{4}\right] & \text{si } N \text{ est impair,} \\ X\left[\dfrac{3\cdot N}{4}\right] & \text{si } N \text{ est pair} \end{cases}$$

Puisque N est impaire $= 27$:

$$\begin{aligned} Q_1 &= X\left[\frac{27+1}{4}\right] \\ &= X[7] \\ &= \boxed{20\ ans} \end{aligned}$$

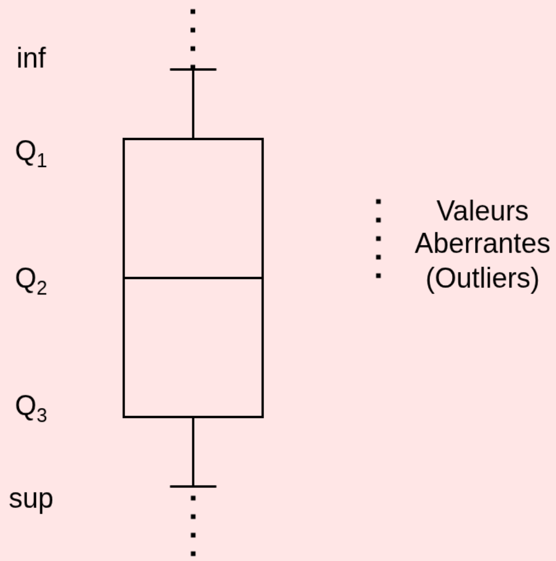
$$\begin{aligned} Q_3 &= X\left[\frac{3\cdot(27+1)}{4}\right] \\ &= X[21] \\ &= \boxed{35\ ans} \end{aligned}$$

4. les cinq nombre de donne sont definie par : min Q_1 Q_2 Q_3 max

$$\min = 13\ ans\ ,\ Q_1 = 20\ ans\ ,\ Q_2 = 25\ ans\ ,\ Q_3 = 35\ ans\ ,\ \max = 70\ ans$$

5. boxPlot

BoxPlot



- Elle nous permet de repérer les valeurs aberrantes.
- La médiane n'est pas forcément située au milieu de la boxplot.
- Les formules :

$$\boxed{\lim_{inf} = Q_1 - (1.5 \times EIQ)}$$

$$\boxed{\lim_{sup} = Q_3 + (1.5 \times EIQ)}$$

$$\boxed{EIQ(\text{Écart interquartile}) = Q_3 - Q_1}$$

- Si $\lim_{inf}$ ou $\lim_{sup}$ ont des valeurs incohérentes ou trop différentes du min ou max, alors on prendra comme limites min et max respectivement.

On a :

$$\begin{cases} X_{min} &= 13 \ ans \\ Q_1 &= 20 \ ans \\ Q_2 &= 25 \ ans \\ Q_3 &= 35 \ ans \\ X_{max} &= 70 \ ans \end{cases}$$

On calcule les limites :

$$EIQ = Q_3 - Q_1 = 35 - 20 = \boxed{15 \ ans}$$

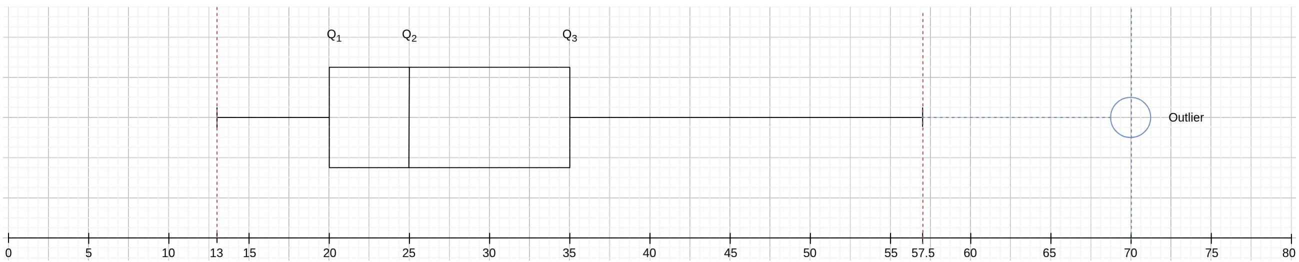
$$\lim_{inf} = Q_1 - (1.5 \times EIQ) = 20 - (1.5 \times 15) = \boxed{-2.5 \ ans}$$

$$\lim_{sup} = Q_3 + (1.5 \times EIQ) = 35 + (1.5 \times 15) = \boxed{57.5 \ ans}$$

Limites

- On a $\lim_{inf} = -2.5$ ans, ce qui est une valeur incohérente dans le contexte âge $\Rightarrow$ on prend comme limite le min $= 13$ ans.
- On a $\lim_{sup} = 57$ ans, valeur normale et pas loin du max $\Rightarrow$ on la prend comme limite.

On prend comme unité :
8 carreaux $\rightarrow$ 5 ans



Conclusion

70 est un outlier

6. Le Q-plot permet de visualiser la distribution d'une seule variable, tandis que le QQ-plot permet de comparer la distribution de deux variables différentes.

Exercice 2

On dispose des données de l'âge et du taux de graisse de 18 adultes dans un hôpital, sélectionnés au hasard :

Âge	23	23	27	27	39	41	47	49	50
Fat %	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
Âge	52	54	54	56	57	58	58	60	61
Fat %	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

Remarques sur les données

- Deux attributs : *âge* et *graisse*.
- Les données sur l'*âge* sont croissantes, ce qui facilite les calculs.
- Les données sur la *graisse* ne sont pas croissantes, donc on doit les ordonner.

1. Formule de l'écart-type :

$$\sigma = \begin{cases} \sqrt{\dfrac{\sum(x_i - \overline{x})^2}{N-1}} & \text{si } N \text{ représente un échantillon} \\ \sqrt{\dfrac{\sum(x_i - \overline{x})^2}{N}} & \text{si } N \text{ représente la population complète} \end{cases}$$

Âge :

$$\begin{aligned} \overline{x} &= \frac{23 \cdot 2 + 27 \cdot 2 + 39 + 40 + 49 + 50 + 52 + 54 \cdot 2 + 56 + 57 + 58 + 60 + 61}{18} \\ &= \frac{836}{18} \\ &= \boxed{46.44 \ ans} \end{aligned}$$

Puisque N est pair $= 18$:

$$\begin{aligned} \text{Med}(X) &= \frac{X\left[\frac{18}{2}\right] + X\left[\frac{18}{2}+1\right]}{2} \\ &= \frac{X[9] + X[10]}{2} \\ &= \frac{50+52}{2} \\ &= \boxed{51 \ ans} \end{aligned}$$

x_i	23	27	39	41	47	49	50
$(x_i - \overline{x})^2$	549.433	377.913	55.353	29.593	0.313	6.553	12.673
x_i	52	54	56	57	58	60	61
$(x_i - \overline{x})^2$	30.913	57.153	91.393	111.513	133.633	183.873	211.993

Puisque N représente un échantillon :

$$\begin{aligned} \sigma &= \sqrt{\frac{\sum(x_i - \overline{x})^2}{18-1}} \\ &= \sqrt{\frac{2970.434}{17}} \\ &= \boxed{13.218 \ ans} \end{aligned}$$

Graisse :

7.8	9.5	17.8	25.9	26.5	27.2	27.4	28.8	30.2
31.2	31.4	32.9	33.4	34.1	34.6	35.7	41.2	42.5

$$\begin{aligned}\overline{y} &= \frac{\substack{7.8+9.5+17.8+25.9+26.5+27.2+27.4+28.8+30.2+31.2+31.4+32.9+33.4+34.1+34.6+35.7\\+41.2+42.5}}{18}\\ &= \frac{518.1}{18}\\ &= \boxed{28.78\ \%}\end{aligned}$$

Puisque N est pair $= 18$:

$$\begin{aligned}\text{Med}(Y) &= \frac{Y\left[\frac{18}{2}\right] + Y\left[\frac{18}{2}+1\right]}{2}\\ &= \frac{Y[9]+Y[10]}{2}\\ &= \frac{30.2+31.2}{2}\\ &= \boxed{30.7\ \%}\end{aligned}$$

y_i	7.8	9.5	17.8	25.9	26.5	27.2	27.4	28.8	30.2
$(y_i-\overline{y})^2$	440.160	371.718	120.560	8.294	5.198	2.496	1.904	0.0001	2.016
y_i	31.2	31.4	32.9	33.4	34.1	34.6	35.7	41.2	42.5
$(y_i-\overline{y})^2$	5.856	6.864	16.974	21.344	28.302	33.872	47.886	154.256	188.238

Puisque N représente un échantillon :

$$\begin{aligned}\sigma &= \sqrt{\frac{\sum(y_i-\overline{y})^2}{18-1}}\\ &= \sqrt{\frac{1455.938}{17}}\\ &= \boxed{9.254\ \%}\end{aligned}$$

2. boxPlots :

Âge

On a :

$$\begin{cases} X_{min} &= 23\ ans\\ Q_1 &= X\left[\frac{N}{4}\right] = X\left[\frac{18}{4}\right] = X[4.5] = X[5] = 39\ ans\\ Q_2 &= 51\ ans\\ Q_3 &= X\left[\frac{3\cdot N}{4}\right] = X\left[\frac{3\cdot 18}{4}\right] = X[13.5] = X[14] = 57\ ans\\ X_{max} &= 61\ ans\end{cases}$$

On calcule les limits :

$$EIQ = Q_3 - Q_1 = 57 - 39 = \boxed{18\ ans}$$

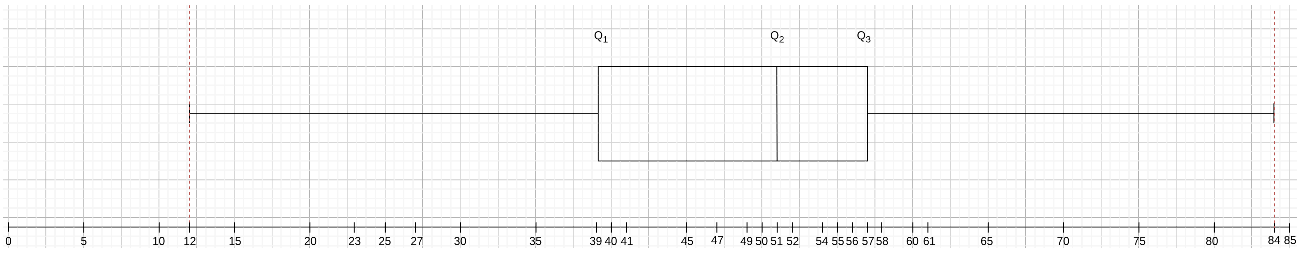
$$\lim_{inf} = 39 - (1.5 \times EIQ) = 20 - (1.5 \times 18) = \boxed{12\ ans}$$

$$\lim_{sup} = 57 + (1.5 \times EIQ) = 35 + (1.5 \times 18) = \boxed{84\ ans}$$

Limits

- On a $\lim_{inf} = 12$ ans, valeur normale et pas loin du min $\Rightarrow$ on la prend comme limite.
- On a $\lim_{sup} = 84$ ans, valeur normale et pas loin du max $\Rightarrow$ on la prend comme limite.

On prend comme unité :
8 carreaux $\rightarrow$ 5 ans



Conclusion

aucun outlier

Graisse
On a :

$$\begin{cases} Y_{min} &= 7.8 \ \% \\ Q_1 &= Y\left[\frac{N}{4}\right] = Y\left[\frac{18}{4}\right] = Y[4.5] = Y[5] = 26.5 \ \% \\ Q_2 &= 30.7 \ \% \\ Q_3 &= Y\left[\frac{3 \cdot N}{4}\right] = Y\left[\frac{3 \cdot 18}{4}\right] = Y[13.5] = X[14] = 34.1 \ \% \\ Y_{max} &= 42.5 \ \% \end{cases}$$

On calcule les limits :

$$EIQ = Q_3 - Q_1 = 34.1 - 26.5 = \boxed{7.6 \ \%}$$

$$\lim_{inf} = 26.5 - (1.5 \times EIQ) = 26.5 - (1.5 \times 7.6) = \boxed{15.1 \ \%}$$

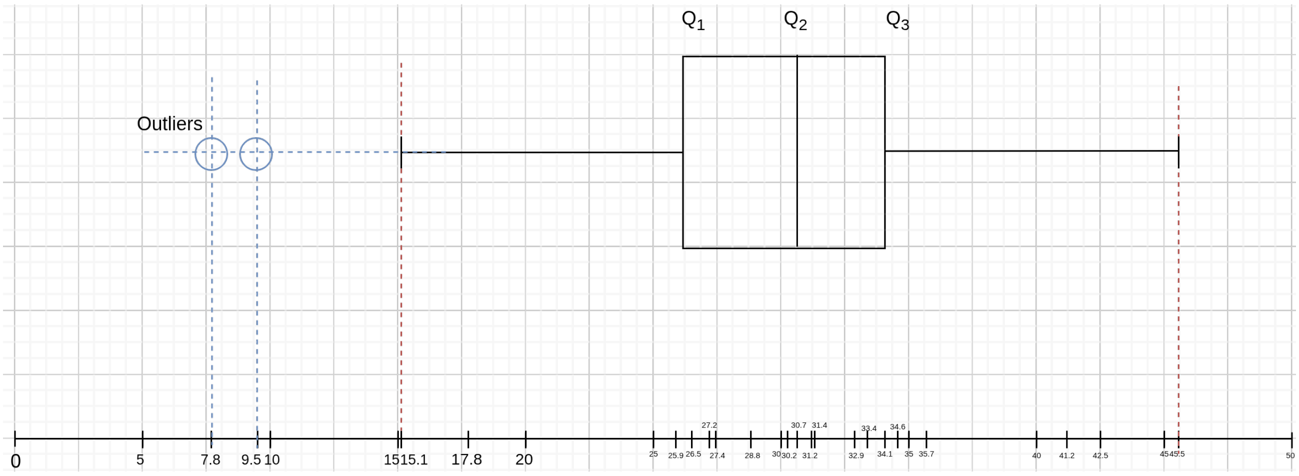
$$\lim_{sup} = 34.1 + (1.5 \times EIQ) = 34.1 + (1.5 \times 7.6) = \boxed{45.5\%}$$

Limits

- On a $\lim_{inf} = 15.1$ %, valeur normale et pas loin du min $\Rightarrow$ on la prend comme limite.
- On a $\lim_{sup} = 45.5$ %, valeur normale et pas loin du max $\Rightarrow$ on la prend comme limite.

On prend comme unité :

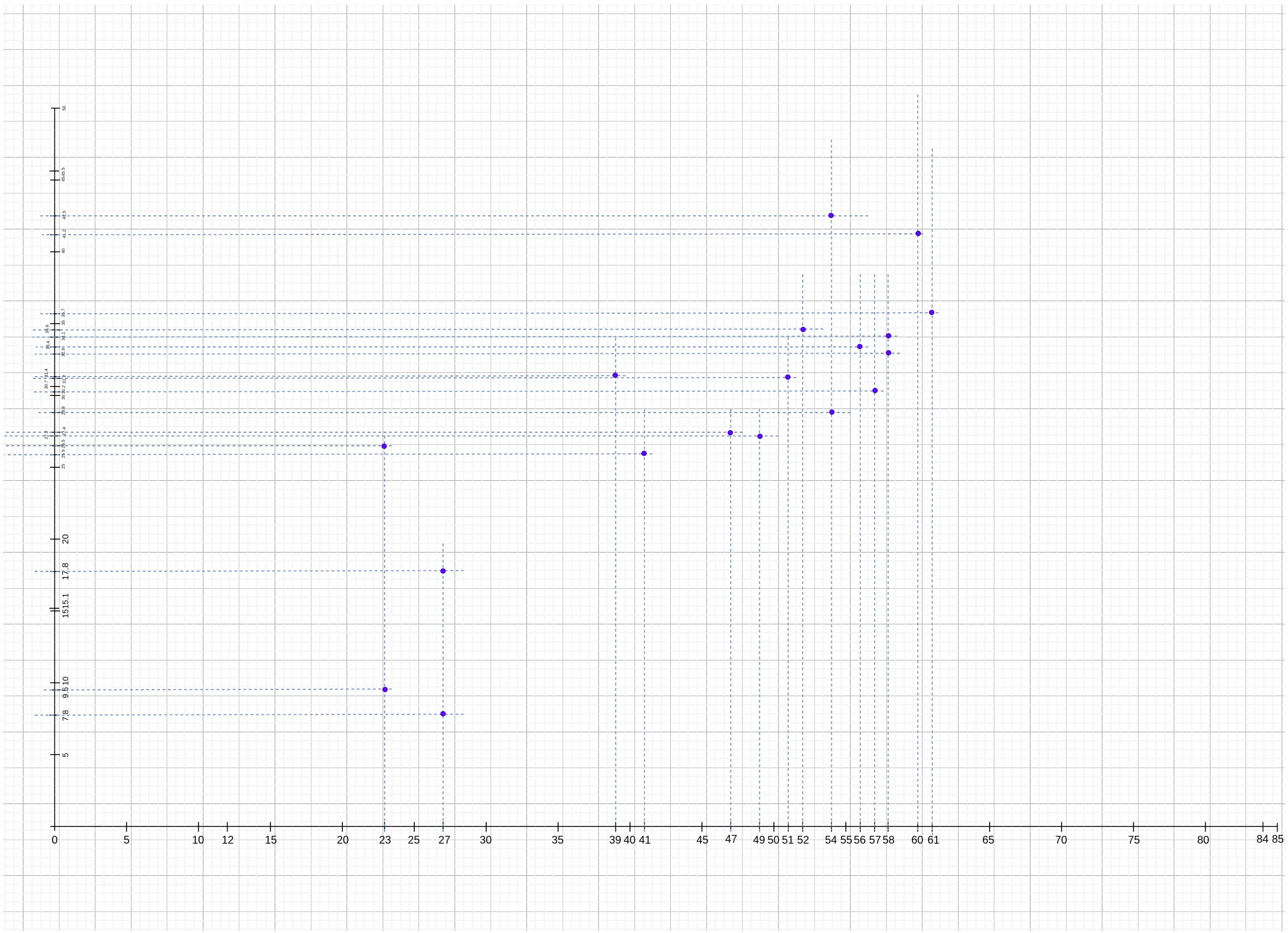
8 carreaux → 5 %



Conclusion

9.5 et 7.8 sont des outliers

3. Scatter Plot



Corrélation

La corrélation dans un scatter plot, c'est la relation entre l'attribut de l'axe X et l'axe Y :

- Corrélation positive : quand X augmente, Y augmente ; quand X diminue, Y diminue aussi.
- Corrélation négative : quand X augmente, Y diminue ; quand X diminue, Y augmente.
- Pas de corrélation : aucune relation.

Dans notre scatter plot on remarque plus la perssone est age plus a la de griasse donc c'est une corelation positive.

4. Normalization Z-score:

Z-score

$$Z = \frac{x_i - \overline{x}}{\sigma}$$

Les valeurs sont entre $[0, 1]$. On normalise parce que les algorithmes (modèles) sont plus performants dans cet intervalle.

Âge

x_i	23	27	39	41	47	49	50
Z	-1.77	-1.47	-0.56	-0.41	0.04	0.19	0.26
x_i	52	54	56	57	58	60	61
Z	0.42	0.57	0.73	0.79	0.87	1.02	1.10

Graisse

y_i	7.8	9.5	17.8	25.9	26.5	27.2	27.4	28.8	30.2
Z	-2.26	-2.08	-1.18	-0.31	-0.24	-0.17	-0.14	0.002	0.15
y_i	31.2	31.4	32.9	33.4	34.1	34.6	35.7	41.2	42.5
Z	0.26	0.28	0.44	0.49	0.57	0.62	0.74	1.34	1.48

5. coefficient de corrélation:

$$r = \begin{cases} \dfrac{\sum(x_i - \overline{x}) \cdot (y_i - \overline{y})}{(N-1) \cdot (\sigma_x \cdot \sigma_y)} & \text{si } N \text{ représente un échantillon} \\[2ex] \dfrac{\sum(x_i - \overline{x}) \cdot (y_i - \overline{y})}{N \cdot (\sigma_x \cdot \sigma_y)} & \text{si } N \text{ représente la population complète} \end{cases}$$

Interprétation de $r \in [-1, 1]$:

$$\begin{cases} r = -1 & \text{corrélation négative linéaire parfaite} \\ -1 < r < 0 & \text{corrélation négative} \\ r = 0 & \text{aucune corrélation} \\ 0 < r < 1 & \text{corrélation positive} \\ r = 1 & \text{corrélation positive linéaire parfaite} \end{cases}$$

$$\begin{aligned} r &= \frac{\sum(x_i - \overline{x}) \cdot (y_i - \overline{y})}{(N-1) \cdot (\sigma_x \cdot \sigma_y)} \\ &= \frac{\sum(x_i - \overline{x}) \cdot (y_i - \overline{y})}{(18-1) \cdot (13.218 \cdot 9.254)} \\ &= \frac{\substack{\sqrt{549.433} \cdot (\sqrt{5.198} + \sqrt{371.718}) + \sqrt{377.913} \cdot (\sqrt{440.160} + \sqrt{120.560}) + \sqrt{55.353} \cdot \sqrt{6.864} + \sqrt{29.593} \cdot \sqrt{8.294} + \\ \sqrt{0.313} \cdot \sqrt{1.904} + \sqrt{6.553} \cdot \sqrt{2.496} + \sqrt{12.673} \cdot \sqrt{5.853} + \sqrt{30.913} \cdot \sqrt{33.872} + \sqrt{57.153} \cdot (\sqrt{188.238} + \sqrt{10^{-4}}) + \\ \sqrt{111.513} \cdot \sqrt{2.016} + \sqrt{183.873} \cdot (\sqrt{28.302} + \sqrt{16.974}) + \sqrt{183.873} \cdot \sqrt{154.246} + \sqrt{211.993} \cdot \sqrt{47.886}}}{2079.429} \\ &= \frac{1723.577}{2079.429} \\ &= \boxed{0.828} \end{aligned}$$

Conclusion

On a $r = 0.828 > 0 \Rightarrow$ corrélation positive, ce qui confirme l'interprétation du scatter plot.

Exercice 3

On a comme data set les notes suivantes d'un groupe de 10 étudiants pours les cours respectifs d'un module X et un module Y :

	X				Y			
	TP1 /10	Test1 /10	EMD1 /20	Final1 /10	TP2 /10	Test2 /10	EMD2 /20	Final2 /10
1	7.75	5	13.5	13.13	7.25	6.5	16	14.88
2	6	6	11	11.50	7.75	5.50	11.5	12.38
3	6	4	7.5	8.75	7.87	6.25	10	12.06
4	6.17	4	5.5	7.84	7.75	4.00	11.5	11.44
5	8	5	11	12	6.62	6.5	13	13.63
6	7.75	2	10.5	10.13	7.5	7.5	12.5	13.31
7	6	4	5.5	7.75	7.37	3.5	12.5	11.75
8	6.75	5	9	10.38	7.62	7.5	5.5	10.19
9	5.67	6	6	8.84	8.37	4	8.5	10.06
10	6.5	6	9	10.75	7.37	5.5	17	15.44

1. Oui car l'intervalle des valeurs de TP, Test est de [0-10] alors que celui de la note final et EMD est de [0-20] donc on doit normalise de tell sorte que tout attributs $\in$ [0-20].

$$\begin{cases} \text{TP\_NORMALISE}_i = \text{TP}_i \times 2 \\ \text{TEST\_NORMALISE}_i = TEST_i \times 2 \\ \text{FINAL\_NORMALIZE}_i = \dfrac{\text{TP\_NORMALISE}_i + \text{TEST\_NORMALIZE} + \text{EMD}_1}{3} \end{cases}$$

	X				Y			
	TP1 /10	Test1 /10	EMD1 /20	Final1 /10	TP2 /10	Test2 /10	EMD2 /20	Final2 /10
1	15.5	10	13.5	13	14.5	13	16	14.5
2	12	12	11	11.66	15.5	11	11.5	12.66
3	12	8	7.5	9.16	15.74	12.5	10	12.74
4	12.34	8	5.5	8.61	15.5	8	11.5	11.66
5	16	10	11	12.33	13.24	13	13	13.08
6	15.5	4	10.5	10	15	15	12.5	14.16
7	12	8	5.5	8.5	14.74	7	12.5	11.41
8	13.5	10	9	10.83	15.24	15	5.5	11.91
9	11.34	12	6	9.78	16.74	8	8.5	11.08
10	13	12	9	11.33	14.74	11	17	14.24

2. Final représente la **moyenne** par rapport aux notes de TP, Test et EMD.

3. on met x final 1 et y final 2 et on ordone les valeurs :

   (a) Moyenne :

	1	2	3	4	5	6	7	8	9	10
final 1	8.5	8.61	9.16	9.78	10	10.83	11.33	11.66	12.33	13
final 2	11.08	11.41	11.66	11.91	12.66	12.74	13.08	14.16	14.24	14.5

$$\begin{aligned}\overline{x} &= \frac{13+11.66+9.16+8.61+12.33+10+8.5+10.83+9.78+11.33}{10}\\ &= \frac{105.2}{10}\\ &= \boxed{10.52}\end{aligned}$$

$$\begin{aligned}\overline{y} &= \frac{14.5+12.66+12.74+11.66+13.08+14.16+11.41+11.91+11.08+14.24}{10}\\ &= \frac{127.44}{10}\\ &= \boxed{12.74}\end{aligned}$$

   (b) Mediane :

$$\begin{aligned}\mathrm{Med}(X) &= \frac{X\left[\frac{N}{2}\right]+X\left[\frac{N}{2}+1\right]}{2}\\ &= \frac{X\left[\frac{10}{2}\right]+X\left[\frac{10}{2}+1\right]}{2}\\ &= \frac{X\left[5\right]+X\left[6\right]}{2}\\ &= \frac{10+10.83}{2}\\ &= \boxed{10.415}\end{aligned}$$

$$\begin{aligned}\mathrm{Med}(Y) &= \frac{Y\left[\frac{N}{2}\right]+Y\left[\frac{N}{2}+1\right]}{2}\\ &= \frac{Y\left[\frac{10}{2}\right]+Y\left[\frac{10}{2}+1\right]}{2}\\ &= \frac{Y\left[5\right]+Y\left[6\right]}{2}\\ &= \frac{12.66+12.74}{2}\\ &= \boxed{12.7}\end{aligned}$$

(c) Mode :

$$\text{Mode}(X) = \boxed{13} \qquad\qquad \text{Mode}(Y) = \boxed{14.5}$$

Conclusion

$$\overline{x} < \overline{y}$$
$$\text{Med}(X) < \text{Med}(Y)$$
$$\text{Mode}(X) < \text{Mode}(Y)$$

(d) Q_1 et Q_3 :

$$\begin{aligned} Q_1(X) &= X\left[\frac{N}{4}\right] \\ &= X\left[\frac{10}{4}\right] \\ &= X[2.5] \\ &= X[3] \\ &= \boxed{9.16} \end{aligned}$$

$$\begin{aligned} Q_3(X) &= X\left[\frac{3\cdot N}{4}\right] \\ &= X\left[\frac{3\cdot 10}{4}\right] \\ &= X[7.5] \\ &= X[8] \\ &= \boxed{11.66} \end{aligned}$$

$$\begin{aligned} Q_1(Y) &= Y\left[\frac{N}{4}\right] \\ &= Y[3] \\ &= \boxed{11.66} \end{aligned}$$

$$\begin{aligned} Q_3(Y) &= Y\left[\frac{3\cdot N}{4}\right] \\ &= Y[8] \\ &= \boxed{14.16} \end{aligned}$$

(e) Le résumé des cinq nombres :

$$X \Rightarrow \min = 8.5 \quad Q_1 = 9.16 \quad Q_2 = 10.415 \quad Q_3 = 11.66 \quad \max = 13$$

$$Y \Rightarrow \min = 11.08 \quad Q_1 = 11.66 \quad Q_2 = 12.7 \quad Q_3 = 14.16 \quad \max = 14.5$$

4. boxplot :

   On a pour final 1:

$$
\begin{cases}
X_{min} & = 8.5 \\
Q_1 & = 9.16 \\
Q_2 & = 10.415 \\
Q_3 & = 11.66 \\
X_{max} & = 13
\end{cases}
$$

   On calcule les limites :

$$
EIQ = Q_3 - Q_1 = 11.66 - 9.16 = \boxed{2.5}
$$

$$
\lim_{inf} = Q_1 - (1.5 \times EIQ) = 9.16 - (1.5 \times 2.5) = \boxed{5.41}
$$

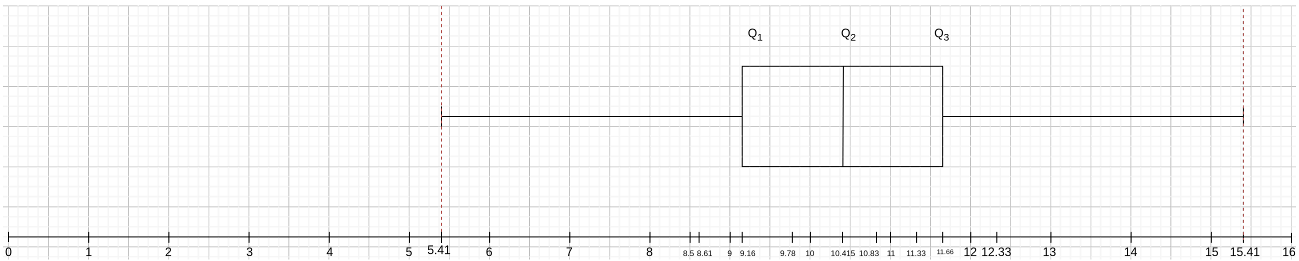
$$
\lim_{sup} = Q_3 + (1.5 \times EIQ) = 11.66 + (1.5 \times 2.5) = \boxed{15.41}
$$

Limites

- On a $\lim_{inf} = 5.41$, valeur normale et pas loin du min $\Rightarrow$ on la prend comme limite.
- On a $\lim_{sup} = 15.41$, valeur normale et pas loin du max $\Rightarrow$ on la prend comme limite.

On prend comme unité :

8 carreaux $\rightarrow$ 1



Conclusion

aucun outlier

On a pour final 2:

$$\begin{cases} Y_{min} & = 11.08 \\ Q_1 & = 11.66 \\ Q_2 & = 12.7 \\ Q_3 & = 14.16 \\ Y_{max} & = 14.5 \end{cases}$$

On calcule les limites :

$$EIQ = Q_3 - Q_1 = 14.16 - 11.66 = \boxed{2.5}$$

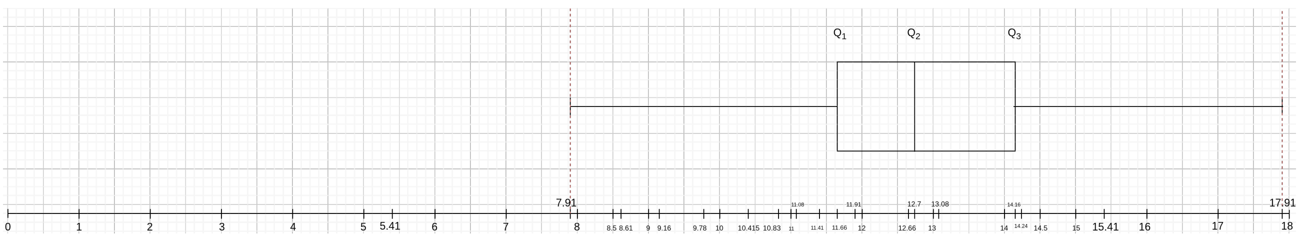
$$\lim_{inf} = Q_1 - (1.5 \times EIQ) = 11.66 - (1.5 \times 2.5) = \boxed{7.91}$$

$$\lim_{sup} = Q_3 + (1.5 \times EIQ) = 14.16 + (1.5 \times 2.5) = \boxed{17.91}$$

Limites

- On a $\lim_{inf} = 7.91$, valeur normale et pas loin du min $\Rightarrow$ on la prend comme limite.
- On a $\lim_{sup} = 17.91$, valeur normale et pas loin du max $\Rightarrow$ on la prend comme limite.

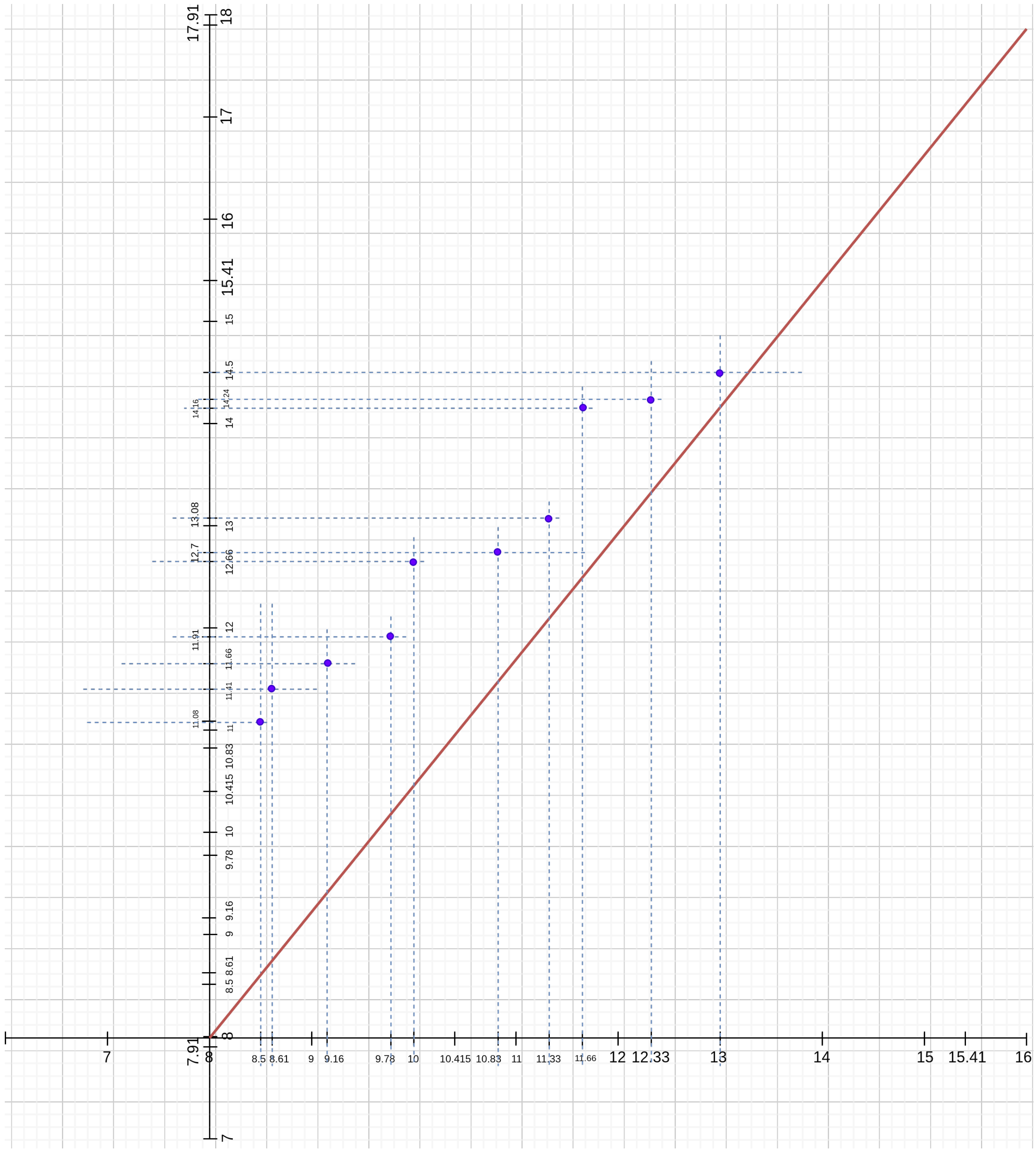
On prend comme unité :
8 carreaux $\rightarrow$ 1



Conclusion

aucun outlier

5. QQ-Plot :



Conclusion

Toutes les moyennes de final 1 sont inférieures à celles de final 2.

6. Coefficient Corrélation :

x_i	8.5	8.61	9.16	9.78	10	10.83	11.33	11.66	12.33	13
$(x_i - \overline{x})^2$	4.08	3.64	1.84	0.54	0.27	0.09	0.65	1.29	3.27	6.15

y_i	11.08	11.41	11.66	11.91	12.66	12.74	13.08	14.16	14.24	14.5
$(y_i - \overline{y})^2$	2.75	1.76	1.16	0.68	0.006	0	0.11	2.01	2.25	3.09

Ecart-type :

$$\begin{aligned}\sigma_x &= \sqrt{\frac{\sum(x_i - \overline{x})^2}{N-1}}\\ &= \sqrt{\frac{4.08+3.64+1.84+0.54+0.27+0.09+0.65+1.29+3.27+6.15}{10-1}}\\ &= \sqrt{\frac{21.82}{9}}\\ &= \sqrt{2.424}\\ &= \boxed{1.556}\end{aligned}$$

$$\begin{aligned}\sigma_y &= \sqrt{\frac{\sum(y_i - \overline{y})^2}{N-1}}\\ &= \sqrt{\frac{2.75+1.76+1.16+0.68+0.006+0+0.11+2.01+2.25+3.09}{10-1}}\\ &= \sqrt{\frac{13.81}{9}}\\ &= \sqrt{1.535}\\ &= \boxed{1.238}\end{aligned}$$

$$\begin{aligned}r &= \frac{\sum(x_i - \overline{x})\cdot(y_i - \overline{y})}{(N-1)\cdot(\sigma_x \cdot \sigma_y)}\\ &= \frac{\sum(x_i - \overline{x})\cdot(y_i - \overline{y})}{(10-1)\cdot(1.556 \cdot 1.238)}\\ &= \frac{\sqrt{4.08}\cdot\sqrt{2.75}+\sqrt{3.64}\cdot\sqrt{1.76}+\sqrt{1.84}\cdot\sqrt{1.16}+\sqrt{0.54}\cdot\sqrt{0.68}+\sqrt{0.27}\cdot\sqrt{0.006}+\sqrt{0.09}\cdot 0+\sqrt{0.65}\cdot\sqrt{0.11}+\sqrt{1.29}\cdot\sqrt{2.01}+\sqrt{3.27}\cdot\sqrt{2.25}+\sqrt{6.15}\cdot\sqrt{3.09}}{17.336}\\ &= \frac{16.937}{17.336}\\ &= \boxed{0.977}\end{aligned}$$

Conclusion

On a $r = 0.977 > 0 \Rightarrow$ corrélation positive