

Analisi di dati ambientali tramite risorse open source per la Data Science

Introduzione alla Statistica Descrittiva con R

Roberto Ascari – roberto.ascari@unimib.it

Notazione

- **Popolazione:** insieme di tutte le entità che siamo interessati a studiare (persone, città, transazioni, rilevazioni, ecc.).
- **Unità statistica:** singole entità che compongono la popolazione.
- **Variabile:** caratteristica delle unità statistiche che siamo interessati a studiare. I valori di una variabile tendono a *variare* da una unità statistica all'altra. Le indicheremo con lettere maiuscole: X , Y e Z .
- **Modalità:** sono i valori che le variabili possono assumere. Le indicheremo con lettere minuscole: x , y e z .

Tipologia di variabili (1)

Le variabili possono essere suddivise sulla base delle modalità che possono assumere.

Variabili quantitative hanno modalità numeriche.

- **Quantitative continue:** variabili le cui modalità possono assumere qualsiasi valore all'interno di un intervallo reale (es. la temperatura, l'altezza di una persona).
- **Quantitative discrete:** variabili le cui modalità possono assumere solo un numero finito o un'infinità numerabile di valori (es. il # di figli in una famiglia, l'età in anni compiuti).

Tipologia di variabili (2)

Variabili qualitative hanno per modalità delle etichette/categorie.

- **Qualitative ordinabili:** variabili le cui modalità sono categorie che possono essere ordinate secondo un criterio naturale (es. il livello di istruzione, il grado di soddisfazione).
- **Qualitative nominale:** variabili le cui modalità sono etichette che non hanno un ordine naturale (es. il colore degli occhi, il gruppo sanguigno).

Matrice dei dati

- I dati possono essere raccolti in una matrice avente le **unità statistiche** sulle righe e le **variabili** sulle colonne.

	Variabile 1 X	Variabile 2 Y	Variabile K Z
Unità 1	x_1	y_1	...	z_1
Unità 2	x_2	y_2	...	z_2
...
Unità i	x_i	y_i	...	z_i
...
Unità N	x_N	y_N	...	z_N

Particolato atmosferico

- Il particolato atmosferico è formato da una miscela complessa di particelle solide e liquide di natura organica o inorganica, sospese nell'aria.
- Il particolato si distingue, in base al diametro aerodinamico, in:
 - PM_{10} con diametro aerodinamico inferiore a $10\text{ }\mu\text{m}$, in grado di penetrare nel tratto superiore dell'apparato respiratorio;
 - $PM_{2.5}$ con diametro aerodinamico inferiore a $2.5\text{ }\mu\text{m}$, in grado di raggiungere i polmoni ed i bronchi secondari.
- La direttiva europea 2008/50/CE indica come soglia per il PM_{10} il valore $50\text{ }\mu\text{g}/\text{m}^3$, da non superare per più di 35 giorni in un anno.

Condizioni Logiche in R

- Uguaglianza: $x == y$
- Diverso da: $x != y$
- Disuguaglianze: $x > y$; $x >= y$; $x < y$; $x <= y$

Connettori logici:

- **And:** restituisce TRUE se entrambe le condizioni sono TRUE
 $(x == y) \& (x >= z)$
- **Or:** restituisce TRUE se almeno una condizione è TRUE
 $(x == y) \mid (x >= z)$
- **Not:** nega la condizione
 $!(x >= y)$

Indici di posizione (1)

- **Media aritmetica**

$$\mu_X = \frac{1}{N} (x_1 + x_2 + \cdots + x_N) = \frac{1}{N} \sum_{i=1}^N x_i$$

Tipologia	Interrogazione	Verifica	Interrogazione	Lavoro di gruppo	Interrogazione	Verifica
Voto	2	8	7	8.5	7	6.5

$$\mu = \frac{1}{6} (2 + 8 + 7 + 8.5 + 7 + 6.5) = \frac{39}{6} = 6.5$$

Indici di posizione (2)

- **Media aritmetica pesata** (o ponderata). Anziché trattare tutti i valori equamente, diamo più importanza ad alcuni tramite un sistema di pesi **non negativi** w_1, w_2, \dots, w_N .

$$\mu_X^W = \frac{\textcolor{red}{w}_1 x_1 + \textcolor{green}{w}_2 x_2 + \dots + \textcolor{blue}{w}_N x_N}{\textcolor{red}{w}_1 + \textcolor{green}{w}_2 + \dots + \textcolor{blue}{w}_N} = \sum_{i=1}^N w_i^* x_i,$$

dove $w_i^* = \frac{w_i}{w_1 + w_2 + \dots + w_N}$.

Dato che $w_i \geq 0$, si ha che $w_i^* \geq 0$ e $\sum_{i=1}^N w_i^* = 1$.

Indici di posizione (2)

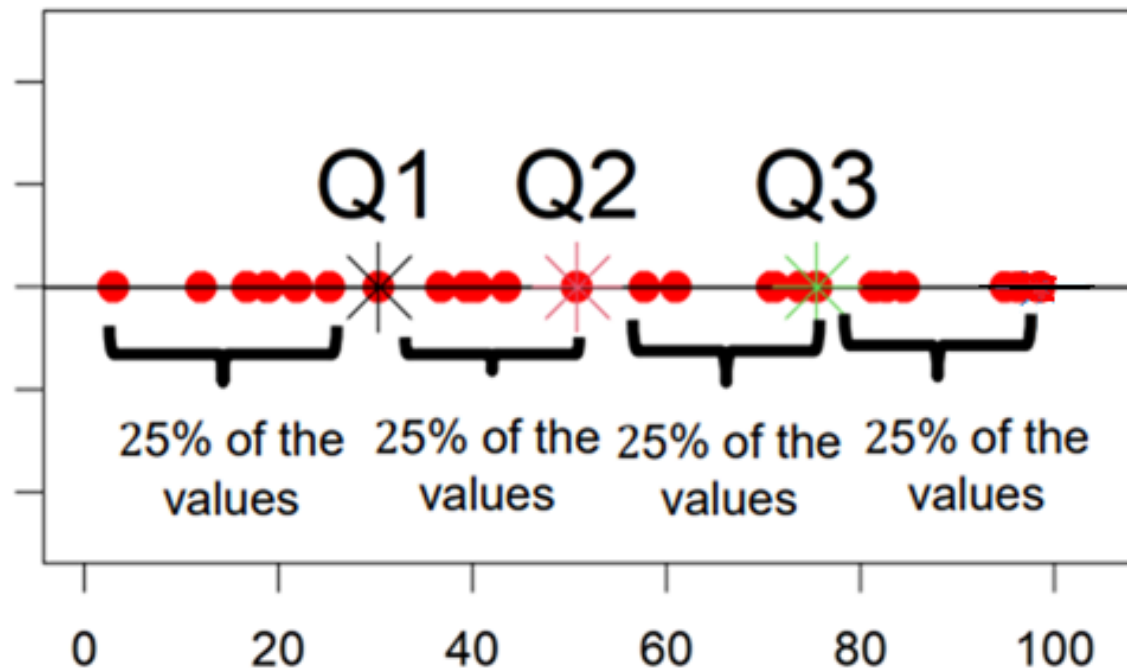
Ad esempio, supponiamo che il peso delle interrogazioni sia una volta e mezza quello delle verifiche e che il lavoro di gruppo abbia un peso doppio rispetto ad una verifica:

Tipologia	Interrogazione	Verifica	Interrogazione	Lavoro di gruppo	Interrogazione	Verifica
Peso	1.5	1	1.5	2	1.5	1
Voto	2	8	7	8.5	7	6.5

$$\mu^W = \frac{((1.5*2)+(1*8)+(1.5*7)+(2*8.5)+(1.5*7)+(1*6.5))}{(1.5+1+2+1.5+2+1)}$$
$$= \frac{1.5}{8.5} * 2 + \frac{1}{8.5} * 8 + \frac{1.5}{8.5} * 7 + \frac{2}{8.5} * 8.5 + \frac{1.5}{8.5} * 7 + \frac{1}{8.5} * 6.5 = 6.529$$

Indici di posizione (3)

- **Quantili.** Il quantile di ordine p è quel valore che, **nella successione ordinata dei dati**, lascia a sinistra il $p\%$ dei dati. In altre parole, è quel valore che è **più grande o uguale** del $p\%$ dei dati.
- Se $p \in \{0.25, 0.5, 0.75\}$, i quantili prendono il nome di **quartili**, dato che suddividono la variabile in 4 parti, ciascuna contenente il 25% dei dati.



Indici di posizione (3)

- Se $p \in \{0.1, 0.2, 0.3, \dots, 0.8, 0.9\}$, i quantili prendono il nome di **decili**.
- Se $p \in \{0.01, 0.02, \dots, 0.98, 0.99\}$, i quantili prendono il nome di **percentili**.
- Il quantile di ordine $p = 0.5$ viene chiamato **mediana**, la quale rappresenta il valore che, nella successione ordinata dei dati, occupa la posizione centrale.

$$Me(X) = \begin{cases} x_{\left(\frac{N+1}{2}\right)} & \text{se } N \text{ è dispari} \\ \frac{1}{2} \left(x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)} \right) & \text{se } N \text{ è pari} \end{cases}$$

dove $x_{(k)}$ rappresenta l'elemento in posizione k nella successione ordinata dei dati.

Indici di posizione (3)

Tipologia	Interrogazione	Verifica	Interrogazione	Lavoro di gruppo	Interrogazione	Verifica
Voto	2	8	7	8.5	7	6.5

Tipologia	Interrogazione	Verifica	Interrogazione	Interrogazione	Verifica	Lavoro di gruppo
Voto	2	6.5	7	7	8	8.5
	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$

- $\mu = 6.5$
- $\mu^W = 6.529$
- $Me(X) = x_{\left(\frac{6}{2}\right)} + x_{\left(\frac{6}{2}+1\right)} = \frac{x_{(3)} + x_{(4)}}{2} = \frac{7+7}{2} = 7$

Indici di variabilità (1)

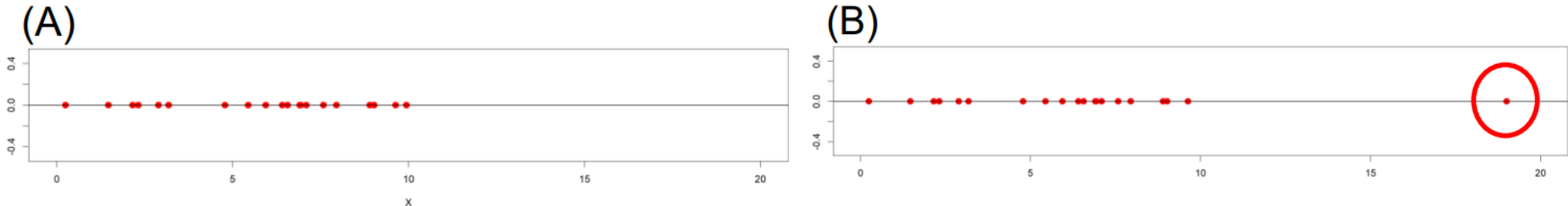
- **Range** (campo di variazione). Differenza tra la modalità massima osservata e quella minima:

$$R_X = \max(X) - \min(X).$$

Si tratta di un indice molto sensibile a valori anomali.

- **Range inter-quartilico** (IQR). Differenza tra il terzo ed il primo quartile:

$$IQR_X = Q_3(X) - Q_1(X).$$



Indici di variabilità (2)

- **Varianza.** La varianza è la media del quadrato degli scarti di ogni x_i dalla media di X .

$$\sigma_X^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu_X)^2$$

- La varianza è sempre non-negativa.
- Assume valore 0 quando non c'è variabilità.
- Valori maggiori indicano una maggiore variabilità.
- L'unità di misura della varianza è il quadrato dell'unità di misura dei dati.

Indici di variabilità (3)

- **Deviazione standard.** È la radice quadrata della varianza:

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu_X)^2}$$

Continuano a valere le prime tre osservazioni viste per la varianza, ma l'unità di misura della deviazione standard coincide con quella dei dati.

$$\sigma^2 = \frac{1}{6} [(2 - 6.5)^2 + \dots + (6.5 - 6.5)^2] = 27/6 = 4.5$$

$$\sigma = \sqrt{4.5} = 2.121$$

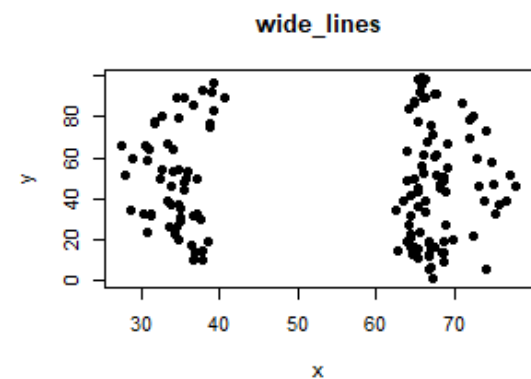
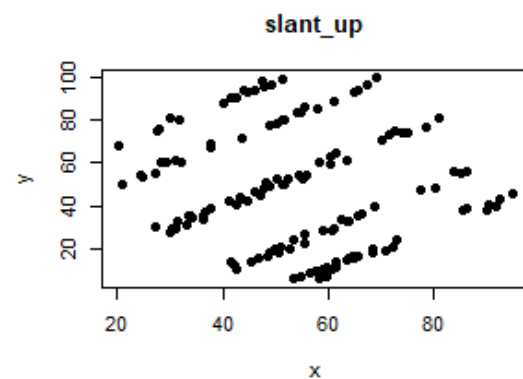
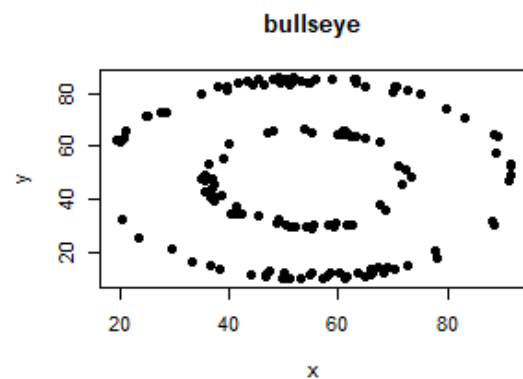
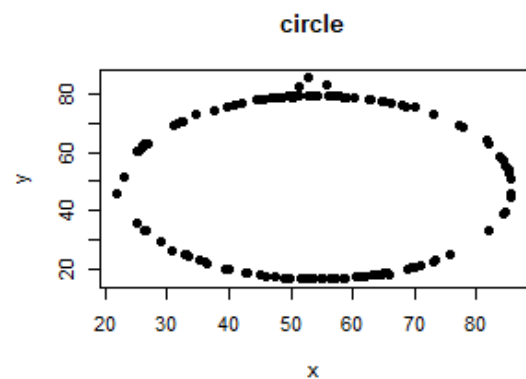
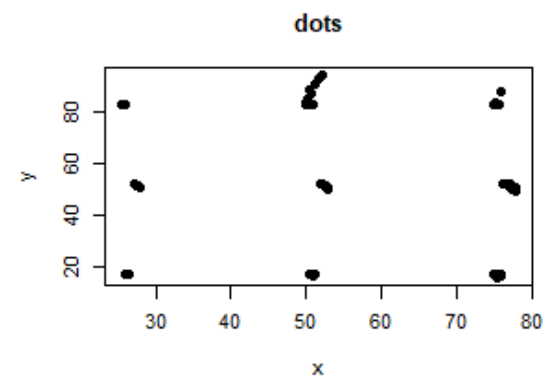
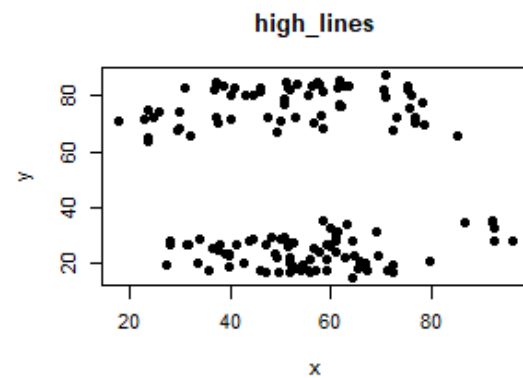
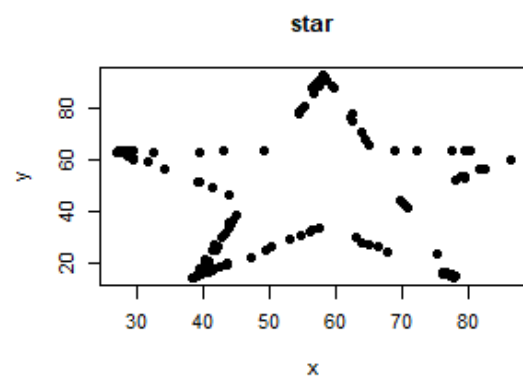
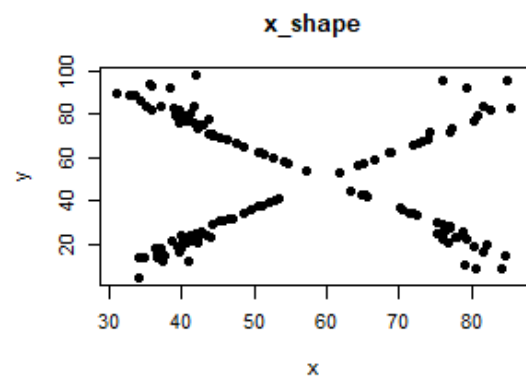
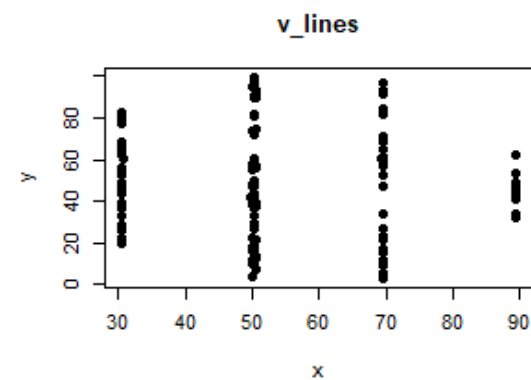
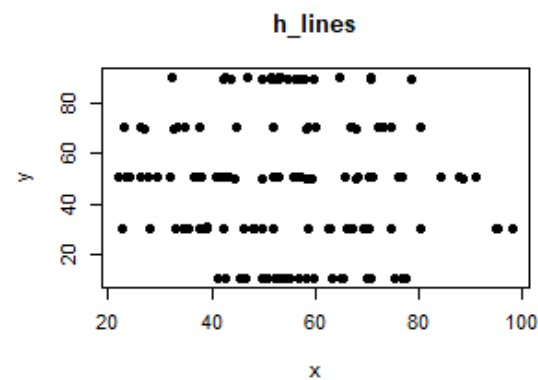
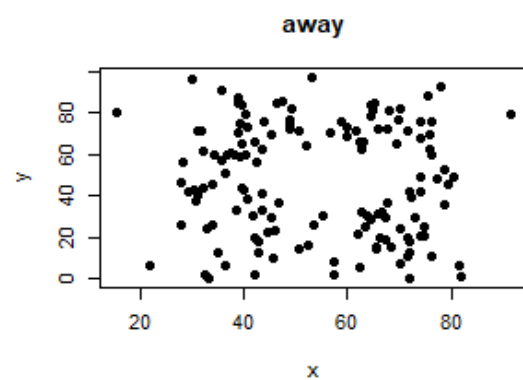
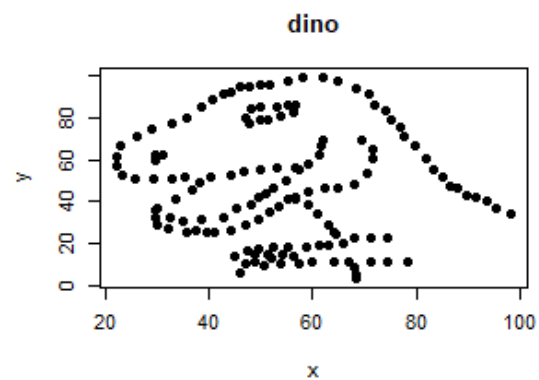
I voti ottenuti si scostano dalla propria media di circa 2.121 punti.

Rappresentazioni grafiche

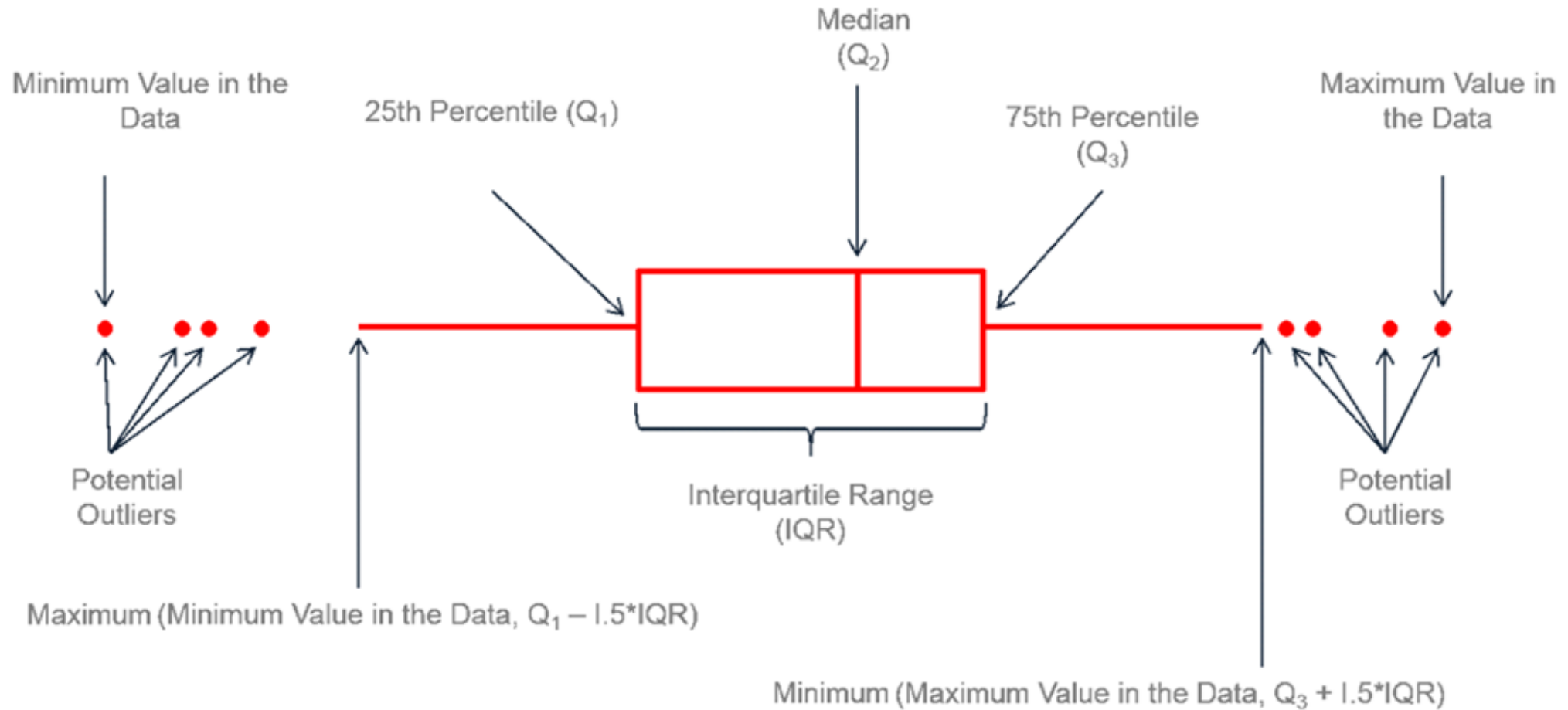
Sintetizzare i dati anche tramite grafici è utile per diversi motivi:

- **Identificare pattern e valori anomali.** Rappresentazioni grafiche permettono di individuare andamenti e anomalie che le statistiche di sintesi da sole non possono evidenziare.
- **Confrontare** dataset diversi o distribuzioni nel tempo è più intuitivo con grafici sovrapposti rispetto a una tabella di numeri.
- **Comunicare in modo efficace.** Un'immagine è più immediata di una lista di statistiche, rendendo le informazioni più accessibili anche a chi non ha esperienza avanzata in statistica.
- **Paradosso di Anscombe:** dataset differenti possono portare alle stesse statistiche di sintesi, ma mostrare strutture anche molto diverse nei dati.

Rappresentazioni grafiche

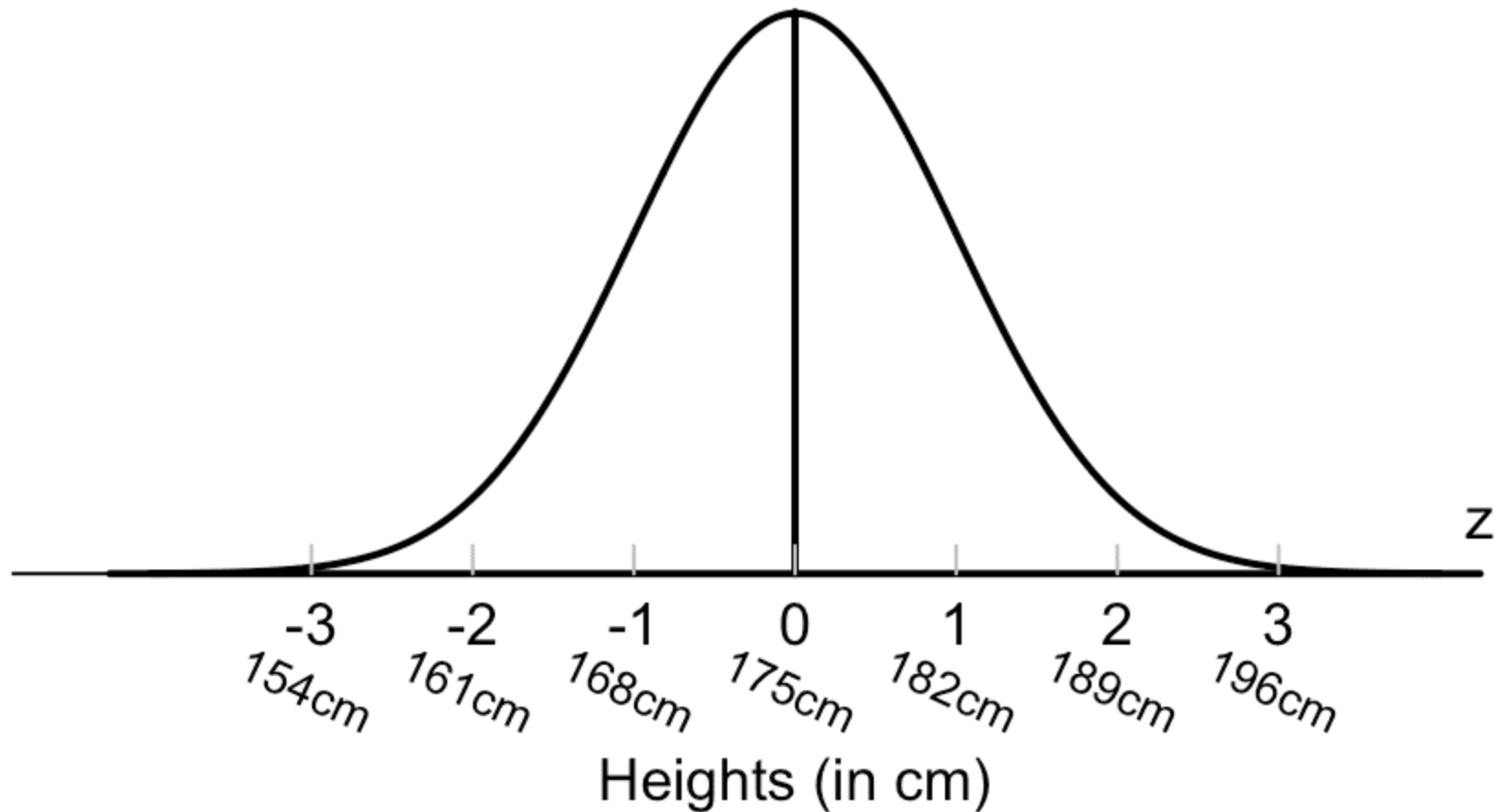


Boxplot

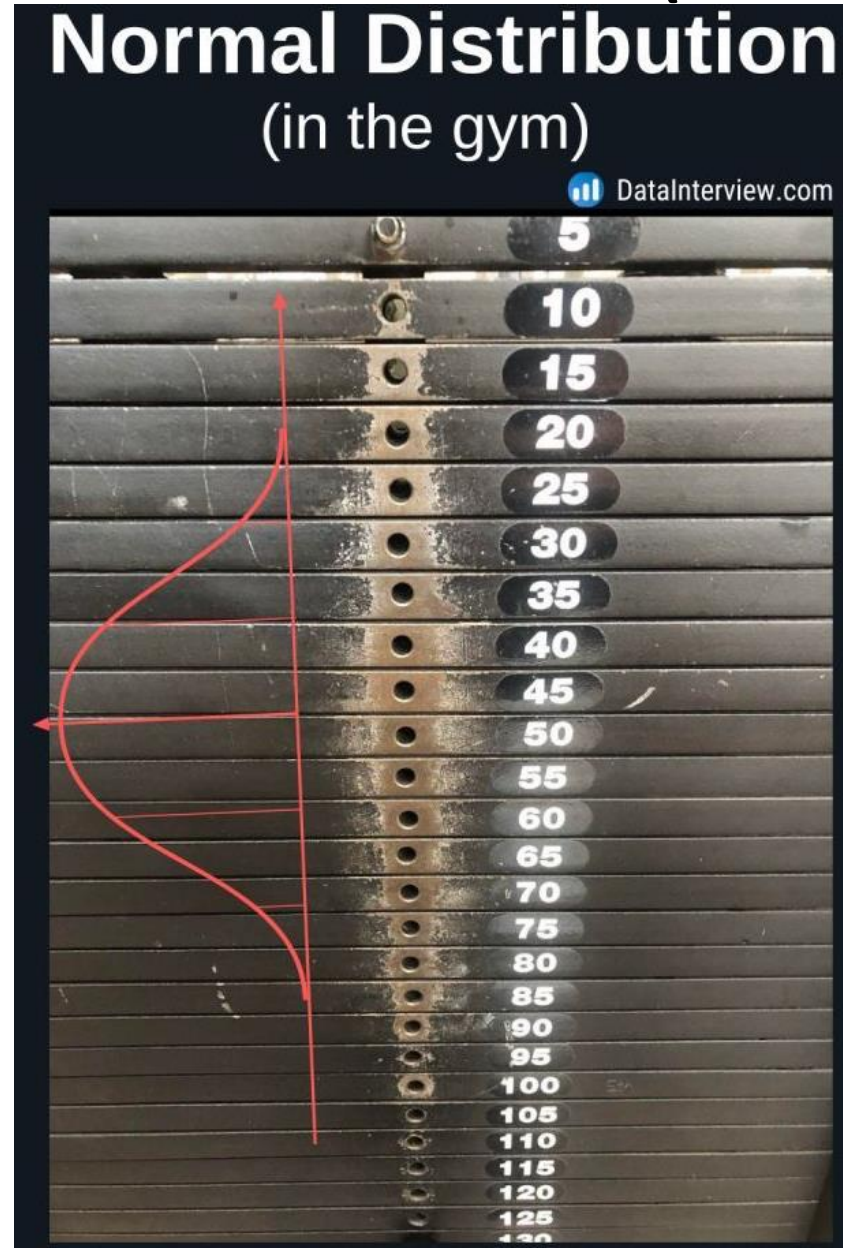


La distribuzione Normale (o Gaussiana)

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in (-\infty, +\infty)$$

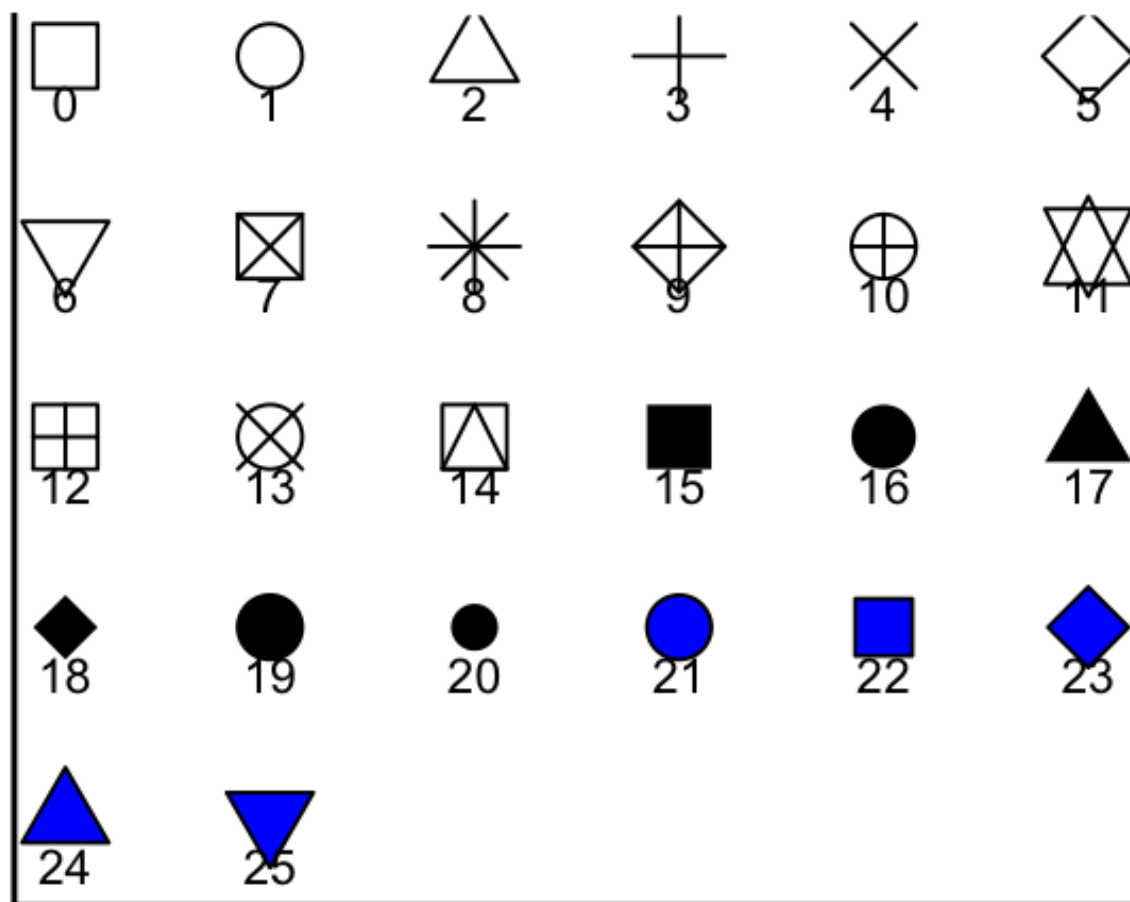


La distribuzione Normale (o Gaussiana)









Tipologia punti

Point shapes available in R



Tipologia linee

6.'twodash'	
5.'longdash'	
4.'dotdash'	
3.'dotted'	
2.'dashed'	
1.'solid'	
0.'blank'	

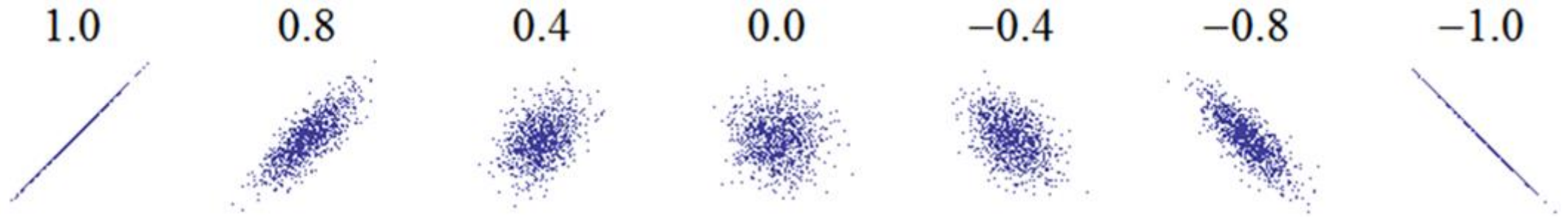
Coefficiente di correlazione lineare

- Il **coefficiente di correlazione lineare** $\rho_{X,Y}$ misura la dipendenza lineare tra le variabili X e Y .

$$\rho_{X,Y} = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\sigma_X^2 \cdot \sigma_Y^2}}.$$

- $\rho_{X,Y}$ assume valori nell'intervallo $[-1, 1]$ dove:
 - -1 indica una perfetta relazione lineare negativa;
 - 0 indica assenza di legame **lineare**;
 - 1 indica una perfetta relazione lineare positiva.

Coefficiente di correlazione lineare



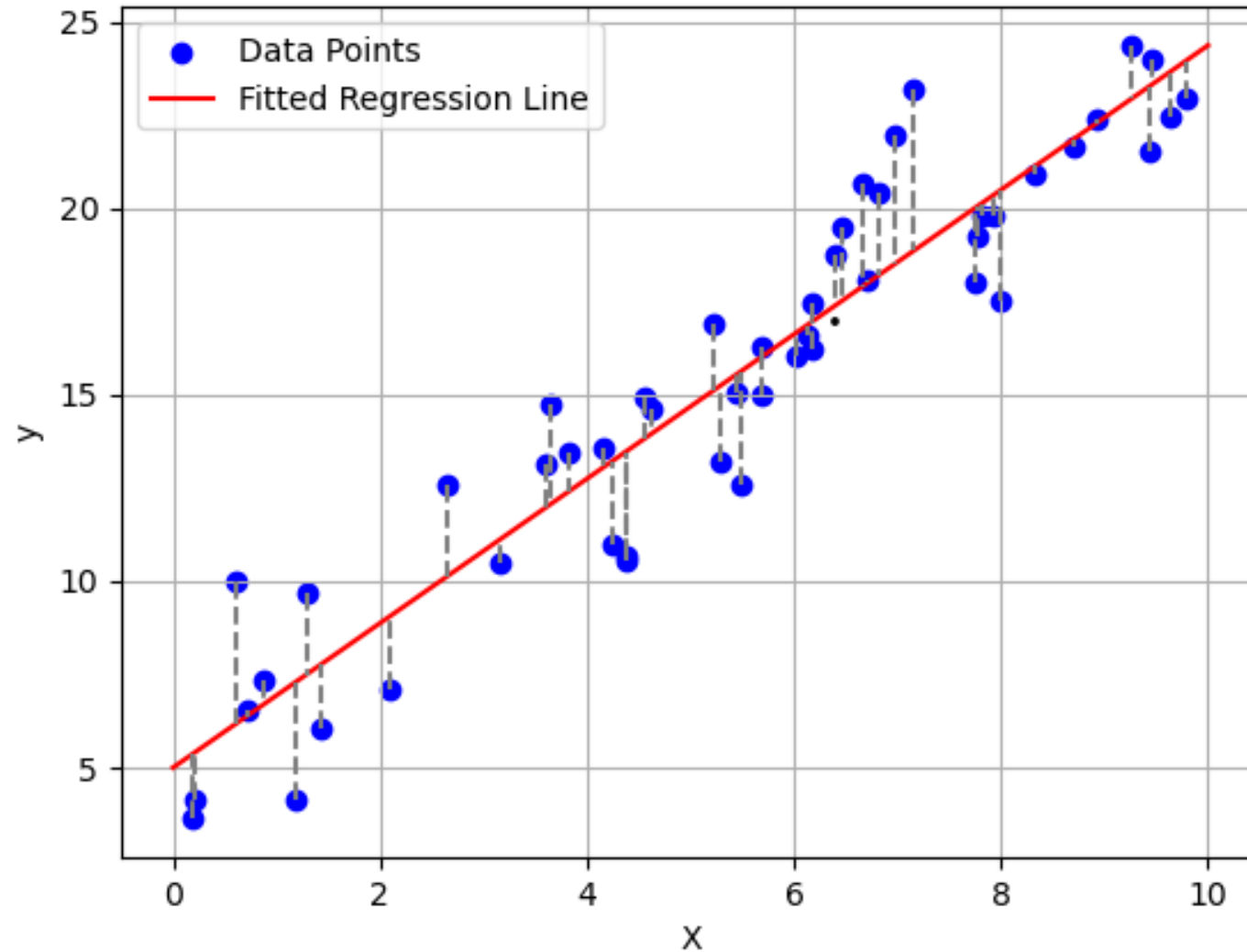
Retta dei Minimi Quadrati

- Si tratta di quella retta $y = a + bx$ che meglio approssima la relazione tra le variabili X e Y .
- Si trovano quei valori a e b che minimizzano la seguente quantità:

$$Q(a, b) = \sum_{i=1}^N [y_i - (a + bx_i)]^2.$$

- $b = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \mu_X)(y_i - \mu_Y)}{\sigma_X^2}$
- $a = \mu_Y - b \mu_X$

Retta dei Minimi Quadrati




Inner Join

Table: Customers

customer_id	first_name
1	John
2	Robert
<u>3</u>	David
4	John
<u>5</u>	Betty

Table: Orders

order_id	amount	customer_id
1	200	10
2	500	<u>3</u>
3	300	6
4	800	<u>5</u>
5	150	8



customer_id	first_name	amount
3	David	500
5	Betty	800