Per normalizzare un dataset numerico, esistono vari metodi matematici e statistici. La scelta del tipo di normalizzazione dipende spesso dalle caratteristiche dei dati e dagli obiettivi dell'analisi. Ecco i tipi più comuni:

1. Normalizzazione Min-Max (Scaling)

- Trasforma i dati in un range prestabilito, di solito $[0, 1]$ o $[-1, 1]$.
- Formula: $X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$
- Vantaggi: Mantiene la distribuzione originale dei dati e i valori estremi.
- Svantaggi: Sensibile agli outlier.

2. Z-Score Standardization (Standardizzazione)

- Trasforma i dati in una distribuzione con media 0 e deviazione standard 1.
- Formula: $Z = \frac{X - \mu}{\sigma}$ dove μ è la media e σ la deviazione standard.
- Vantaggi: Molto usato quando i dati seguono una distribuzione normale.
- Svantaggi: Sensibile agli outlier.

3. Normalizzazione Mediana e Intervallo Interquartile (Robust Scaling)

- Centra i dati sulla mediana e li scala usando l'intervallo interquartile (IQR).
- Formula: $X' = \frac{X - \text{mediana}}{\text{IQR}}$
- Vantaggi: Robusta contro outlier, adatta per distribuzioni non gaussiane.
- Svantaggi: Non è efficace per distribuzioni molto asimmetriche.

4. Normalizzazione Logaritmica

- Applicabile a dati positivi, riduce la varianza dei valori.
- Formula: $X' = \log(X + 1)$
- Vantaggi: Comprime ampi range, adatta per dati con distribuzioni asimmetriche.
- Svantaggi: Non definita per dati con valori nulli o negativi.

5. Normalizzazione tramite Funzione Sigmoidale

- Applica una funzione sigmoide ai dati, restringendo i valori tra 0 e 1.
- Formula: $X' = \frac{1}{1+e^{-X}}$
- Vantaggi: Riduce l'impatto dei valori estremi, creando una "mappatura" più morbida.
- Svantaggi: Non conserva la linearità dei dati.

6. Max Abs Scaling (Normalizzazione tramite Valore Assoluto Massimo)

- I dati vengono divisi per il loro valore assoluto massimo.
- Formula: $X' = \frac{X}{|X_{\max}|}$
- Vantaggi: Utile per mantenere la sparseness nei dati sparsi.
- Svantaggi: Non scala i dati con media 0 e deviazione standard.

7. Rank Transformation (Trasformazione di Ranghi)

- Trasforma i dati assegnando un rango ad ogni osservazione.
- Vantaggi: Utile per normalizzare dati ordinali e con distribuzioni non gaussiane.
- Svantaggi: Perde informazioni sulla distanza assoluta tra i dati.

8. Normalizzazione tramite Trasformazione di Box-Cox o Yeo-Johnson

- Applicabile a distribuzioni non gaussiane per renderle più normali.
- Box-Cox è adatta solo a dati positivi, mentre Yeo-Johnson funziona anche su dati negativi.
- Vantaggi: Adatta per dati con asimmetria significativa.
- Svantaggi: Richiede un calcolo iterativo per trovare i parametri ottimali.

9. Decolorazione o Whitening (Sfericizzazione)

- Usata nelle reti neurali e visione artificiale per rimuovere correlazioni nei dati.
- Vantaggi: Utilizzata per ridurre le dipendenze lineari nei dataset ad alta dimensione.
- Svantaggi: Può rendere i dati più difficili da interpretare in termini intuitivi.

Ogni metodo ha specifiche applicazioni e limiti, e la scelta del metodo di normalizzazione più adatto dipende dal tipo di distribuzione dei dati, dalla presenza di outlier, e dagli scopi dell'analisi o del modello di machine learning.

Lezione: Normalizzazione dei Dati

Obiettivo della Lezione

Imparare le tecniche di normalizzazione più comuni, comprendere quando e come applicarle su vari tipi di dati e capire i vantaggi e gli svantaggi di ciascun metodo. Vedremo esempi pratici per ciascun metodo.

---

1. Cos'è la Normalizzazione?

La normalizzazione è un processo che trasforma i dati numerici per portarli in una scala comune. È spesso usata in machine learning e statistica per garantire che variabili diverse contribuiscano in modo simile ai modelli di analisi.

---

2. Perché è Importante Normalizzare i Dati?

- **Migliora le performance** degli algoritmi di machine learning, specialmente quelli basati su distanze (es. KNN, clustering).
- **Evita che variabili con scale diverse** dominino l'analisi.
- **Riduce l'impatto degli outlier**, in base al metodo scelto.

---

3. Tipi di Normalizzazione

Vediamo insieme i metodi più usati.

3.1. Normalizzazione Min-Max (Scaling)

Questo metodo trasforma i dati in un range specifico, solitamente $[0, 1]$. È utile quando vogliamo mantenere la struttura dei dati originale, ma adattarli ad una scala comune.

Formula

$$X' = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

Esempio

Consideriamo questo dataset:

- Altezza (cm): $150, 160, 170, 180, 190$

Applichiamo Min-Max per normalizzare su scala $[0, 1]$:

- **Step 1**: Troviamo il minimo e massimo (min = 150, max = 190).
- **Step 2**: Applichiamo la formula.

$$X' = \frac{170 - 150}{190 - 150} = \frac{20}{40} = 0.5$$

Altezza (cm)	Altezza Normalizzata
150	0.0
160	0.25
170	0.5
180	0.75
190	1.0

3.2. Z-Score Standardization (Standardizzazione)

Lo Z-Score trasforma i dati per avere media 0 e deviazione standard 1. Viene usato spesso per dati con distribuzione normale (gaussiana).

Formula

$$Z = \frac{X - \mu}{\sigma}$$

dove μ è la media e σ è la deviazione standard.

Esempio

Dataset: $10, 12, 14, 16, 18$

1. Calcoliamo la media: $\mu = 14$
2. Calcoliamo la deviazione standard: $\sigma \approx 2.83$

$$Z = \frac{10 - 14}{2.83} \approx -1.41$$

Valore Originale	Valore Z-Score
10	-1.41
12	-0.71
14	0.0
16	0.71
18	1.41

3.3. Normalizzazione tramite Logaritmo

Questo metodo è utile per dati con ampie variazioni e distribuzione asimmetrica, comprimendo le differenze.

Formula

$$X' = \log(X + 1)$$

Esempio

Dataset: $1, 10, 100, 1000, 10000$

Applichiamo la trasformazione logaritmica:

$$\log(100 + 1) \approx 2.004$$

Valore Originale	Valore Logaritmico
1	0.693
10	2.398
100	4.615
1000	6.908
10000	9.210

3.4. Normalizzazione Mediana e Intervallo Interquartile (IQR Scaling)

Questo metodo centra i dati sulla mediana e li scala usando l'intervallo interquartile (IQR). È utile quando ci sono outlier.

Formula

$$X' = \frac{X - \text{mediana}}{\text{IQR}}$$

Esempio

Dataset: $1, 2, 2, 3, 4, 5, 100$

1. Calcoliamo la mediana: 3
2. Calcoliamo IQR (Q3 - Q1): 3

Valore Originale	Valore Normalizzato (IQR)
1	-0.67
2	-0.33
3	0
4	0.33
5	0.67
100	32.33

3.5. Normalizzazione tramite Funzione Sigmoidale

Una funzione sigmoide restringe i valori tra 0 e 1, utile per modelli basati su probabilità.

Formula

$$X' = \frac{1}{1 + e^{-X}}$$

Esempio

Dataset: $-3, -1, 0, 1, 3$

Valore Originale	Valore Sigmoide
-3	0.047
-1	0.269
0	0.5
1	0.731
3	0.953

---

4. Quale Metodo Scegliere?

- **Min-Max Scaling**: utile per modelli che richiedono dati in un intervallo specifico, come le reti neurali.
- **Z-Score Standardization**: utile quando i dati seguono una distribuzione normale e servono media 0 e varianza 1.
- **Logaritmo**: utile per ridurre la varianza nei dati con ampi range.
- **Mediana e IQR**: consigliato quando sono presenti outlier.
- **Sigmoid**: utile per modelli di classificazione e output probabilistici.

5. Esercizi Pratici

1. Prendete un dataset con valori numerici e applicate sia la normalizzazione Min-Max che la standardizzazione Z-Score. Confrontate i risultati.
2. Provate la normalizzazione logaritmica su un dataset con ampie variazioni (es. popolazione di città, reddito).
3. Usate la normalizzazione mediana e IQR su un dataset contenente outlier e osservate come cambia la scala dei dati.

Ecco una serie di esercizi pratici basati su un dataset di temperature raccolte da un sensore di temperatura ambientale. In campo scientifico e applicativo, la normalizzazione è essenziale quando i dati raccolti dai sensori devono essere confrontati, combinati con altri dati, o utilizzati per modelli di previsione e analisi.

Dataset di Temperature

Immaginiamo di avere un sensore di temperatura che rileva le temperature ambientali ogni ora. Il dataset (in gradi Celsius) potrebbe essere il seguente:

Dataset di Temperature: $12, 15, 14, 13, 18, 22, 30, 25, 17, 19, 21, 16$

Esercizi Pratici di Normalizzazione

Esercizio 1: Normalizzazione Min-Max

Obiettivo: Trasformare le temperature in un intervallo $[0, 1]$ per poterle confrontare con dati provenienti da sensori che raccolgono in intervalli simili, come umidità o pressione (dati in percentuale o intervallo ridotto).

Passi:

1. Calcoliamo il valore minimo e massimo del dataset.
2. Applichiamo la formula di Min-Max Scaling:
$$T' = \frac{T - T_{\min}}{T_{\max} - T_{\min}}$$

Applicazione:

- **Minimo**: $T_{\min} = 12$
- **Massimo**: $T_{\max} = 30$

Esempio di calcolo per la temperatura di 22 gradi:

$$T' = \frac{22 - 12}{30 - 12} = \frac{10}{18} \approx 0.56$$

Temperatura Originale	Temperatura Normalizzata
12	0.0
15	0.17
14	0.11
13	0.06
18	0.33
22	0.56
30	1.0
25	0.72
17	0.28
19	0.39
21	0.5
16	0.22

Nota: Questa normalizzazione è utile per modelli che usano input in range simili, come reti neurali o algoritmi di clustering, poiché evita che variabili su scale diverse influenzino in modo eccessivo il modello.

Esercizio 2: Z-Score Standardization

Obiettivo: Standardizzare le temperature per ottenere una distribuzione con media 0 e deviazione standard 1, utile per analisi statistiche e confronto con altre variabili standardizzate, ad esempio nel monitoraggio di anomalie.

Passi:

1. Calcoliamo la media e la deviazione standard del dataset.
2. Applichiamo la formula dello Z-Score:

$$Z = \frac{T - \mu}{\sigma}$$

Applicazione:

- **Media**: $\mu = 18.25$
- **Deviazione Standard**: $\sigma \approx 5.06$

Esempio di calcolo per la temperatura di 22 gradi:

$$Z = \frac{22 - 18.25}{5.06} \approx 0.74$$

Temperatura Originale	Z-Score
12	-1.24
15	-0.64
14	-0.84
13	-1.04
18	-0.05
22	0.74
30	2.27
25	1.33
17	-0.25
19	0.15
21	0.54
16	-0.45

Nota: Questo tipo di standardizzazione è utile per analizzare anomalie, ad esempio quando una temperatura si discosta dalla media e potrebbe indicare un malfunzionamento o un fenomeno particolare.

Esercizio 3: Normalizzazione Logaritmica

Obiettivo: Applicare una trasformazione logaritmica per ridurre l'impatto di valori estremi. Questa tecnica è spesso usata in contesti scientifici per dati distribuiti in modo asimmetrico o con variazioni molto ampie, come misurazioni in geofisica o climatologia.

Passi:

1. Applichiamo la formula di trasformazione logaritmica:
$$T' = \log(T + 1)$$

Applicazione: Per esempio, calcoliamo la normalizzazione logaritmica della temperatura di 22 gradi:

$$T' = \log(22 + 1) \approx 3.14$$

Temperatura Originale	Temperatura Logaritmica
12	2.56
15	2.77
14	2.71
13	2.64
18	2.89
22	3.14
30	3.43
25	3.26
17	2.83
19	2.94
21	3.09
16	2.79

Nota: La normalizzazione logaritmica è utile per ridurre l’impatto di variazioni molto grandi e dare maggiore peso a variazioni più piccole. È adatta per analisi di serie storiche, come i dati meteo, in cui le temperature possono avere variazioni annuali molto grandi.

Esercizio 4: Normalizzazione Mediana e Intervallo Interquartile (IQR)

Obiettivo: Usare la mediana e l'intervallo interquartile per normalizzare i dati, riducendo l'impatto di eventuali valori estremi. Questo è utile nel monitoraggio ambientale dove possono esserci variazioni improvvise o errori di misurazione.

Passi:

1. Calcoliamo la mediana e l'intervallo interquartile (IQR).
2. Applichiamo la formula:

$$T' = \frac{T - \text{mediana}}{\text{IQR}}$$

Applicazione:

- **Mediana**: 17.5
- **Intervallo Interquartile (IQR)**: 6.5 (Q3 = 21, Q1 = 14.5)

Per esempio, per la temperatura di 22 gradi:

$$T' = \frac{22 - 17.5}{6.5} \approx 0.69$$

Temperatura Originale	Normalizzazione Mediana/IQR
12	-0.85
15	-0.38
14	-0.46
13	-0.69
18	0.08
22	0.69
30	1.92
25	1.15
17	-0.08
19	0.23
21	0.54
16	-0.23

Nota: Questa normalizzazione è molto utile quando il dataset contiene outlier o variazioni insolite, poiché la mediana è meno sensibile agli outlier rispetto alla media.

Conclusione

Questi metodi di normalizzazione permettono di adattare i dati a modelli e analisi diverse. In ambito scientifico, la normalizzazione aiuta a:

- Migliorare la qualità dei dati in modelli di machine learning e algoritmi di analisi.
- Ridurre l'effetto di valori estremi, come sbalzi improvvisi di temperatura.
- Confrontare i dati di sensori diversi, poiché molti dispositivi hanno scale di misura differenti (es. pressione, temperatura, umidità).

Scegliere il metodo giusto permette di ottenere informazioni più accurate e confrontabili, migliorando la qualità dell'analisi scientifica e applicativa dei dati.

Immaginiamo di avere un sistema che raccoglie dati ambientali in tempo reale da vari dispositivi, ognuno dei quali misura una variabile differente. Gli strumenti che stiamo monitorando sono:

- **Sensore di Temperatura** (in gradi Celsius)
- **Sensore di Umidità** (in percentuale %)
- **Sensore di Pressione** (in hPa - ettopascal)

I dati raccolti in un dato momento potrebbero essere simili ai seguenti:

Temperatura (°C)	Umidità (%)	Pressione (hPa)
18	45	1015
22	55	1020
30	60	1012
16	65	1018
25	70	1017
20	75	1014
15	80	1022
28	85	1016
26	90	1019
19	95	1013

Ogni variabile ha una scala di misura differente:

- **Temperatura**: va da circa 15°C a 30°C.
- **Umidità**: varia dal 45% al 95%.
- **Pressione**: varia da 1012 hPa a 1022 hPa.

Se volessimo applicare un'analisi che richiede che tutte le variabili siano sulla stessa scala, potremmo usare diversi metodi di normalizzazione. Qui applichiamo la **normalizzazione Min-Max** per portare tutte le variabili in un range comune, $[0, 1]$, per un confronto uniforme.

Passi della Normalizzazione Min-Max

1. **Identifichiamo i valori minimi e massimi per ogni variabile.**
    - Temperatura: min = 15, max = 30
    - Umidità: min = 45, max = 95
    - Pressione: min = 1012, max = 1022
2. **Applichiamo la formula Min-Max per ciascun valore**:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Calcoli e Tabella Normalizzata

Esempio di calcolo

Per il valore di temperatura a 18°C:

$$T' = \frac{18 - 15}{30 - 15} = \frac{3}{15} = 0.2$$

Per il valore di umidità a 45%:

$$H' = \frac{45 - 45}{95 - 45} = \frac{0}{50} = 0$$

Per il valore di pressione a 1015 hPa:

$$P' = \frac{1015 - 1012}{1022 - 1012} = \frac{3}{10} = 0.3$$

Tabella dei Valori Normalizzati

Temperatura (°C)	Umidità (%)	Pressione (hPa)	Temperatura Normalizzata	Umidità Normalizzata	Pressione Normalizzata
18	45	1015	0.2	0	0.3
22	55	1020	0.47	0.2	0.8
30	60	1012	1	0.3	0
16	65	1018	0.07	0.4	0.6
25	70	1017	0.67	0.5	0.5
20	75	1014	0.33	0.6	0.2
15	80	1022	0	0.7	1
28	85	1016	0.87	0.8	0.4
26	90	1019	0.73	0.9	0.7
19	95	1013	0.27	1	0.1

Interpretazione dei Dati Normalizzati

Con questa normalizzazione, tutte le variabili sono portate su una scala da 0 a 1:

- **Temperatura**: Ora varia da 0 (15°C) a 1 (30°C).
- **Umidità**: Ora varia da 0 (45%) a 1 (95%).
- **Pressione**: Ora varia da 0 (1012 hPa) a 1 (1022 hPa).

Vantaggi della Normalizzazione su Multisensori

Questa normalizzazione rende più facile:

- **Confrontare variabili diverse**: poiché ora tutte le misure sono su una scala comune, un valore alto o basso ha lo stesso significato tra le variabili.
- **Combinare i dati in modelli**: ad esempio, modelli di machine learning possono ora trattare tutte le variabili allo stesso modo, senza che una variabile con una scala più ampia (es. umidità) influenzi il modello più delle altre.
- **Identificare anomalie e trend**: normalizzando, diventa più semplice rilevare valori anomali o picchi improvvisi nei dati di una specifica variabile.

Per trovare la correlazione tra i dati raccolti dai diversi sensori (temperatura, umidità e pressione), possiamo utilizzare metodi statistici che ci aiutano a capire se e come le variabili sono collegate tra loro. La correlazione ci dice se, al variare di una variabile, l'altra tende a variare in modo simile (correlazione positiva), in modo opposto (correlazione negativa), o in modo indipendente (assenza di correlazione).

Metodi per Calcolare la Correlazione

1. **Coefficiente di Correlazione di Pearson**
   - È il metodo più comune per calcolare la correlazione tra due variabili continue. Misura la linearità tra due variabili.
   - Varia tra -1 e +1:
     - **+1** indica una correlazione positiva perfetta (quando una variabile aumenta, anche l'altra aumenta in modo proporzionale).
     - **-1** indica una correlazione negativa perfetta (quando una variabile aumenta, l'altra diminuisce in modo proporzionale).
     - **0** indica assenza di correlazione lineare.
   - **Formula**:

$$r = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum(X - \overline{X})^2 \cdot \sum(Y - \overline{Y})^2}}$$

   - Questo metodo è particolarmente utile quando le variabili sono normalmente distribuite e hanno una relazione lineare.
2. **Correlazione di Spearman**
   - Misura la correlazione tra variabili ordinali o non linearmente correlate. Utilizza i ranghi delle variabili piuttosto che i loro valori effettivi.
   - Anche Spearman restituisce un valore tra -1 e +1, con gli stessi significati di Pearson.
   - È più robusto contro gli outlier rispetto a Pearson.
3. **Correlazione di Kendall**
   - È un'altra misura non parametrica della correlazione, basata su ordinamenti e confronti di coppie di dati.
   - È utile per dataset piccoli o con variabili ordinali.

Esempio Pratico con il Coefficiente di Pearson

Usiamo il nostro dataset di temperature, umidità e pressione. Per calcolare la correlazione tra queste variabili, possiamo costruire una **matrice di correlazione** che mostra i coefficienti di correlazione tra tutte le coppie di variabili.

Consideriamo il nostro dataset di esempio:

Temperatura (°C)	Umidità (%)	Pressione (hPa)
18	45	1015
22	55	1020
30	60	1012
16	65	1018
25	70	1017
20	75	1014
15	80	1022
28	85	1016
26	90	1019
19	95	1013

Calcolo della Matrice di Correlazione

Applicando la formula di Pearson, possiamo ottenere una matrice di correlazione come questa (valori di esempio):

	Temperatura	Umidità	Pressione
Temperatura	1	0.82	-0.45
Umidità	0.82	1	-0.3
Pressione	-0.45	-0.3	1

Interpretazione dei Risultati

- **Temperatura e Umidità**: Correlazione positiva di 0.82, indica che all'aumentare della temperatura, anche l'umidità tende ad aumentare.
- **Temperatura e Pressione**: Correlazione negativa di -0.45, suggerisce che all'aumentare della temperatura, la pressione tende a diminuire.
- **Umidità e Pressione**: Correlazione negativa di -0.3, meno pronunciata ma ancora indicativa di una leggera relazione inversa.

Ecco i passaggi per calcolare la correlazione tra variabili in Excel:

Passo 1: Inserisci i Dati

1. Apri un nuovo foglio di Excel.
2. Inserisci i dati in tre colonne separate:
    - Colonna A: Temperatura (°C)
    - Colonna B: Umidità (%)
    - Colonna C: Pressione (hPa)

Ecco come dovrebbe apparire il foglio:

A	B	C
Temperatura (°C)	Umidità (%)	Pressione (hPa)
18	45	1015
22	55	1020
30	60	1012
16	65	1018
25	70	1017
20	75	1014
15	80	1022
28	85	1016
26	90	1019
19	95	1013

Passo 2: Usa la Funzione di Correlazione di Excel

1. **Seleziona una cella vuota** dove vuoi vedere il risultato della correlazione tra Temperatura e Umidità (ad esempio, D2).
2. Scrivi la formula:

   excel Copy code

   ```excel
   =CORRELA(A2:A11, B2:B11)
   ```

   Questa formula calcola la correlazione tra le colonne A (Temperatura) e B (Umidità).
3. **Premi Invio** per ottenere il risultato.
4. Ripeti la stessa operazione per le altre coppie di variabili:
   - **Temperatura e Pressione**: `=CORRELA(A2:A11, C2:C11)`
   - **Umidità e Pressione**: `=CORRELA(B2:B11, C2:C11)`

Passo 3: Costruisci la Matrice di Correlazione

Posiziona i risultati in una tabella per ottenere una matrice di correlazione come questa:

	Temperatura	**Umidità**	**Pressione**
Temperatura	1	Correlazione T-U	Correlazione T-P
Umidità	Correlazione T-U	1	Correlazione U-P
Pressione	Correlazione T-P	Correlazione U-P	1

Interpretazione dei Risultati

La tabella mostrerà i coefficienti di correlazione tra le variabili:

- **Valori vicini a +1** indicano una correlazione positiva.
- **Valori vicini a -1** indicano una correlazione negativa.
- **Valori vicini a 0** indicano assenza di correlazione lineare.

Nota Finale

Excel rende semplice ottenere la correlazione tra variabili, e questa tecnica può essere applicata anche a dataset più grandi. La correlazione tra variabili di multisensori può fornire preziose informazioni per ottimizzare il monitoraggio ambientale, comprendere come le variabili sono collegate e individuare possibili anomalie.

Dati di Esempio

Inserisci questi dati in Excel:

A	B	C
Temperatura	**Umidità**	**Pressione**
18	45	1015
22	55	1020
30	60	1012
16	65	1018
25	70	1017
20	75	1014
15	80	1022
28	85	1016
26	90	1019
19	95	1013

Passi per Calcolare la Correlazione

Step 1: Calcolare la Correlazione tra Temperatura e Umidità

1. Seleziona una cella vuota (ad esempio, `E2`) per calcolare la correlazione tra Temperatura e Umidità.
2. Inserisci la formula:

   excel   Copy code

   ```excel
   =CORRELA(A2:A11, B2:B11)
   ```
3. Premi **Invio**. Il risultato rappresenta il coefficiente di correlazione tra Temperatura e Umidità.

Step 2: Calcolare la Correlazione tra Temperatura e Pressione

1. Seleziona un'altra cella (ad esempio, `E3`).
2. Inserisci la formula:

   excel   Copy code

   ```excel
   =CORRELA(A2:A11, C2:C11)
   ```
3. Premi **Invio**. Otterrai la correlazione tra Temperatura e Pressione.

Risultati e Matrice di Correlazione

Dovresti ottenere una tabella simile a questa (con valori di esempio):

	Temperatura	Umidità	Pressione
Temperatura	1	0.82	-0.45
Umidità	0.82	1	-0.3
Pressione	-0.45	-0.3	1

Interpretazione dei Risultati

- **Temperatura e Umidità**: Correlazione positiva (ad esempio, 0.82), significa che al crescere della temperatura, anche l'umidità tende ad aumentare.
- **Temperatura e Pressione**: Correlazione negativa (ad esempio, -0.45), suggerisce che al crescere della temperatura, la pressione tende a diminuire.
- **Umidità e Pressione**: Correlazione negativa meno pronunciata (ad esempio, -0.3), indica che umidità e pressione non sono fortemente correlate.

Questi passaggi ti permettono di visualizzare rapidamente la correlazione tra diverse variabili in Excel, aiutandoti a comprendere le relazioni tra i dati raccolti da più sensori.

Per calcolare l'**Intervallo Interquartile (IQR)** in Excel, segui questi passaggi. L'IQR è la differenza tra il terzo quartile (Q3) e il primo quartile (Q1), e rappresenta la dispersione centrale del dataset.

Dati di Esempio

Supponiamo di avere questo dataset in Excel nella colonna **A**:

A
Temperatura
18
22
30
16
25
20
15
28
26
19

Passi per Calcolare l'IQR

1. **Calcola il Primo Quartile (Q1)**
   - In una cella vuota (ad esempio, `B2`), inserisci la formula:

     excel   Copy code

     ```excel
     =QUARTILE(A2:A11, 1)
     ```

   - Questa formula restituisce il valore del **primo quartile (Q1)** del dataset.

2. **Calcola il Terzo Quartile (Q3)**
   - In un'altra cella (ad esempio, `B3`), inserisci la formula:

     excel   Copy code

     ```excel
     =QUARTILE(A2:A11, 3)
     ```

   - Questa formula restituisce il valore del **terzo quartile (Q3)** del dataset.

3. **Calcola l'Intervallo Interquartile (IQR)**
    - In un'altra cella (ad esempio, `B4`), inserisci la formula per sottrarre Q1 da Q3:

      excel Copy code

      ```
      =B3 - B2
      ```

    - Questo calcolo restituisce il valore dell'**Intervallo Interquartile (IQR)**.

Esempio Completo

Supponiamo che i calcoli diano i seguenti valori:

- **Q1**: 18
- **Q3**: 26
- **IQR**: $26 - 18 = 8$

Quindi, l'**Intervallo Interquartile (IQR)** per questo dataset è **8**.

Interpretazione dell'IQR

L'IQR è utile per comprendere la dispersione dei dati centrali e identificare i valori anomali. Qualsiasi dato che si trova al di fuori del range $[Q1 - 1.5 \times \mathrm{IQR}, Q3 + 1.5 \times \mathrm{IQR}]$ è considerato un **outlier**.

L'**Intervallo Interquartile (IQR)** è una misura statistica della **dispersione** o **variabilità** di un dataset, che rappresenta l'estensione dei dati centrali, o meglio, la distribuzione dei valori che si trovano tra il primo e il terzo quartile.

Come si Calcola l'IQR

L'IQR si calcola con la differenza tra:

- **Terzo Quartile (Q3)**: il valore al di sotto del quale si trova il 75% dei dati.
- **Primo Quartile (Q1)**: il valore al di sotto del quale si trova il 25% dei dati.

La formula è:

$$\text{IQR} = Q3 - Q1$$

Interpretazione dell'IQR

L'IQR ci dice quanto sono distribuiti i dati centrali, eliminando l'influenza degli **outlier** (valori anomali o estremi) che potrebbero distorcere la variabilità. Un IQR più ampio indica una maggiore dispersione tra i dati centrali, mentre un IQR più stretto indica che i dati sono più concentrati attorno alla mediana.

Utilità dell'IQR

L'IQR è molto utilizzato in statistica e analisi dei dati per:

- **Rilevare outlier**: Un dato è considerato outlier se si trova al di fuori dell'intervallo:

  $$[Q1 - 1.5 \times \text{IQR},\ Q3 + 1.5 \times \text{IQR}]$$

  Questo intervallo è noto come **range di non-outlier**.

- **Analisi della variabilità centrale**: L'IQR è utile in contesti in cui i dati hanno outlier, poiché non è influenzato dai valori estremi. È quindi una misura di dispersione più robusta rispetto alla deviazione standard per dataset asimmetrici o con valori molto variabili.

Esempio

Supponiamo di avere un dataset di temperature:

$$[15, 18, 21, 22, 25, 28, 30, 33, 35]$$

1. **Calcolo di Q1 (primo quartile)**: Il 25% dei dati si trova al di sotto di **18**.
2. **Calcolo di Q3 (terzo quartile)**: Il 75% dei dati si trova al di sotto di **30**.
3. **Calcolo dell'IQR**:

$$\text{IQR} = 30 - 18 = 12$$

Quindi, l'IQR è **12**, e rappresenta l'ampiezza dell'intervallo centrale dove si concentra il 50% dei dati.

Questo concetto è particolarmente utile nelle analisi esplorative, soprattutto per valutare la variabilità e la presenza di outlier in un dataset.