

Metodi Statistici per le decisioni

2025-2026

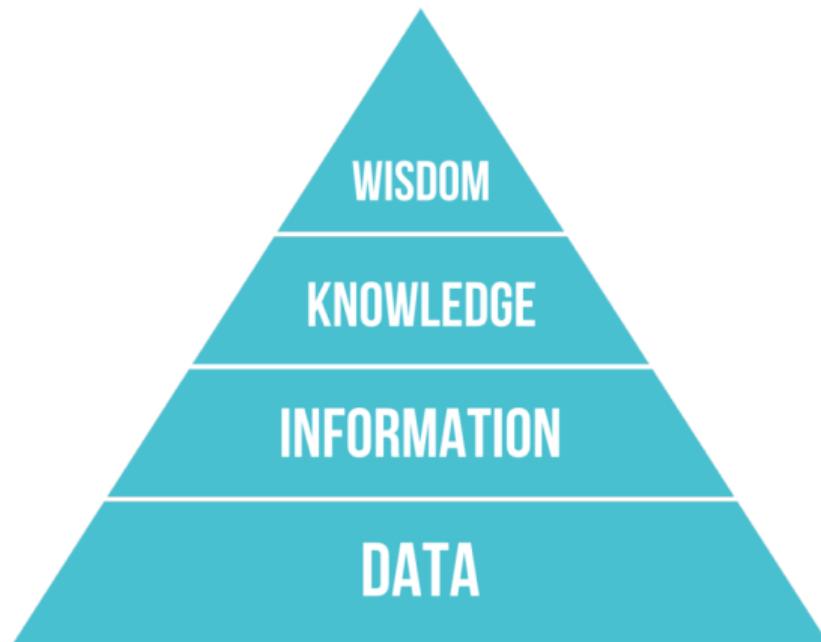
Vincenzo Nardelli



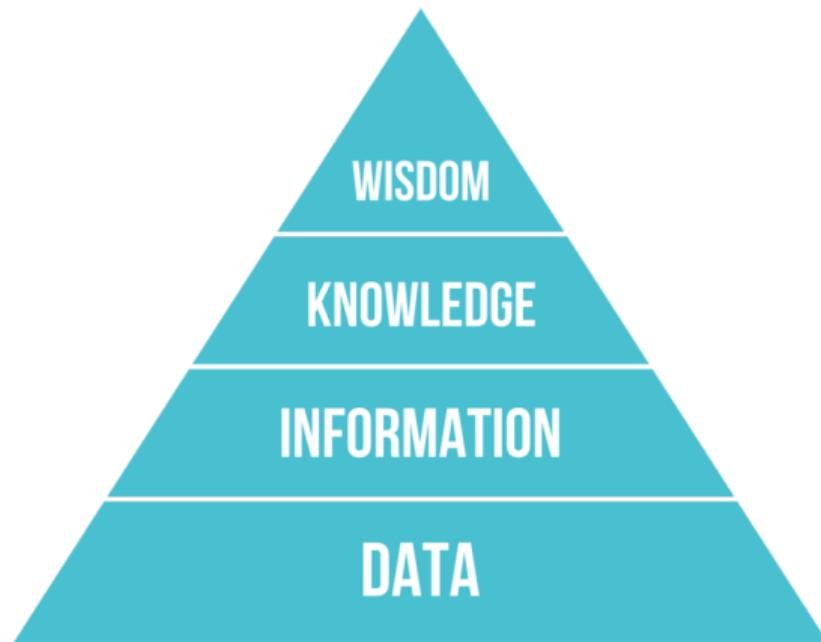
vincenzo.nardelli@unicatt.it



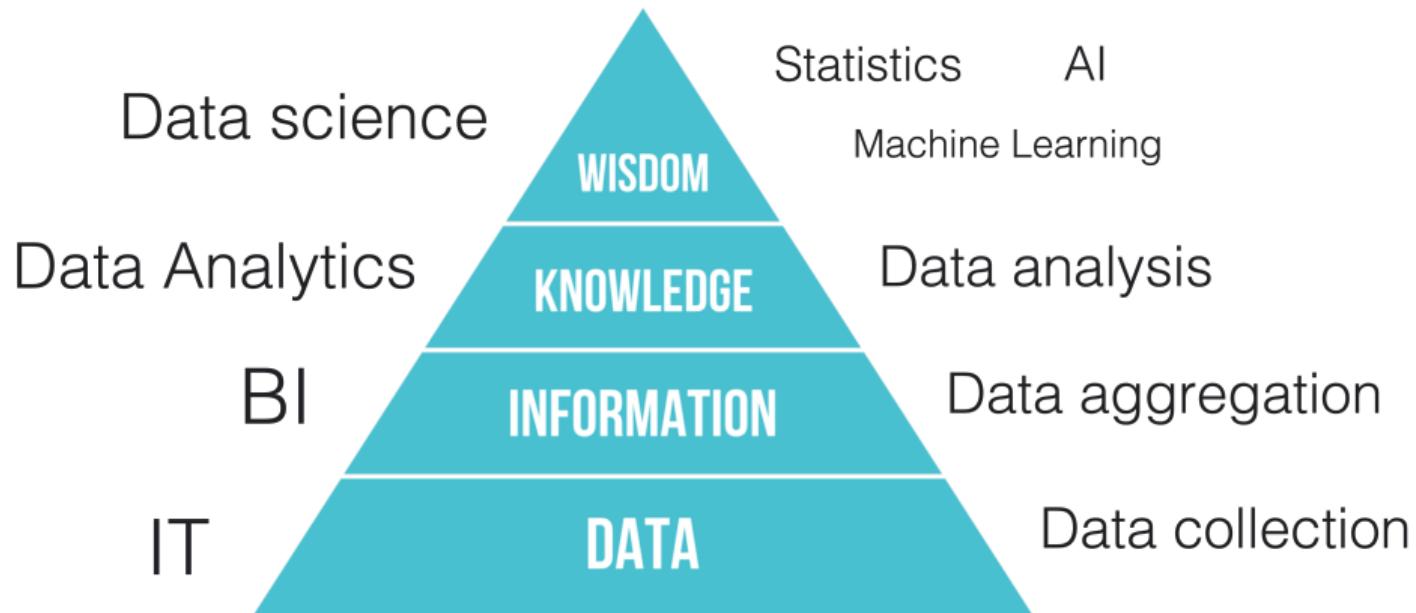
Framework DIKW



Framework DIKW



Framework DIKW ext



Framework DIKW ext



Rischi Operativi nell'Uso di Modelli di AI/Machine Learning

- ▶ **Campione Distorto:** previsioni errate a causa di dati non rappresentativi.
- ▶ **Bias nei Dati:** discriminazioni o pregiudizi riflessi nelle decisioni.
- ▶ **Assenza di Verifica nei Modelli:** mancanza di monitoraggio che può portare a decisioni non allineate con il contesto.
- ▶ **Scarsa Interpretabilità dei Modelli:** difficoltà nel comprendere le logiche interne dei modelli.
- ▶ **Etica e Privacy:** rischi legati alla violazione della privacy e alla gestione responsabile dei dati.

Campione Distorto

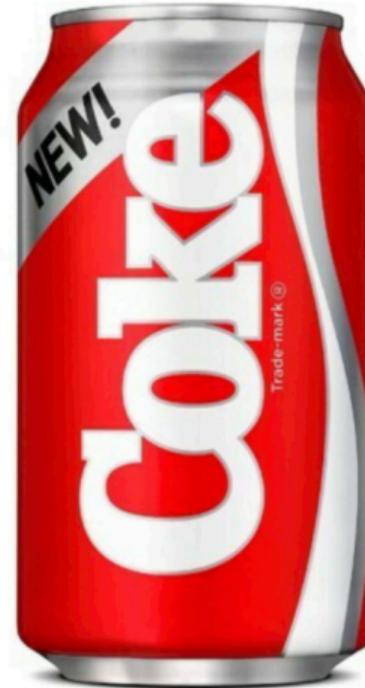
Coca-Cola e il Lancio di New Coke (1985) - Settore Bevande

Problema Operativo: Fallimento del lancio di un nuovo prodotto con perdite milionarie.

Causa: Metodo di test che non rifletteva il comportamento reale dei consumatori.

Errori metodologici: "Sip test" non rappresentativi del consumo reale, ignorato il campione critico (10-12% fortemente contrario), mancata considerazione del contesto e dell'attaccamento emotivo al brand.

Fonte: The Coca-Cola Company



Bias nei Dati

Amazon e il Sistema di Recruiting (2018) - Settore Risorse Umane

Problema Operativo: Discriminazione di genere nei processi di selezione.

Causa: Pregiudizi storici nei dati di addestramento.

Fonte: Reuters - Amazon scraps secret AI recruiting tool that showed bias against women



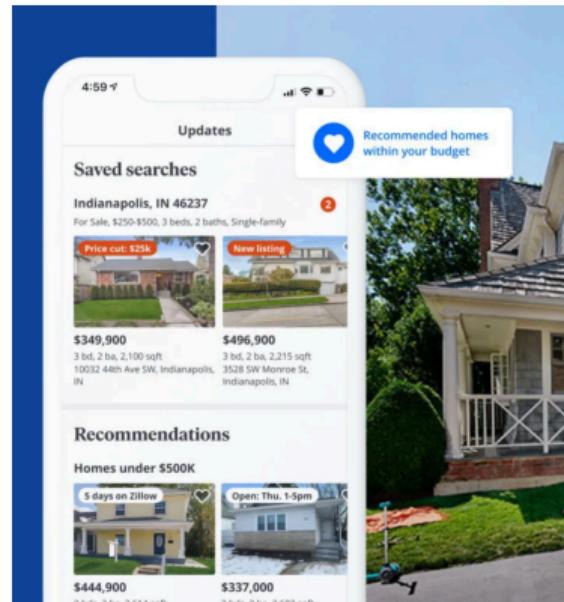
Assenza di Verifica nei Modelli

Zillow (2021) - Settore Immobiliare

Problema Operativo: Perdita di milioni di dollari nel settore immobiliare.

Causa: Modello non allenato per gestire le fluttuazioni di mercato.

Fonte: CNN - Zillow's home-buying debacle shows how hard it is to use AI to value real estate



Scarsa Interpretabilità dei Modelli

Apple Card e Limiti di Credito (2019) - Settore Finanziario

Problema Operativo: Disparità nei limiti di credito assegnati.

Causa: Modello non trasparente e difficile da interpretare.

Fonte: BBC - Apple's 'sexist' credit card investigated by US regulator



Target e la profilazione dei clienti (2012) - Settore Retail e Marketing

Problema Operativo: Violazione della privacy dei clienti e conseguente reazione negativa.

Causa: Profilazione avanzata dei clienti senza consenso esplicito.

Fonte: Forbes - How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did



Misure Base della Statistica Descrittiva

Le misure di statistica descrittiva sono utili per riassumere, descrivere e comprendere i dati. Esistono vari tipi di misure, tra cui:

- ▶ **Indici di Centralità** – indicano la tendenza centrale dei dati.
- ▶ **Indici di Variabilità** – misurano quanto i dati si discostano dalla tendenza centrale.
- ▶ **Analisi Bivariata** – esplora le relazioni tra due variabili.

Indici di Centralità

Gli indici di centralità forniscono un valore rappresentativo che descrive il “centro” di un insieme di dati.

Media

La media è la somma di tutti i valori divisa per il numero di valori.

$$\text{Media} = \frac{1}{n} \sum_{i=1}^n x_i$$

Indici di Centralità

Gli indici di centralità forniscono un valore rappresentativo che descrive il “centro” di un insieme di dati.

Media

La media è la somma di tutti i valori divisa per il numero di valori.

$$\text{Media} = \frac{1}{n} \sum_{i=1}^n x_i$$

Mediana

La mediana è il valore che si trova nel mezzo di un insieme di dati ordinati.

Nota: Se il numero dei dati è dispari, la mediana è il valore centrale; se è pari, è la media dei due valori centrali.

Indici di Centralità

Gli indici di centralità forniscono un valore rappresentativo che descrive il “centro” di un insieme di dati.

Media

La media è la somma di tutti i valori divisa per il numero di valori.

$$\text{Media} = \frac{1}{n} \sum_{i=1}^n x_i$$

Mediana

La mediana è il valore che si trova nel mezzo di un insieme di dati ordinati.
Nota: Se il numero dei dati è dispari, la mediana è il valore centrale; se è pari, è la media dei due valori centrali.

Moda

La moda è il valore che compare più frequentemente in un insieme di dati.

Nota: I dati possono avere una o più mode, o nessuna.

Indici di Variabilità

Gli indici di variabilità mostrano quanto i valori dei dati si discostano dal valore centrale.

Varianza

La varianza è la media dei quadrati delle deviazioni dalla media.

$$\text{Varianza} = \frac{1}{n} \sum_{i=1}^n (x_i - \text{Media})^2$$

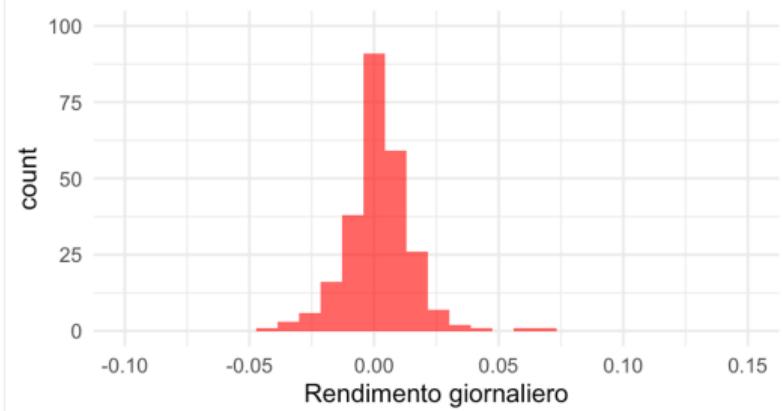
Deviazione Standard

La deviazione standard è la radice quadrata della varianza e rappresenta la dispersione media rispetto alla media.

$$\text{Deviazione Standard} = \sqrt{\text{Varianza}}$$

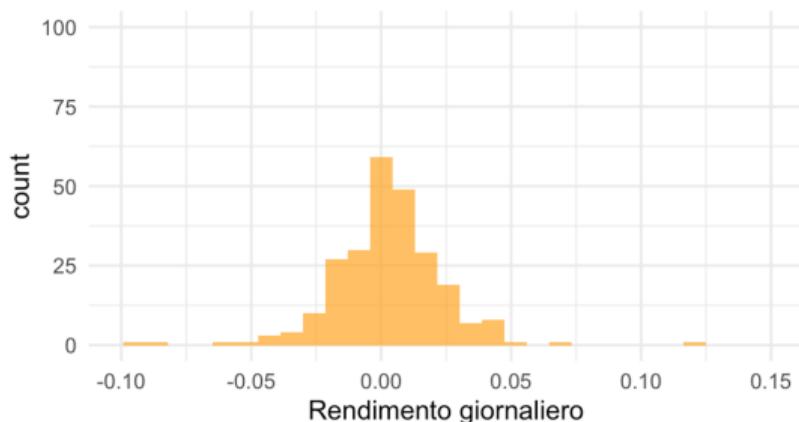
Esempio Indici univariati

Rendimenti giornalieri delle azioni Ferrari ed Unicredit in Borsa Italiana nel 2023



Media: 0.17% - Dev. Std. 1.3%

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.046658	-0.005166	0.001038	0.001722	0.008665	0.073043



Media: 0.26% - Dev. Std. 2.1%

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.090631	-0.009232	0.002348	0.002699	0.013447	0.122944

Covarianza

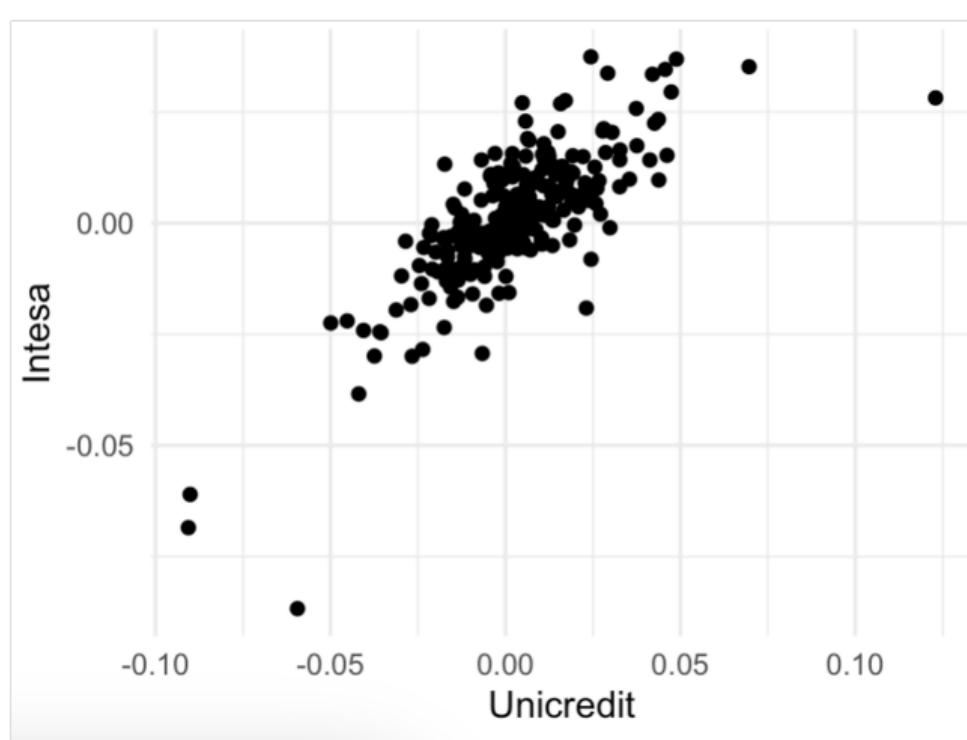
- ▶ Definizione: La covarianza misura la direzione della relazione lineare tra due variabili. Una covarianza positiva indica che le variabili aumentano insieme, mentre una covarianza negativa indica una relazione inversa.
- ▶ Formula:

$$\text{Covarianza} = \frac{1}{n} \sum_{i=1}^n (x_i - \text{Media}_x)(y_i - \text{Media}_y)$$

- ▶ Esempio: Una covarianza positiva tra reddito e spesa può indicare che all'aumentare del reddito aumenta anche la spesa.

Esempio Analisi bivariata

Rendimenti giornalieri delle azioni Intesa ed Unicredit in Borsa Italiana nel 2023



Analisi Bivariata

Correlazione

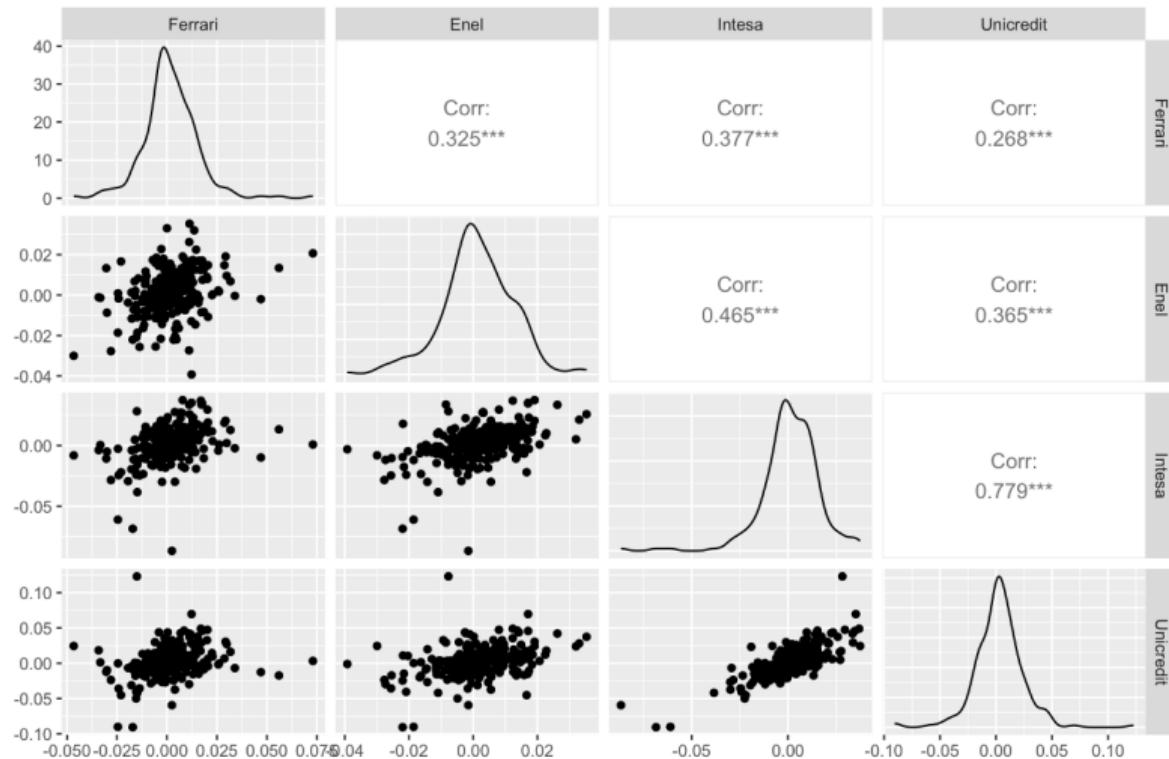
- ▶ Definizione: La correlazione è una misura normalizzata della covarianza, che assume valori tra -1 e +1 e indica sia la forza che la direzione della relazione tra due variabili.
- ▶ Formula:

$$\text{Correlazione} = \frac{\text{Covarianza}(X, Y)}{\text{Deviazione Standard}_X \cdot \text{Deviazione Standard}_Y}$$

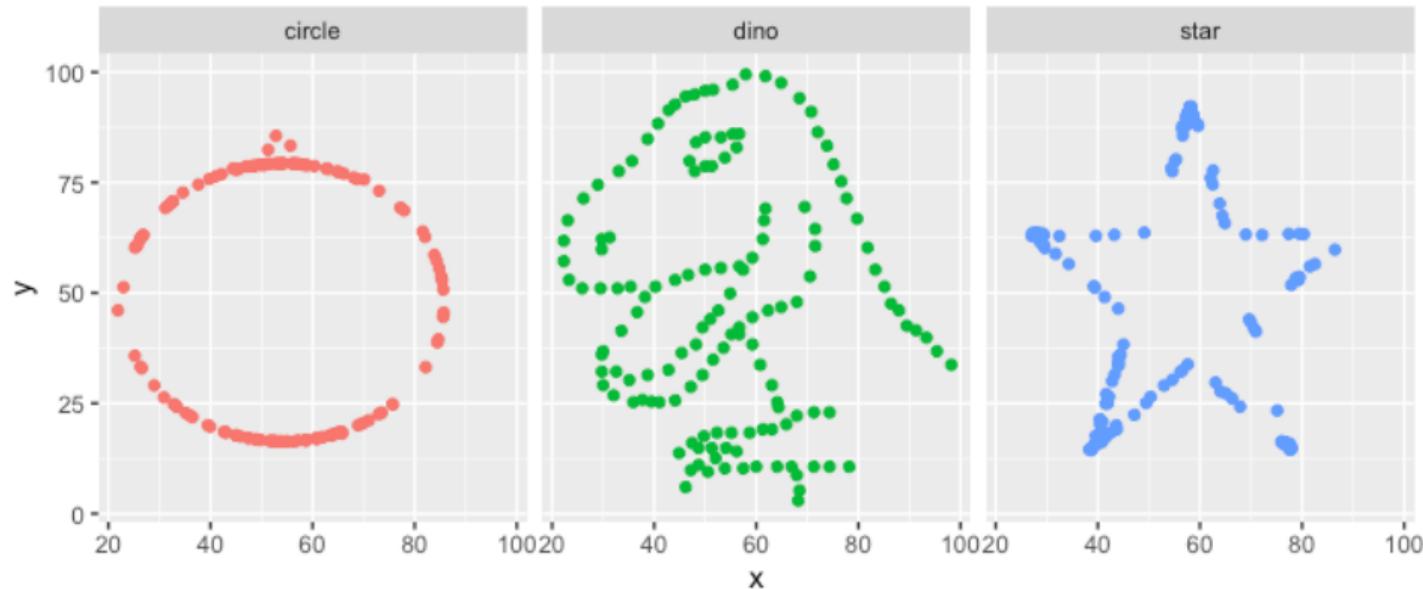
- ▶ Nota: Una correlazione vicina a +1 indica una forte relazione positiva, una vicina a -1 una forte relazione negativa, e una vicina a 0 nessuna relazione.

Esempio Analisi bivariata

Rendimenti giornalieri delle azioni in Borsa Italiana nel 2023

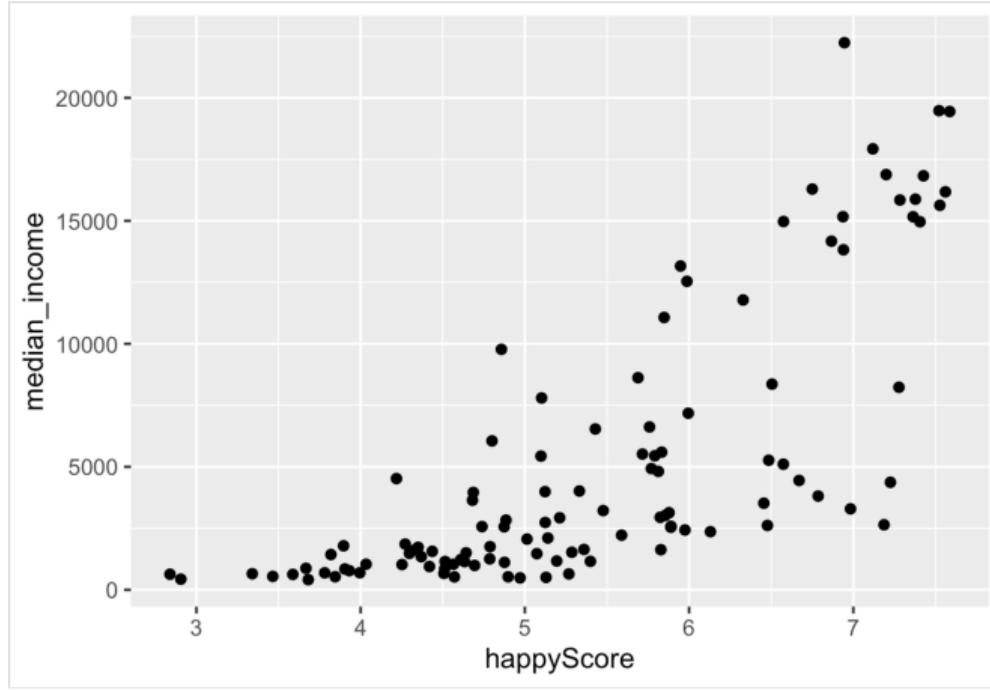


Perchè non basta la statistica descrittiva?

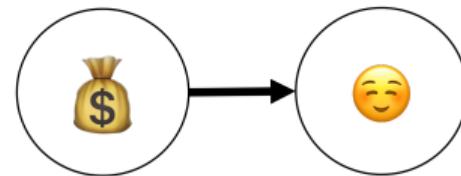


Dataset	Mean X	Mean Y	Std Dev X	Std Dev Y	Corr X-Y
Circle	54.3	47.8	16.8	26.9	-0.0683
Dino	54.3	47.8	16.8	26.9	-0.0645
Star	54.3	47.8	16.8	26.9	-0.0630

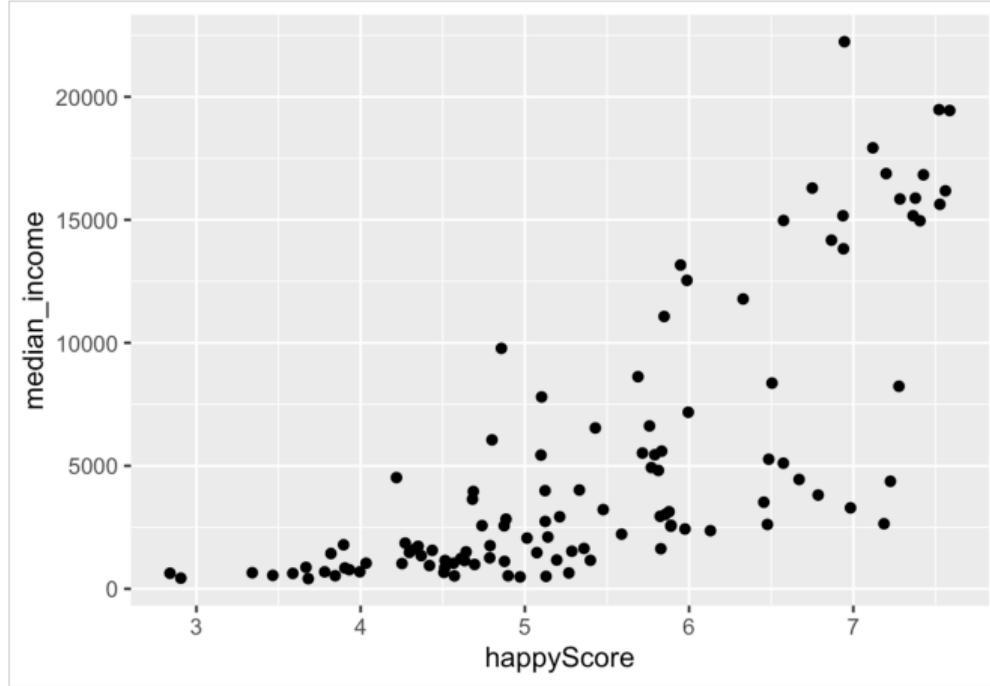
I soldi fanno la felicità?



Correlazione = 0.76

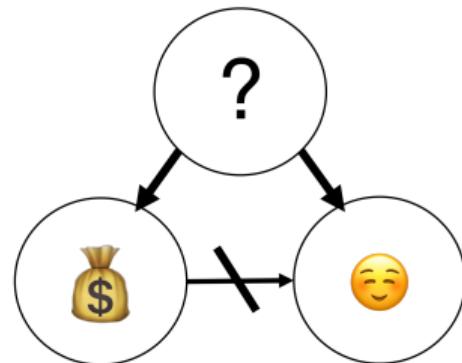


Correlazione Spuria



Correlazione = 0.76

Correlation is not causation!



Apprendimento Statistico vs Machine Learning

- ▶ **Machine Learning:** Enfasi su grandi applicazioni e accuratezza delle predizioni.
- ▶ **Apprendimento Statistico:** Enfasi su modelli, interpretabilità e precisione.
- ▶ Distinzione sfumata.

Il Problema dell'Apprendimento Supervisionato

- ▶ Misura dell'esito Y (detto anche variabile dipendente, risposta, obiettivo).
- ▶ Vettore di p predittori X (input, regressori, covariate, caratteristiche, variabili indipendenti).
- ▶ **Problema di regressione:** Y quantitativo (es. prezzo, pressione sanguigna).
- ▶ **Problema di classificazione:** Y prende valori in un insieme finito (es. sopravvissuto/morto, cifre da 0 a 9).

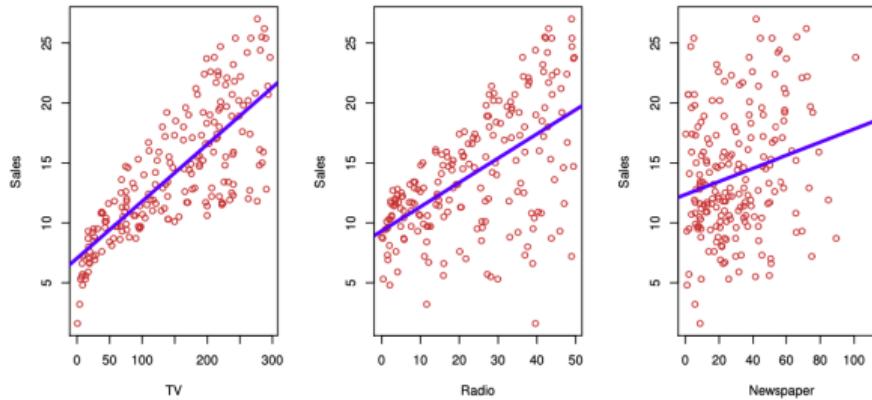
Obiettivi

- ▶ Predire accuratamente i casi di test non visti.
- ▶ Comprendere quali input influenzano l'esito e in che modo.
- ▶ Valutare la qualità delle predizioni e inferenze.

Apprendimento Non Supervisionato

- ▶ Nessuna variabile di esito, solo predittori misurati su un insieme di campioni.
- ▶ Obiettivo più indefinito: trovare gruppi di campioni o combinazioni lineari con maggiore variazione.
- ▶ Utile come fase di pre-elaborazione per l'apprendimento supervisionato.

Che cos'è l'Apprendimento Statistico?



- ▶ Sono mostrati i grafici di *Vendite* rispetto a *TV*, *Radio* e *Giornale*, con una linea di regressione lineare in blu adattata separatamente per ciascuno.
- ▶ Possiamo prevedere le vendite usando questi tre parametri?
- ▶ Forse possiamo fare di meglio usando un modello

$$\text{Vendite} = f(\text{TV}, \text{Radio}, \text{Giornale})$$

Notazione

- ▶ Qui *Vendite* è una risposta o obiettivo che desideriamo prevedere. Ci riferiamo genericamente alla risposta come Y .
- ▶ TV è una caratteristica, o input, o predittore; lo chiamiamo X_1 .
- ▶ Allo stesso modo, chiamiamo *Radio* come X_2 , e così via.
- ▶ Possiamo fare riferimento al vettore degli input collettivamente come

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

- ▶ Ora scriviamo il nostro modello come

$$Y = f(X) + \epsilon$$

dove ϵ cattura errori di misurazione e altre discrepanze.

A cosa serve $f(X)$?

- ▶ Con una buona funzione f possiamo fare previsioni di Y in nuovi punti $X = x$.
- ▶ Possiamo capire quali componenti di $X = (X_1, X_2, \dots, X_p)$ sono importanti per spiegare Y , e quali sono irrilevanti.
- ▶ A seconda della complessità di f , potremmo essere in grado di capire come ogni componente X_j di X influenza Y .