

Data Wrangling & Analisi Statistica

L'essenziale per i nostri dati — e dove torna nei tre casi di oggi

Questa scheda riassume i due concetti chiave della mattinata e li collega subito ai tre casi pratici: prenotazioni Feratel (Daniele), presenze provinciali (Marco) e campagne social (marketing). Tienila accanto mentre lavori con l'AI.

Glossario — termini e abbreviazioni

Una breve guida ai termini tecnici usati in questo documento, per seguire facilmente anche senza essere esperti.

Termine	Significato
Dato grezzo	Dato non ancora pulito, così come esce dal sistema.
Data Wrangling	Le 4 fasi: scoperta, trasformazione, convalida, pubblicazione.
Pulizia dati	Correggere errori, valori mancanti, incoerenze e outlier.
Strutturazione	Cambiare forma/schema dei dati.
Join / Union	Join = unire tabelle per colonne; Union = impilare righe.
Normalizzazione	Ridurre dati ridondanti; la denormalizzazione li riunisce per l'analisi.
Outlier	Valore anomalo, molto diverso dagli altri.
Descrittiva / Inferenziale	Descrivere il campione vs generalizzare alla popolazione.
Media / Mediana / Moda	Indici del valore tipico; la mediana resiste agli outlier.
Dispersione	Quanto i dati sono 'sparsi' attorno al centro (dev. standard, range).
Correlazione vs Causalità	Muoversi insieme NON implica un rapporto di causa-effetto.

1 - Data Wrangling: preparare il dato in 4 fasi

Il data wrangling è il processo (iterativo) che trasforma dati grezzi in dati pronti per un'analisi affidabile. Quattro fasi:

- **1. Scoperta:** capire i dati rispetto all'obiettivo: cosa pulire, come strutturare, cosa manca.
- **2. Trasformazione:** il cuore del processo — strutturazione, normalizzazione/denormalizzazione, pulizia, arricchimento.
- **3. Convalida:** verificare coerenza, qualità e conformità con regole ripetibili.
- **4. Pubblicazione:** rendere disponibile il dataset finale con metadati e documentazione (per replicarlo).

Due principi: il processo è iterativo (si torna indietro a raffinare) e ogni scelta va documentata, così è ripetibile il mese dopo.

Dove torna nei nostri casi

Attività di wrangling	Cosa significa	Nei nostri casi
Pulizia	Correggere errori, togliere righe inutili e outlier, gestire valori mancanti	Daniele: escludere 'Cancellazione' e 'Prenotaz. richiesta', uniformare CAP/città scritti in modo diverso
Strutturazione	Cambiare forma/schema dei dati; Join (colonne) e Union (righe)	Marco: la colonna FORMULA che separa Campeggio/Extralberghiero è strutturazione; impilare più mesi = Union
Arricchimento	Aggiungere dati esterni o di altri sistemi per dare senso all'analisi	Marco: unire la consistenza (letti) per il tasso di occupazione; Marketing: incrociare la spesa adv con stagione/meteo
Convalida + Pubblicazione	Regole di qualità + documentazione del procedimento	Marco: le regole di aggregazione per la privacy = convalida; il prompt riutilizzabile = documentazione

2 - Analisi statistica: l'essenziale

Due famiglie: la statistica DESCRITTIVA riassume il campione (tabelle, grafici, indici) — è quella che useremo di più; la statistica INFERENZIALE generalizza alla popolazione. Gli indici di base:

- **Tendenza centrale:** media (somma/numero), mediana (valore centrale, NON influenzata dagli outlier), moda (valore più frequente).
- **Dispersione:** varianza, deviazione standard e range: quanto i dati sono dispersi attorno al centro.
- **Outlier e asimmetria:** valori anomali e distribuzioni inclinate (asimmetria) che possono distorcere le medie.
- **Correlazione $\neq$ causalità:** due variabili che si muovono insieme NON significa che una causi l'altra. Distinzione cruciale per il marketing.

Dove torna nei nostri casi

Concetto	Perché conta	Nei nostri casi
Media vs mediana	La media si lascia distorcere da pochi valori estremi	Daniele: soggiorno medio 3,79 notti, ma un gruppo molto lungo alza la media → guardare anche la mediana
Dispersione e outlier	Un picco isolato va spiegato, non nascosto	Marco: un mese con un evento genera un picco di presenze = outlier da annotare
Top N / classifiche	Sintetizza grandi tabelle in poche righe leggibili	Daniele e Marco: Top mercati / Top 10 mercati per presenze
Correlazione $\neq$ causalità	Evita conclusioni sbagliate sulle cause	Marketing: budget e vendite salgono d'estate, ma è la stagione; un CTR alto non significa più conversioni

3 - Cosa chiedo all'AI (3 frammenti che uniscono wrangling e statistica)

- "Descrivimi ogni colonna e segnala valori mancanti, incoerenti o anomali (outlier)." → scoperta + pulizia
- "Pulisci i dati secondo queste regole e, in una colonna a parte, spiega cosa hai corretto." → pulizia + documentazione
- "Dammi media, mediana e distribuzione, evidenzia le anomalie; mostra le correlazioni ma non dedurre cause." → statistica + causalità

Messaggio chiave

La qualità dell'analisi dipende dalla preparazione del dato. L'AI accelera enormemente pulizia, trasformazione e calcolo — ma le regole, le scelte e la documentazione restano nostre. È questo che rende un'analisi ripetibile e affidabile.