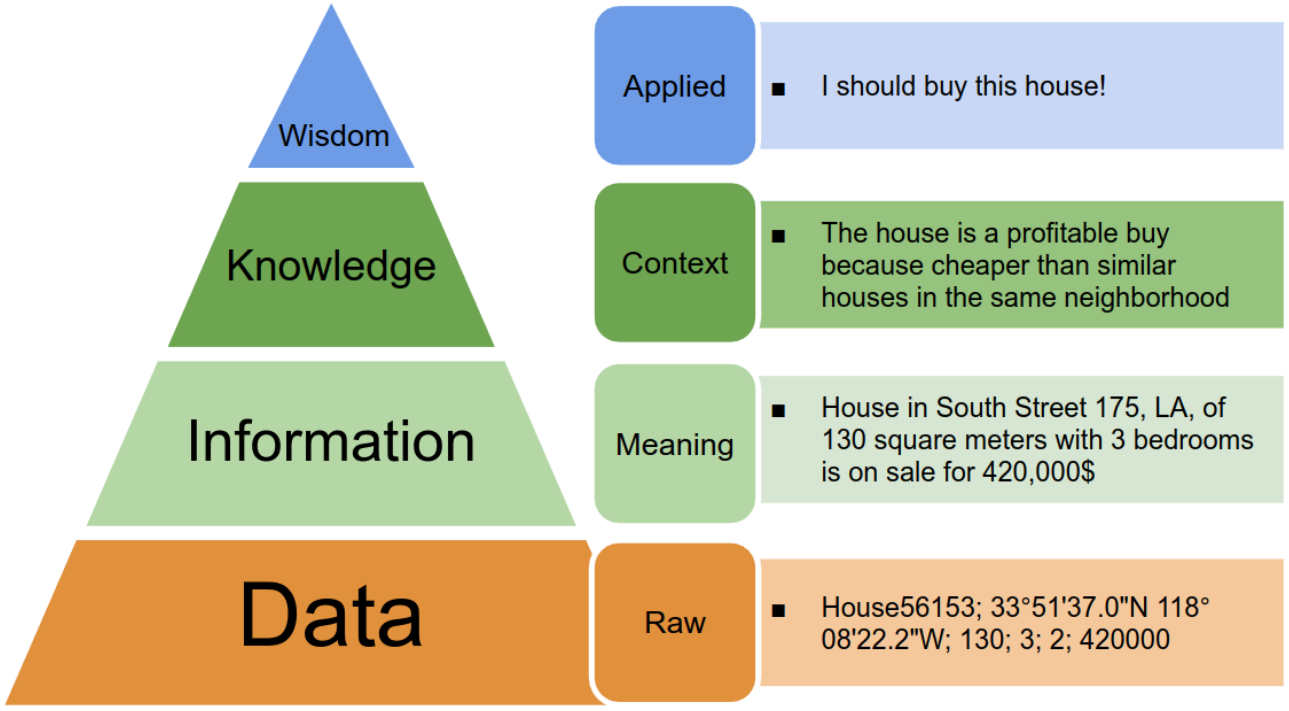
13. Data Analytics

La piramide DIKW

Tipicamente, l'informazione è definita in termini di dati, la conoscenza in termini di informazione, e la saggezza in termini di conoscenza.



In quale livello della piramide ci troviamo?

Fino a questo punto si è rimasti a un livello basso della piramide. I dati semi-strutturati possono fornire informazioni solo definendo manualmente query complesse. Il recupero dell'informazione, invece, consiste nella ricerca e nel ranking di informazioni già presenti sotto forma di documenti in linguaggio naturale, senza un'estrazione diretta dai dati.

Analisi dei dati

Lo scopo dell'analisi dei dati è estrarre informazioni di base da insiemi di dati. Le informazioni estratte possono essere di vario tipo. Si può trattare di informazioni riassuntive, come ad esempio la media calcolata su un insieme di valori numerici, oppure di informazioni di associazione, come la relazione tra due insiemi di valori, ad esempio il prezzo delle case rispetto ai metri quadrati. La base concettuale di queste attività è la statistica descrittiva.

Popolazione

Una popolazione è un insieme di oggetti di interesse. Esempi di popolazione possono essere tutte le case di Los Angeles, tutti gli studenti di un'università o tutti gli scontrini di un negozio di alimentari.

Record

Un record (o osservazione, caso) è una tupla di valori che caratterizza un elemento della popolazione.

Variabile

Una variabile (o campo, caratteristica) è il nome attribuito a un valore di un record e possiede un significato e un tipo comune per tutti i record della popolazione.

Tipi di variabile

Le variabili si possono classificare in base al tipo di valori che possono assumere. La distinzione più importante è tra variabili numeriche (quantitative), sulle quali è possibile applicare operazioni aritmetiche, e variabili categoriche (qualitative), sulle quali ciò non è possibile.

Un esempio di variabile numerica è il prezzo (ad esempio 420000), mentre un esempio di variabile categorica è la città (ad esempio Los Angeles, New York, Roma).

Le variabili numeriche possono essere discrete, se i valori sono numerabili, oppure continue, se sono il risultato di una misurazione continua.

Le variabili categoriche possono essere ordinali, se esiste un ordine naturale tra i valori possibili (ad esempio i voti scolastici: A, B, C, D), oppure nominali, in caso contrario (ad esempio i colori).

Dataset

L'insieme dei record (dataset) assume infine la forma di una singola tabella.

Statistica descrittiva

La statistica descrittiva fornisce indicatori sintetici per identificare, con un singolo valore, proprietà statistiche di una popolazione.

Con riferimento a una singola variabile, si considerano indicatori di centralità, come la media aritmetica, la moda e la mediana, e indicatori di variabilità, come la varianza e la deviazione standard.

Con riferimento a più variabili, si considerano la covarianza e la correlazione.

Centralità: media aritmetica

Sia X una variabile numerica del nostro dataset (non è possibile calcolare la media su valori categorici). Indichiamo con n il numero di record della popolazione e con X_i il valore del i-esimo record. La media aritmetica si calcola come:

$$\text{media} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Proprietà della media aritmetica

Supponiamo di avere un record con un dato mancante (ad esempio, il prezzo). È possibile mantenere il record senza alterare la media della variabile. Una soluzione consiste nel sostituire il dato mancante con la media calcolata per quella variabile. Inoltre, aggiungere un record con un valore pari alla media non modificherà la media aritmetica complessiva del dataset.

Centralità: mediana

Data una popolazione di valori ordinati (ad esempio, la colonna "SquareMeters" ordinata per valore):

$(30, 34, 37, 37, \ldots, 91, 91, 91, 91, 91, \ldots, 525, 600, 670)$

La mediana è il valore che si trova nella posizione centrale. Se n è il numero totale dei valori, la mediana corrisponde a $x_{n/2} = 91$.

Proprietà della mediana

La mediana è un indicatore robusto: anomalie come valori molto grandi o molto piccoli non influenzano significativamente il suo valore. Questo non è vero per la media, che è molto più sensibile alle anomalie (note anche come outlier).

Consideriamo il seguente esempio:

$2, 3, 3, 4, 5, 6, 6 \rightarrow$ Mediana = 4, Media = 4
$2, 3, 3, 4, 5, 6, 80 \rightarrow$ Mediana = 4, Media = 14.6

La mediana rimane 4 (valore in posizione centrale), mentre la media si sposta da 4 a 14.6!

Centralità: moda

Dato un insieme di valori di una variabile (ad esempio, la colonna “bedrooms”):

$(1, 2, 4, 2, 5, 3, 2, 2, 3, 4, 1, 3, 4, 2, 6, 2, 1, 3, 1, 2)$

Per prima cosa si contano le occorrenze di ciascun valore, ovvero la frequenza di quel valore. Ad esempio, “1” appare 4 volte, “2” appare 7 volte, “3” appare 4 volte, ecc. La moda è il valore con la frequenza più alta nell’insieme di osservazioni, cioè il valore “2” in questo esempio.

Proprietà della moda

A differenza della media e della mediana, la moda ha senso anche per dati categorici:

$\text{moda}(\text{Roma}, \text{Roma}, \text{Los Angeles}, \text{Albuquerque}) = \text{Roma}.$

In un sistema di voto (ad esempio, un insieme di classificatori diversi), la moda determina il risultato finale vincente. Come la mediana, è robusta rispetto ad anomalie, ma ha anche senso in assenza di un ordine lineare tra i valori possibili (ad esempio, punti nel piano).

Confronto tra indicatori di centralità

Consideriamo il seguente insieme di osservazioni della variabile “bedrooms”:

$(1, 2, 2, 3, 4, 7, 16)$

La **media aritmetica** (somma dei valori divisa per il numero di valori) è:

$\frac{1+2+2+3+4+7+16}{7} = 5$

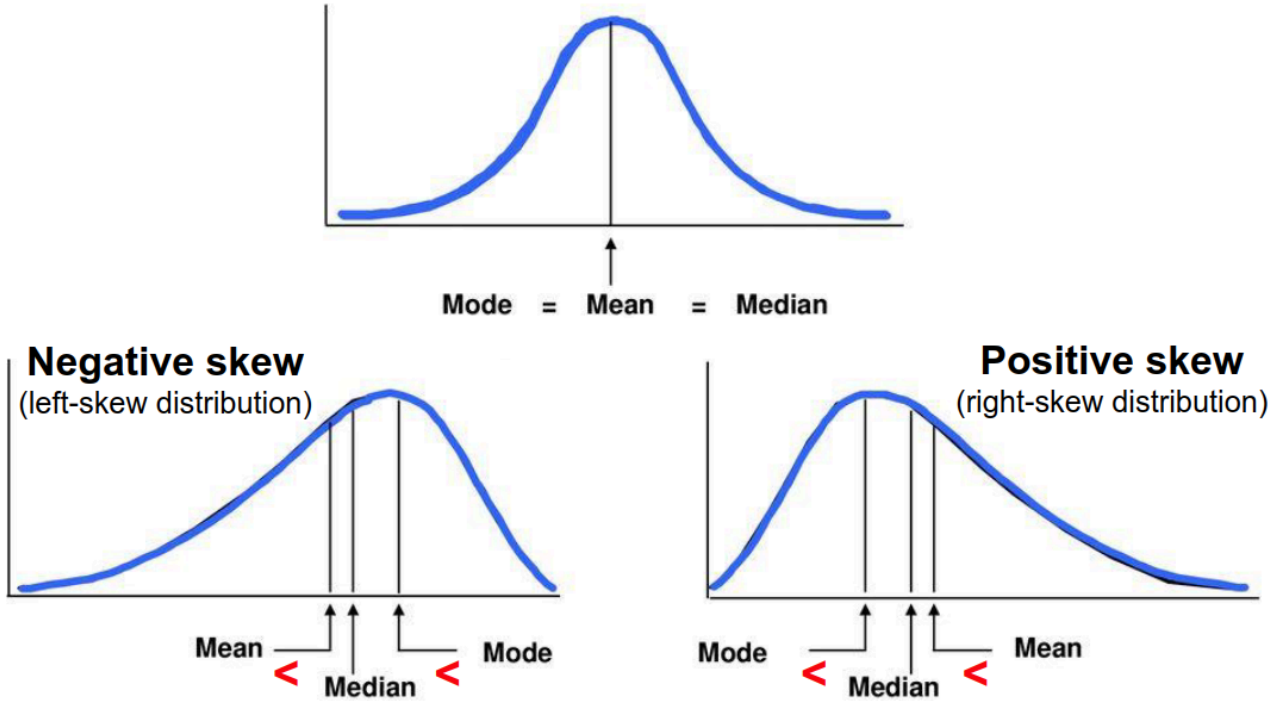
La **mediana** (valore centrale):

$(1, 2, 2, 3, 4, 7, 16)$ → mediana = 3

La **moda** (valore più frequente):

$(1, 2, 2, 3, 4, 7, 16)$ → moda = 2

Skewness (asimmetria): è la distorsione o deviazione dalla curva a campana simmetrica, cioè dalla distribuzione normale, presente in un insieme di dati.



Riassunto sulla centralità

Abbiamo esaminato tre indicatori di centralità: media aritmetica, mediana e moda. Tutti questi indicatori offrono un modo diverso di “riassumere” un insieme di valori in un singolo

valore sintetico.

La media aritmetica è utile anche per sostituire dati mancanti o errati senza alterare la distribuzione complessiva, ma il suo valore non è preso direttamente dai dati osservati ed è sensibile alle anomalie.

La mediana è un valore reale tra le osservazioni, è robusta alle anomalie ma richiede dati ordinali.

La moda è anch'essa un valore reale, il più frequente. È robusta alle anomalie, non necessita di dati ordinati e può essere applicata anche a variabili categoriche.

Variazione: scarto quadratico

Lo scarto quadratico misura la differenza tra ciascun valore x_i e la media delle osservazioni $\bar{x}$, ed è definito come:

$$\mathrm{dev} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Maggiore è la distanza dei valori dalla media, maggiore sarà lo scarto. Consideriamo il seguente esempio, in cui la media è 15:

$$\mathrm{dev}(4, 6, 10, 40) = (4-15)^2 + (6-15)^2 + (10-15)^2 + (40-15)^2 = 121 + 81 + 25 + 625 = 852$$

Variazione: varianza

Lo scarto quadratico è influenzato dal numero di osservazioni: più valori ci sono, maggiore tende ad essere lo scarto. La varianza (spesso rappresentata con s^2, σ^2 o Var) normalizza lo scarto quadratico dividendolo per il numero di osservazioni:

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n} \cdot \mathrm{dev}$$

Nel precedente esempio:

$$s^2(4, 6, 10, 40) = \frac{852}{4} = 213$$

Variazione: deviazione standard

Lo scarto quadratico e la varianza considerano la differenza elevata al quadrato tra i valori e la media, per ottenere differenze sempre non negative. Tuttavia, questo comporta valori elevati che non riflettono realisticamente la distanza stimata dalla media.

La **deviazione standard** (spesso rappresentata con s, σ o Stdev) è la radice quadrata della varianza:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

A differenza dello scarto quadratico e della varianza, la deviazione standard è espressa nelle stesse unità dei dati originali.

Confronto tra indicatori di variazione

Utilizziamo lo stesso esempio per confrontare i tre indicatori di variazione: $(4, 6, 10, 40)$

- Scarto quadratico: $\mathrm{dev}(4, 6, 10, 40) = 852$
- Varianza: $\mathrm{Var}(4, 6, 10, 40) = 213$
- Deviazione standard: $\sigma(4, 6, 10, 40) = 14{,}6$

Altri indicatori per variabile singola

- **Minimo (min):** è il valore minimo tra le osservazioni.
- **Massimo (max):** è il valore massimo tra le osservazioni.
- **Range:** è la differenza tra il valore massimo e il valore minimo.

Indicatori per più variabili

Nella statistica descrittiva, le misure di associazione permettono di descrivere la relazione tra due variabili, cercando connessioni tra di esse. Queste misure sono utili, ad esempio, per determinare se:

- Il prezzo delle case è associato ai metri quadrati.
- Il fumo è associato a malattie cardiache.
- Il budget pubblicitario è associato al numero di vendite.

Vedremo due misure principali:

- **Covarianza:** classifica il tipo di relazione tra due variabili.
- **Correlazione:** misura l'intensità della relazione, con valori compresi tra -1 e 1.

Misure di associazione: covarianza

La **covarianza** classifica il tipo di relazione tra due variabili X e Y. Più precisamente, essa rappresenta la **media dei prodotti degli scarti** dei valori rispetto alle rispettive medie aritmetiche.

La formula generale è:

$$\mathrm{cov}(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

Ogni termine nella somma corrisponde al **prodotto tra lo scarto del valore x_i rispetto alla media di X** e lo **scarto del valore y_i rispetto alla media di Y**. Il prodotto risultante sarà **alto** se entrambi gli scarti sono grandi **nello stesso momento**, in termini assoluti. In altre parole, la covarianza è elevata quando i valori di X e Y si discostano simultaneamente e nella stessa direzione dalle rispettive medie.

Significato della covarianza

L'idea alla base della covarianza è che la **somma dei prodotti degli scarti** tra due variabili sarà elevata se, ogni volta che X si discosta dalla propria media, anche Y si discosta dalla sua media **in modo coerente**.

Perché accade questo?

Se Y **non si discosta** dalla media, il suo scarto sarà basso, e quindi anche il prodotto degli scarti sarà basso.

Se Y **si discosta in modo casuale**, a volte in una direzione e a volte nell'altra, la somma dei prodotti tenderà ad annullarsi, perché i termini positivi e negativi si compenseranno.

Un aspetto importante da notare è che:

- Una **covarianza positiva** indica che X e Y si discostano dalle rispettive medie **nella stessa direzione** (cioè sono direttamente proporzionali).
- Una **covarianza negativa** indica che X e Y si discostano **in direzioni opposte** (cioè sono inversamente proporzionali).

Limiti della covarianza

Un problema della covarianza è che il suo valore **dipende dall'unità di misura**. Infatti:

- Se i valori sono grandi, anche la covarianza tenderà a essere grande, **anche se la relazione tra X e Y è debole**.
- Se i valori sono piccoli, la covarianza sarà piccola, **anche se la relazione tra X e Y è forte**.

Ad esempio, se si utilizza lo stesso insieme di prezzi ma si esprimono in **migliaia di dollari** anziché in dollari, la covarianza può **ridursi di un fattore 1000**, pur mantenendo invariata la relazione tra le variabili.

Correlazione

Per superare il problema legato all'**unità di misura** nella covarianza, si utilizza la **correlazione**. La correlazione risolve questo problema restituendo un risultato **indipendente dall'unità di misura**, poiché tiene conto delle **deviazioni standard** di X e Y.

Dividendo la covarianza per il prodotto delle deviazioni standard delle due variabili, si ottiene un valore **normalizzato**, sempre compreso tra -1 e 1:

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

Significato della correlazione

Un valore di correlazione vicino a -1 indica che le due variabili tendono a variare in **direzioni opposte** (sono **inversamente proporzionali**).

Un valore di correlazione vicino a 1 indica che le due variabili tendono a variare **nella stessa direzione** (sono **direttamente proporzionali**).

Un valore di correlazione vicino a 0 indica che le due variabili hanno **variazioni indipendenti** (almeno per quanto riguarda le relazioni lineari).

Scatter plot e regressione

Oltre al calcolo delle misure di associazione, è utile **visualizzare le coppie di valori** delle due variabili su un piano cartesiano XY. Ad esempio:

Metri quadrati	Prezzo
120	380.000
200	550.000
80	230.000
160	500.000
45	135.000
140	410.000

Mentre le misure di associazione ci indicano **se** esiste una relazione (in questo caso la **correlazione è 0.99**, quindi fortissima), i **modelli di regressione** stimano la **funzione effettiva** che lega le due variabili.

In questo esempio, la relazione stimata è:

$$\text{Prezzo}(\text{MetriQuadrati}) = 3000 \cdot \text{MetriQuadrati}$$

I modelli di regressione verranno trattati nella sezione dedicata al **machine learning**.

Misure di associazione: riepilogo

Analizzare due variabili diverse contemporaneamente aiuta a capire se esiste una qualche forma di relazione tra esse.

La **covarianza** ci dice **che tipo di relazione** esiste tra X e Y (diretta, inversa o nulla).

La **correlazione**, in aggiunta, **non è influenzata dall'unità di misura** delle variabili.

Attenzione. Queste misure di associazione sono affidabili **solo se la relazione è lineare**, ovvero se i punti formano approssimativamente una **linea retta** nel piano XY.

Una **correlazione pari a 0** non implica necessariamente **assenza di relazione**: potrebbe esistere una relazione forte ma **non lineare** tra X e Y.