I Dati Siamo Noi: Introduzione e Teoria

Niccolò Cibei, Julia Maria Wdowinska

Laurea Magistrale in Data Science for Economics, Università degli Studi di Milano
niccolo.cibei@studenti.unimi.it
juliamaria.wdowinska@studenti.unimi.it

CODINGgirls

Introduzione

Questo corso, denominato *I Dati Siamo Noi*, è parte integrante del progetto *Coding Girls*, promosso dalla Fondazione Mondo Digitale in collaborazione con la Società Italiana di Statistica.

Obiettivo del corso:
L'obiettivo di questo corso è quello di dotarvi delle competenze fondamentali nella **statistica**.

- Questo vi consentirà di comprendere e interpretare le informazioni numeriche e statistiche che incontrate nella vita quotidiana.
- Grazie a queste competenze, sarete in grado di prendere decisioni più informate e di sviluppare il vostro pensiero critico e logico.

Programma e Argomenti

1. 11/04, 14:00-16:00: Introduzione alla Teoria Statistica
   - Misure di tendenza centrale, dispersione e associazione.
   - Presentazione grafica dei dati.
2. 02/05, 14:00-16:00: Analisi dei Dati usando R
   - Introduzione all'uso dell'ambiente di programmazione R.
   - Esempi pratici di analisi dei dati.
3. 16/05, 14:00-16:00: Inizio del Lavoro di Gruppo
   - Formazione dei gruppi di lavoro e scelta dell'argomento.
   - Preparazione dei dati per l'analisi.
4. 21/05, 9:30-12:30: Presentazione del Lavoro di Gruppo
   - Evento presso la Sala Lauree in Via Conservatorio, 7.
   - Presentazione dei progetti.

CODINGgirls

Warm-Up Quiz su Dati Statistici

Prima di iniziare, faremo un **warm-up quiz** su alcuni dati statistici.

Immagine per Domanda 3

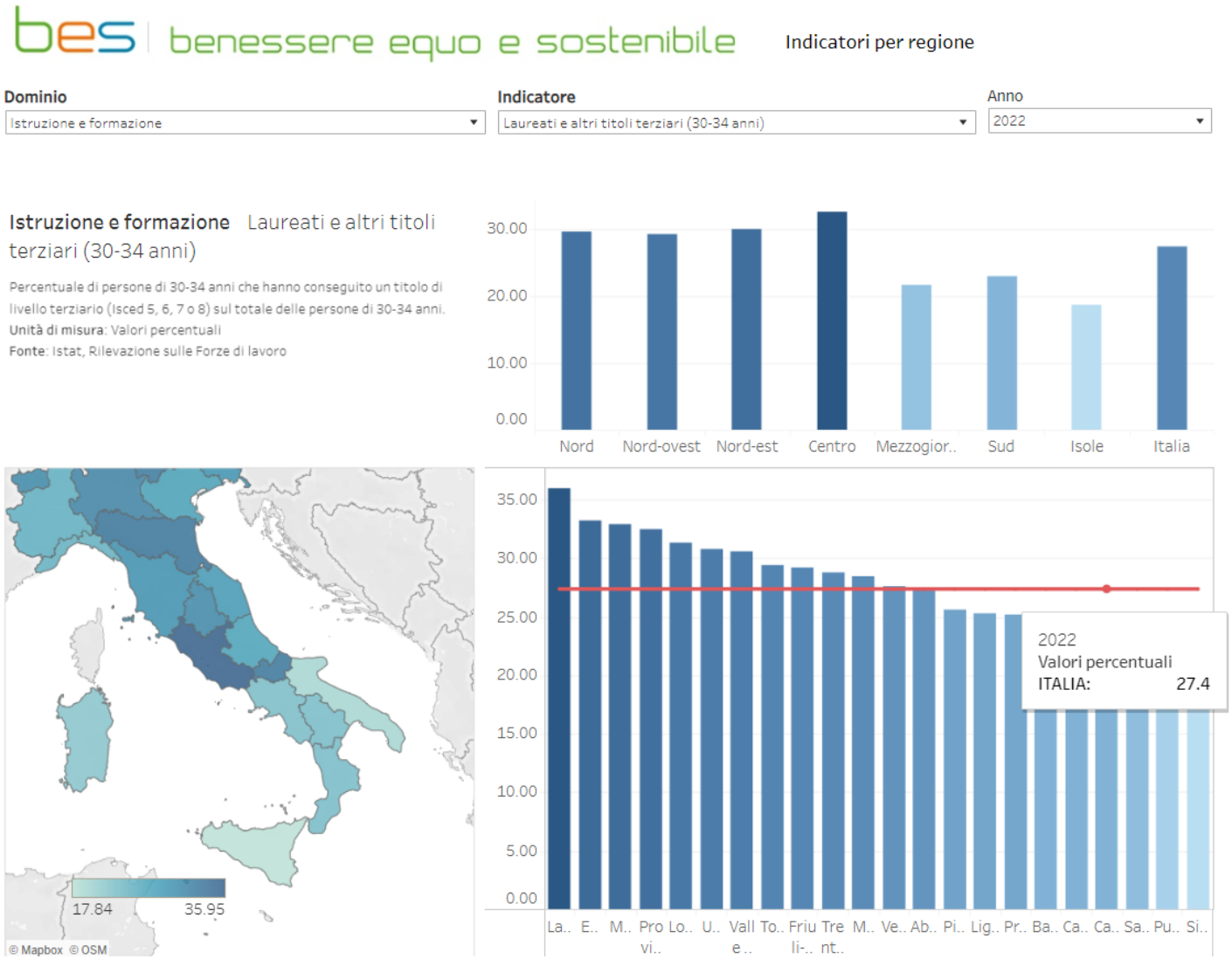


Immagine per Domanda 4

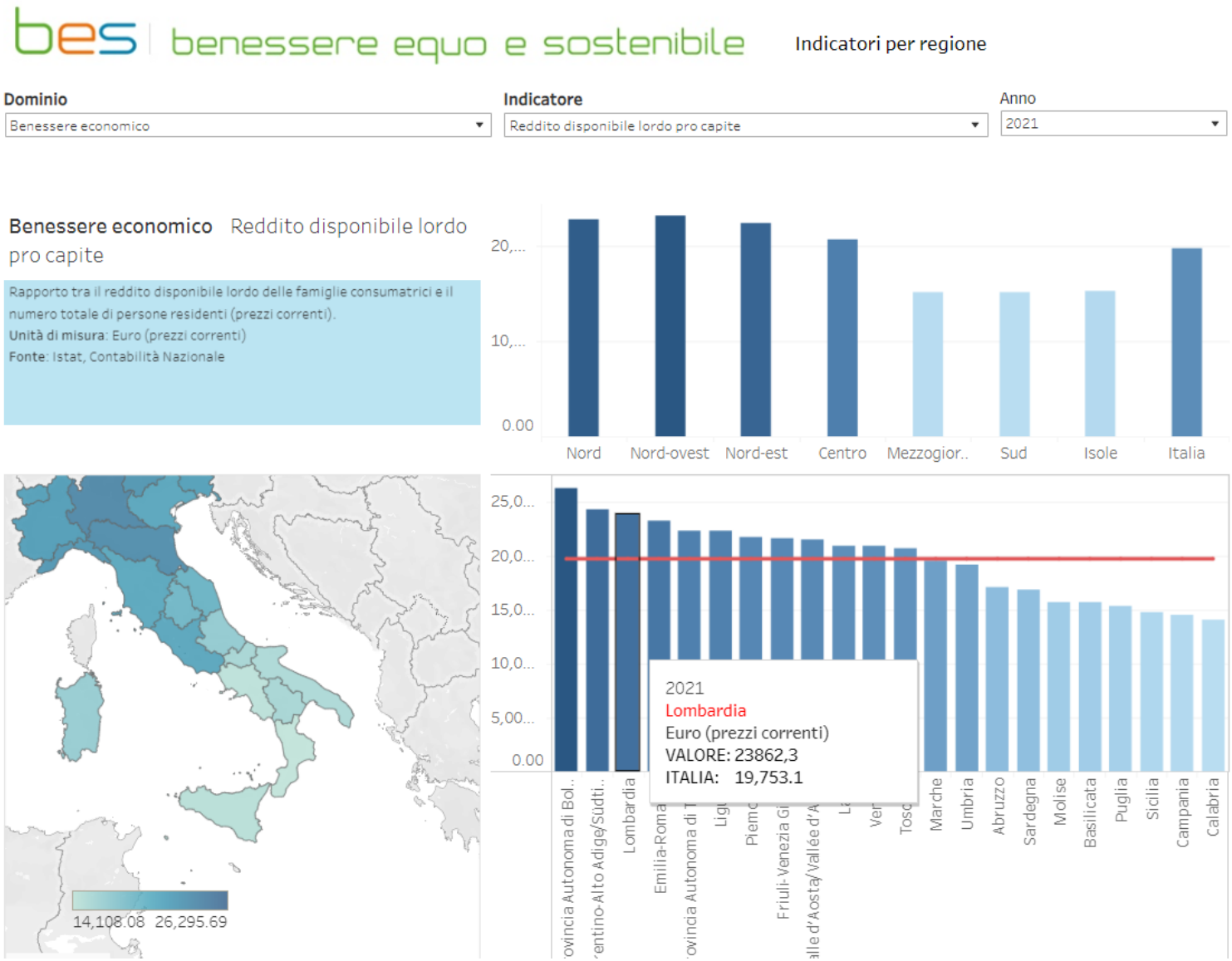


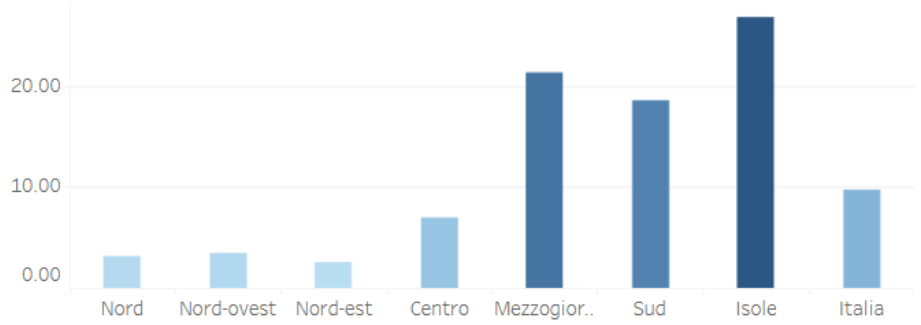
Immagine per Domanda 5

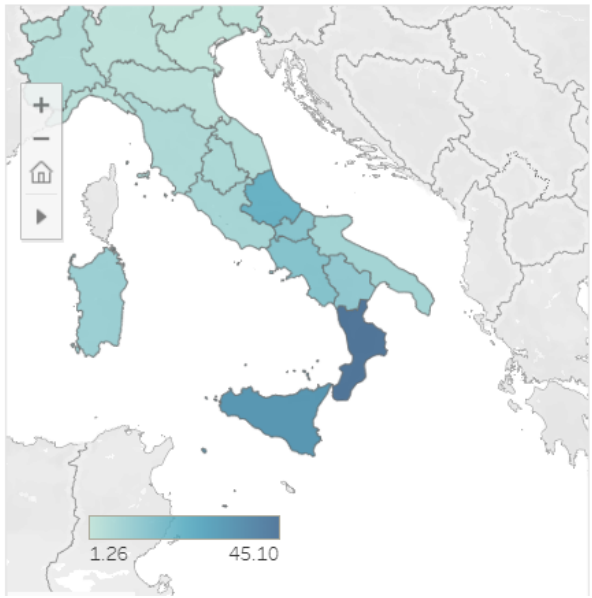
bes | benessere equo e sostenibile — Indicatori per regione

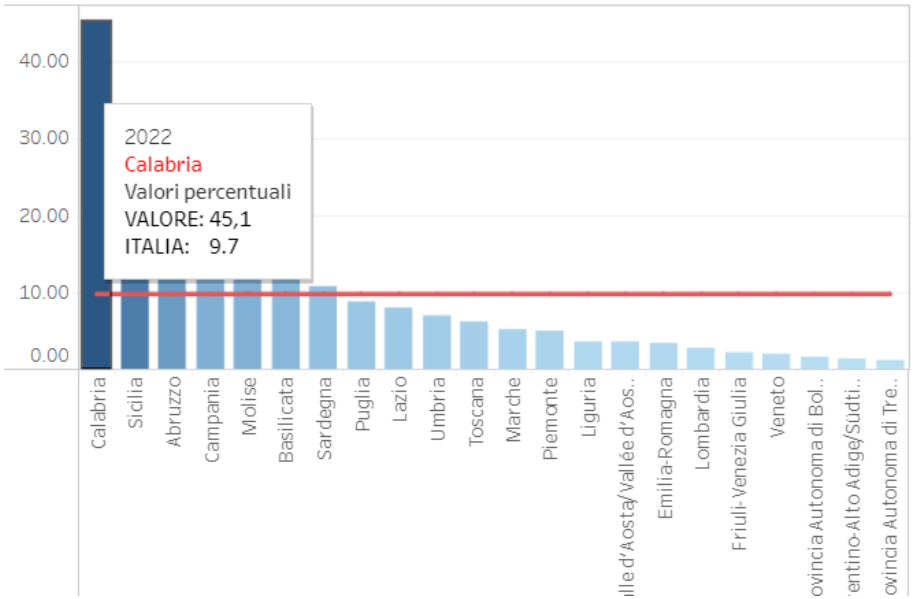
Dominio: Qualità dei servizi | **Indicatore**: Irregolarità nella distribuzione dell'acqua | **Anno**: 2022

Qualità dei servizi Irregolarità nella distribuzione dell'acqua

Percentuale di famiglie che denunciano irregolarità nell'erogazione dell'acqua.
Unità di misura: Valori percentuali
Fonte: Istat, Indagine Aspetti della vita quotidiana







CODINGgirls

Immagine per Domanda 6

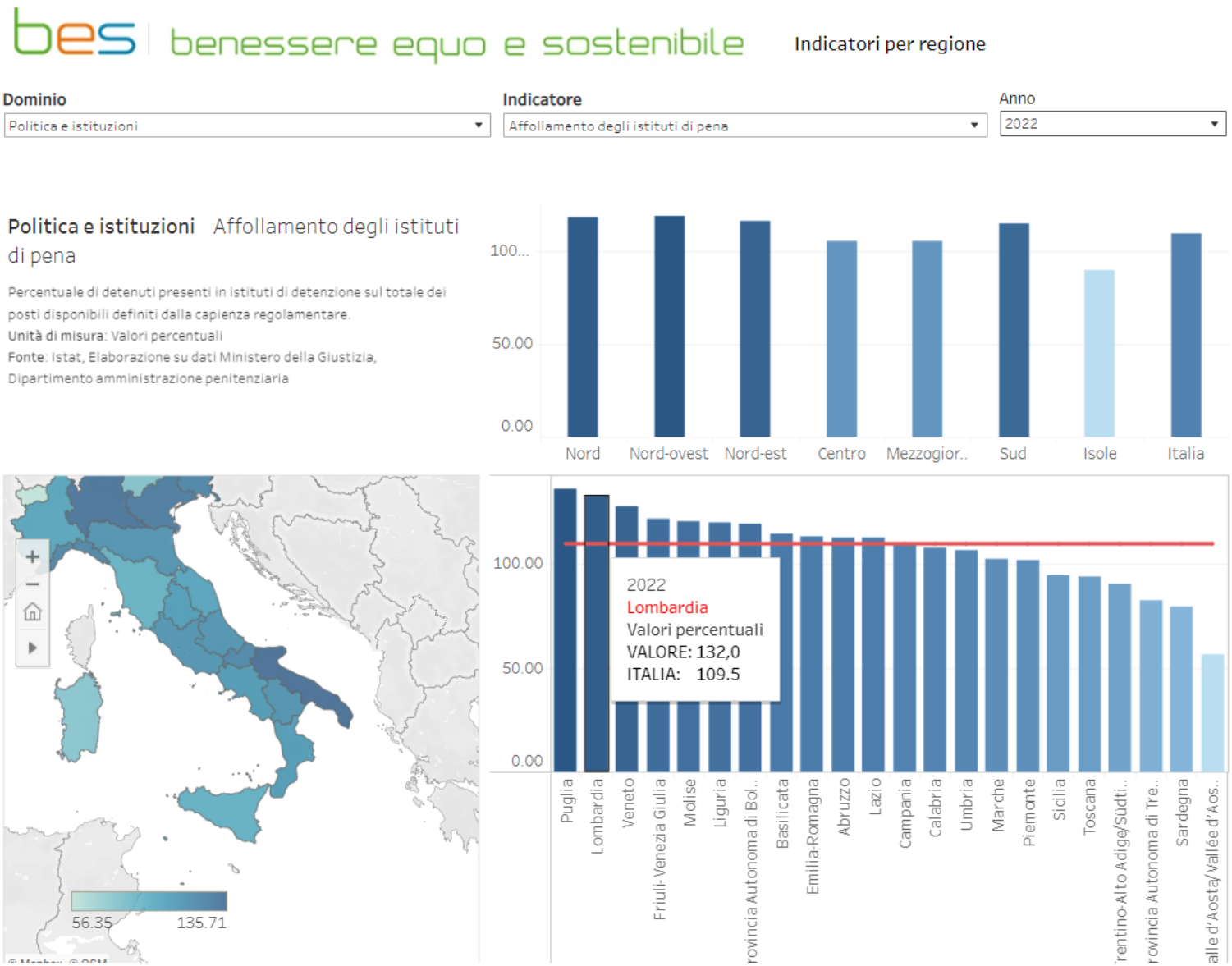


Immagine per Domanda 7

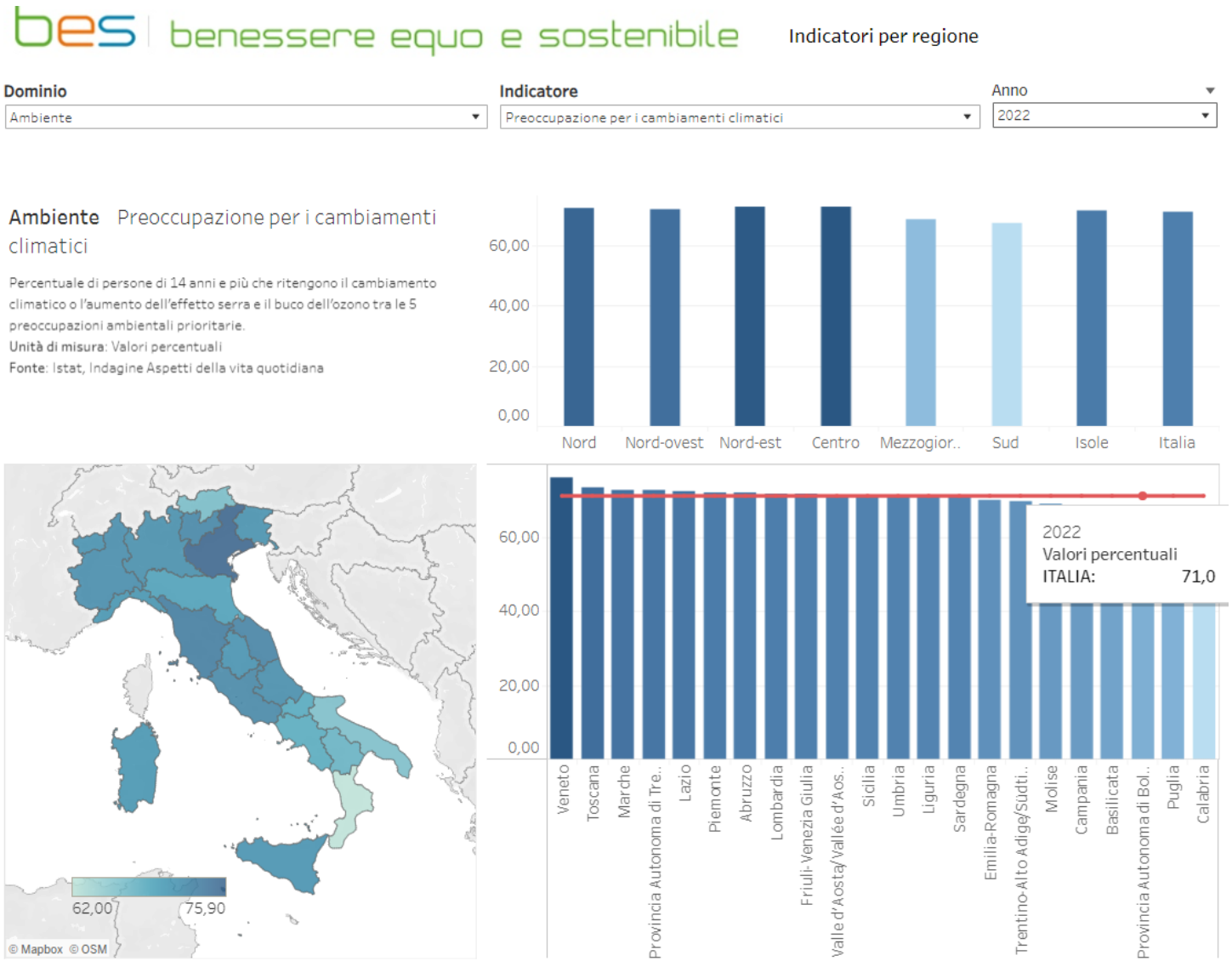


Immagine per Domanda 8

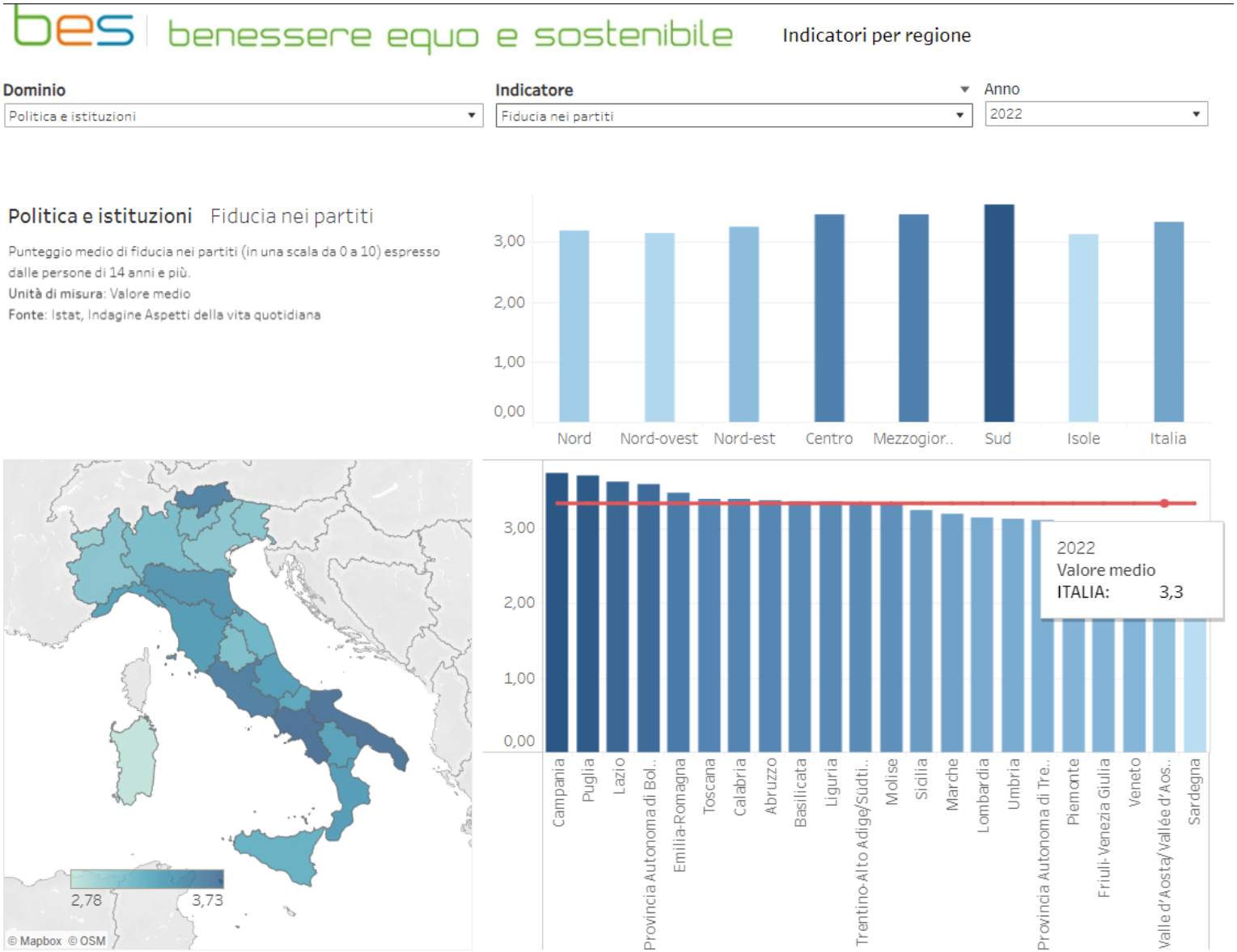
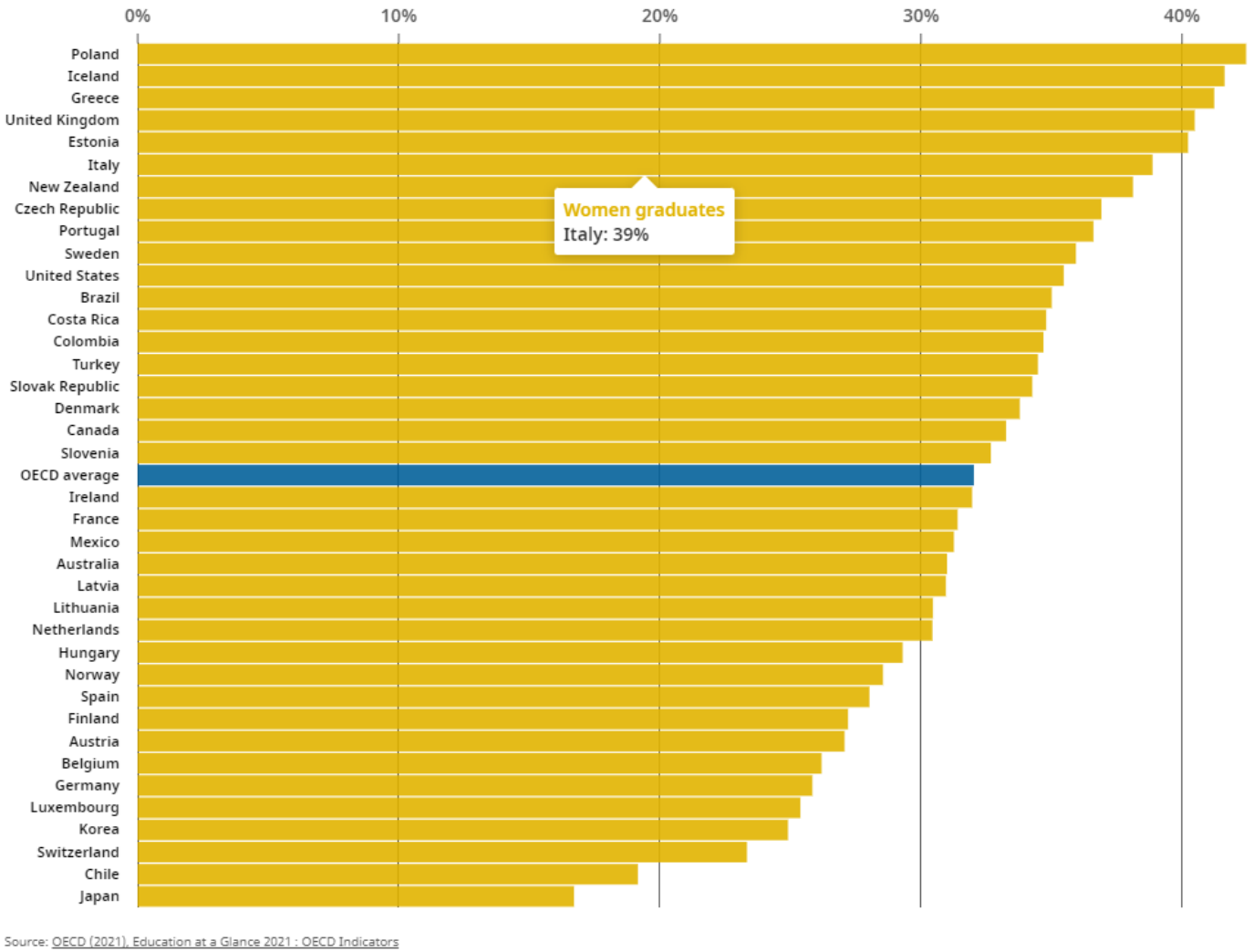


Immagine per Domanda 9



Statistica: Definizione e Tipi

La statistica è la scienza dell'analisi e dell'interpretazione dei dati per prendere decisioni.

Questa disciplina si suddivide in due principali branche:

- La **statistica descrittiva** descrive i dati con un'analisi completa ed esaustiva.
- La **statistica inferenziale** deduce informazioni sulla popolazione generale sulla base di campioni rappresentativi.

In questo corso ci concentreremo sulla **statistica descrittiva**.

Definizioni di Unità Statistica e Variabile

Un'**unità statistica** è un'entità o un elemento individuale su cui vengono misurate o osservate una o più variabili in uno studio statistico.

Una **variabile** è una caratteristica che può assumere diversi valori e che viene misurata o osservata in uno studio statistico.

- Le variabili possono essere suddivise in due categorie principali: **quantitative** e **qualitative**.

Variabili Quantitative

Variabili quantitative rappresentano quantità numeriche.
Possono essere suddivise in variabili **continue** e **discrete**.

- Le **variabili continue** possono assumere un numero infinito di valori all'interno di un intervallo specifico. Ad esempio, altezza, peso, temperatura.
- Le **variabili discrete** possono assumere un numero limitato e contabile di valori. Ad esempio, numero di figli.

Variabili Qualitative

Variabili qualitative rappresentano attributi non numerici o caratteristiche di un'unità statistica. Possono essere suddivise in variabili **nominali** e **ordinali**.

- Le **variabili nominali** rappresentano categorie senza un ordine intrinseco. Ad esempio, colore degli occhi, genere.
- Le **variabili ordinali** rappresentano categorie con un ordine intrinseco. Ad esempio, livelli di istruzione.

Misure di Tendenza Centrale

Le misure di **tendenza centrale** sono utilizzate per rappresentare il valore tipico o centrale di un insieme di dati, fornendo così un'idea dell'ordine di grandezza del fenomeno studiato.

Le principali misure di tendenza centrale sono: **media aritmetica**, **mediana** e **moda**.

Media Aritmetica

La **media aritmetica** è una misura di tendenza centrale che rappresenta il valore tipico di un insieme di dati. Può essere calcolata solo per le **variabili quantitative**.

Formula:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

dove:
x_i - i singoli valori della variabile,
n - il numero totale di osservazioni.

Esempio di interpretazione:
Ad esempio, supponiamo che x rappresenti l'età e $\bar{x} = 24$.
Questo significa che l'età media delle persone nel dataset è di 24 anni.

Mediana

La **mediana** è una misura di tendenza centrale che rappresenta il valore centrale di un insieme di dati. Può essere calcolata per tutti i tipi di variabili **tranne** le **variabili qualitative nominali**.

Calcolo:
Se il numero di osservazioni è **dispari**, la mediana è il valore centrale dei valori ordinati della variabile.
Se il numero di osservazioni è **pari**, la mediana è la media dei due valori centrali dei valori ordinati della variabile.

Esempio di interpretazione:
La mediana del tempo di consegna è di 3 giorni.
Questo significa che il 50% dei pacchi viene consegnato entro 3 giorni e il restante 50% richiede più di 3 giorni per essere consegnato.

Moda

La **moda** è una misura di tendenza centrale che rappresenta il valore più frequente in un insieme di dati. Può essere calcolata per **tutti i tipi di variabili**.

Calcolo:
È il valore della variabile a cui è associata la frequenza più alta.

Esempio di interpretazione:
La moda dei colori preferiti tra i partecipanti è il blu.
Questo significa che il blu è il colore più comune tra i partecipanti.

Scelta della Misura di Tendenza Centrale

È opportuno utilizzare:

- la **media aritmetica** quando si analizzano quantità che variano in modo lineare, evitando valori *anomali* troppo grandi o troppo piccoli;
- la **mediana** quando si desidera conoscere il valore centrale, specialmente in presenza di valori *anomali* che potrebbero influenzare la media;
- la **moda** per evidenziare la caratteristica più diffusa.

Misure di Dispersione

Le misure di **dispersione** sono utilizzate per quantificare la variabilità dei dati attorno alla loro tendenza centrale.

Le principali misure di dispersione sono: **campo di variazione**, **scarto interquartile**, **deviazione standard** e **varianza**.

Campo di Variazione

Il **campo di variazione** è una semplice misura di quanto i dati si estendono da un'estremità all'altra dell'intervallo. Può essere calcolato solo per le **variabili quantitative**.

Formula:

$$x_{max} - x_{min}$$

dove:
x_{max} - il valore massimo della variabile,
x_{min} - il valore minimo della variabile.

Esempio di interpretazione:
Il campo di variazione dei voti degli studenti in un test è 60. Questo significa che la differenza tra il voto massimo e il voto minimo ottenuto dagli studenti è di 60 punti.

Quartili e Scarto Interquartile

I **quartili** dividono un insieme di dati ordinati in quattro parti uguali. Possono essere calcolati per tutti i tipi di variabili **tranne** le **variabili qualitative nominali**.

Lo **scarto interquartile** è la differenza tra il terzo quartile e il primo quartile:

$$IQR = Q3 - Q1$$

Esempio di interpretazione:
IQR delle altezze degli studenti in una classe è di 10 centimetri. Questo indica che il 50% centrale degli studenti ha una differenza di altezza di 10 centimetri.

Deviazione Standard e Varianza

La **deviazione standard** (σ) e la **varianza** (σ^2) misurano la dispersione dei dati rispetto alla media. Possono essere calcolate solo per le **variabili quantitative**.

Formule:

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

Esempio di interpretazione:
Se la deviazione standard del numero di ore di sonno è 4 ore, significa che in media il numero di ore di sonno si discosta di circa 4 ore dalla media del gruppo. La varianza del numero di ore di sonno è 16 ore. Questo indica che c'è una maggiore variabilità nel numero di ore di sonno tra gli individui.

Scelta della Misura di Dispersione

È opportuno utilizzare:

- il **campo di variazione** quando si desidera ottenere una misura semplice della dispersione dei dati, ma si è consapevoli che è sensibile agli estremi;
- lo **scarto interquartile** per ottenere una misura robusta della dispersione che è meno influenzata dagli estremi;
- la **deviazione standard** o la **varianza** quando si desidera una misura precisa della dispersione dei dati rispetto alla media.

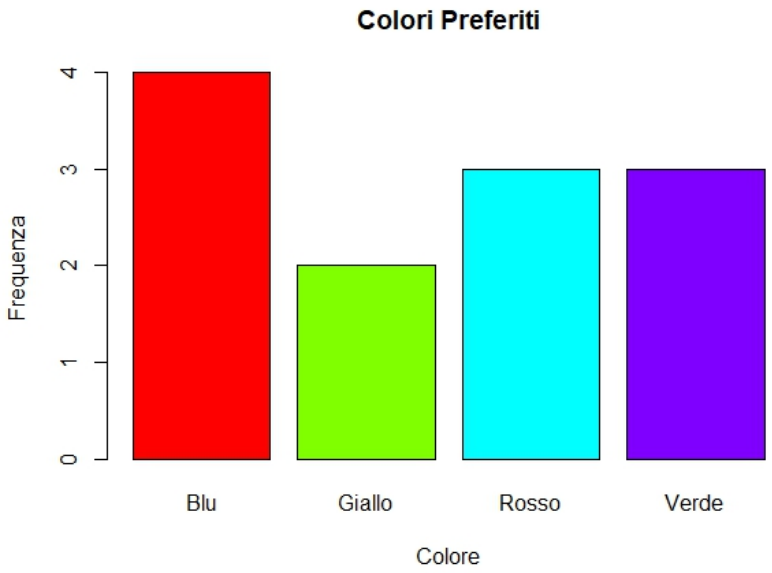
Presentazione Grafica dei Dati

A volte, per comprendere appieno il quadro generale, è necessario guardare oltre un singolo numero. Le **visualizzazioni grafiche** dei dati offrono un'opportunità di esplorare e interpretare le relazioni e i modelli nei dati in modo più intuitivo e immediato.

I grafici più comunemente utilizzati sono: **diagrammi a barre**, **istogrammi** e **diagramma a scatola e baffi**. Un'alternativa ai diagrammi a barre sono i **diagrammi a bastoncini**.

Mentre i diagrammi sono adatti per le **variabili qualitative e quantitative discrete**, gli istogrammi e il diagramma a scatola e baffi sono adatti per le **variabili quantitative continue**.

Grafici per Variabili Qualitative



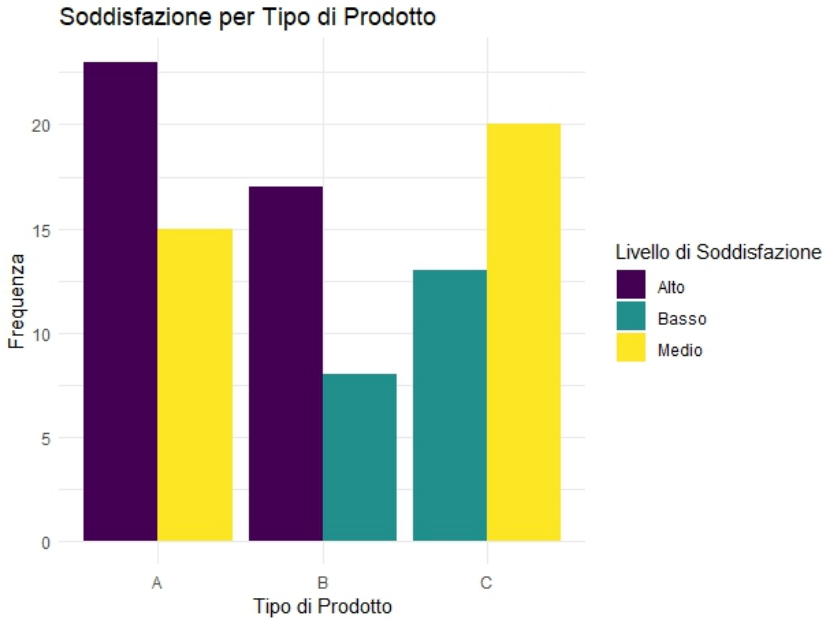
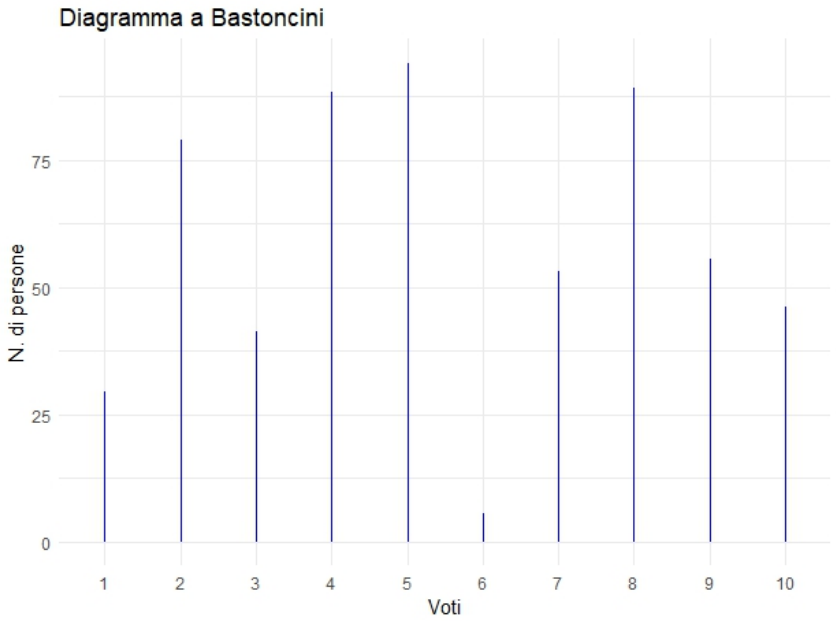


Diagramma a barre per la variabile qualitativa nominale.

Diagramma a barre accostate per la variabile qualitativa ordinale.

Grafici per Variabili Quantitative



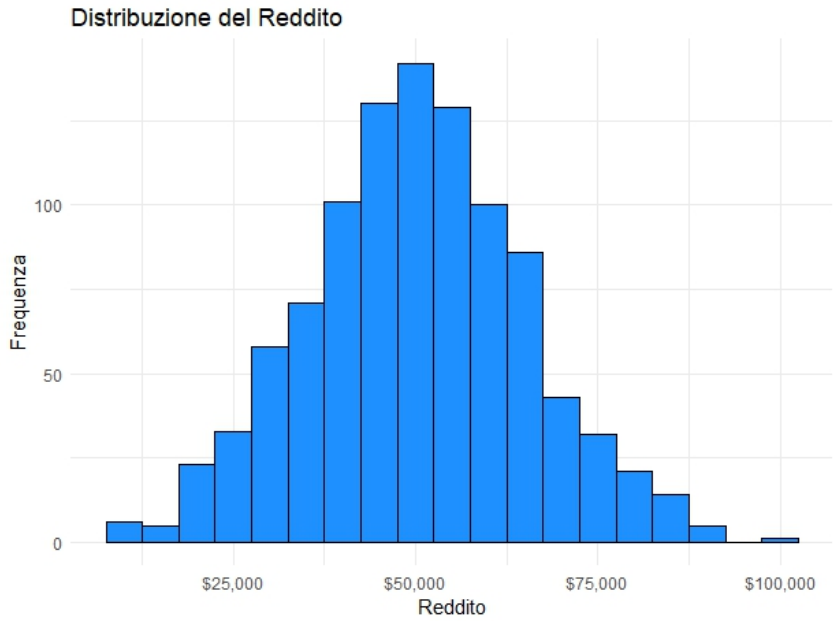


Diagramma a bastoncini per la variabile quantitativa discreta.

Istogramma per la variabile quantitativa continua.

Diagramma a Scatola e Baffi (Boxplot)

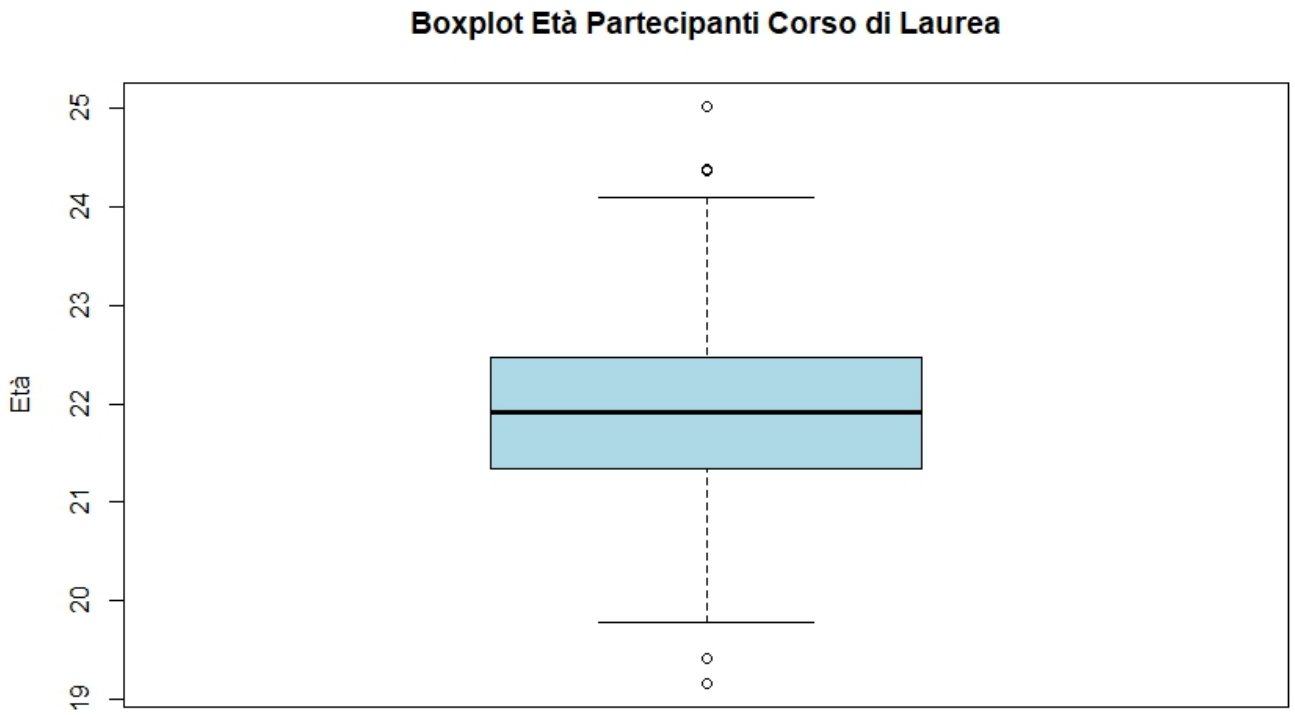
Il **diagramma a scatola e baffi (boxplot)** è un grafico utilizzato per rappresentare la distribuzione di un insieme di dati tramite cinque numeri riassuntivi:

- **Mediana (Q2)**: Linea centrale nella scatola.
- **Primo quartile (Q1)**: Limite inferiore della scatola.
- **Terzo quartile (Q3)**: Limite superiore della scatola.
- **Minimo**: Estremità inferiore del segmento verticale; esclusi i valori estremi.
- **Massimo**: Estremità superiore del segmento verticale; esclusi i valori estremi.

Questo tipo di grafico è particolarmente utile per evidenziare la presenza di **valori anomali**.

CODINGgirls

Esempio di Boxplot



Misure di Associazione

Le misure di **associazione** sono utilizzate per valutare il grado di relazione o dipendenza tra due variabili.

Le principali misure di associazione sono: coefficiente di **correlazione di Pearson** e coefficiente di **correlazione di Spearman**.

Coefficiente di Correlazione di Pearson

Il coefficiente di **correlazione di Pearson** misura la forza e la direzione della relazione lineare tra due variabili. Può essere calcolato solo per le **variabili quantitative**.

Formula:

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

dove:
x_i e y_i - singoli valori delle variabili,
$\bar{x}$ e $\bar{y}$ - medie aritmetiche di x e y.

Un valore di r vicino a 1 indica una correlazione **positiva forte**, mentre un valore vicino a -1 indica una correlazione **negativa forte**. Un valore vicino a 0 indica una correlazione **debole**.

Interpretazione di Correlazione di Pearson

Esempio di interpretazione:
Se il coefficiente di correlazione di Pearson tra il tempo di studio e i risultati degli esami è $r = 0.75$, significa che ogni aumento di una unità nel tempo di studio è associato a un aumento medio di 0.75 unità nei risultati degli esami.
Questo indica una forte relazione positiva tra il tempo di studio e i risultati degli esami, suggerendo che gli studenti che dedicano più tempo allo studio tendono ad ottenere punteggi più alti negli esami.

Nota:
È importante notare che il coefficiente di correlazione misura solo la forza e la direzione di una relazione lineare e **non implica** necessariamente una **relazione di causa-effetto**.

Coefficiente di Correlazione di Spearman

Il coefficiente di **correlazione di Spearman** misura la forza e la direzione della relazione monotona (non necessariamente lineare) tra due variabili. Può essere calcolato per tutti i tipi di variabili **tranne** le **variabili qualitative nominali**.

Formula:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

dove:
d_i - le differenze di rango tra le coppie di osservazioni,
n - il numero totale di osservazioni.

Un valore di r_s vicino a 1 indica una correlazione **positiva forte**, mentre un valore vicino a -1 indica una correlazione **negativa forte**. Un valore vicino a 0 indica una correlazione **debole**.

Esempio di calcolo:

Supponiamo di avere il seguente dataset:

Osservazione	X	Rango di X	Y	Rango di Y	d_i	d_i^2
1	5	3	6	2	1	1
2	7	1.5	7	1	0.5	0.25
3	4	4	4	4	0	0
4	7	1.5	5	3	-1.5	2.25
Somma						3.5

e $n = 4$, quindi:

$$r_s = 1 - \frac{6 \times 3.5}{4(4^2 - 1)} = 1 - \frac{21}{60} = 1 - 0.35 = 0.65$$

Interpretazione di Correlazione di Spearman

Esempio di interpretazione:
Se il coefficiente di correlazione di Spearman tra il livello di istruzione e il reddito mensile è $r_s = 0.6$ significa che un aumento di una posizione nel livello di istruzione è associato a un aumento medio di 0.6 unità nel reddito mensile.
Questo indica una relazione positiva moderatamente forte tra il livello di istruzione e il reddito mensile, suggerendo che individui con un livello di istruzione più alto tendono ad avere un reddito mensile più elevato.

Nota:
Anche qui correlazione **non implica causalità**.

Competizione: Vinci e Scegli il tuo Team!

Ora abbiamo preparato una competizione in cui, risolvendo correttamente e velocemente diversi esercizi sugli argomenti che abbiamo studiato oggi, avrete la possibilità di scegliere i membri del vostro team, mentre gli altri team verranno formati casualmente.

Regole della Competizione

- Ci sono 8 esercizi. Ad ogni punto, riceverete un esercizio.
- Quando diciamo "via", procedete a risolvere l'esercizio.
- Chiunque finisce, deve alzare la mano e dire "fatto". La prima persona a farlo viene alla lavagna e presenta la soluzione.
- Se la soluzione è corretta, questa persona ottiene un punto. Non solo il risultato finale sia valutato, ma anche il modo in cui risolvete l'esercizio. Dovete applicare le formule.
- Per ogni esercizio che riceverete, le regole sono le stesse.
- La persona che raccoglie il maggior numero di punti può scegliere i membri del team per il progetto di gruppo, mentre gli altri team verranno formati casualmente.

Buona fortuna!

CODINGgirls