Statistica e Analisi dei Dati

Perché non è solo roba da matematici

Matteo Dell'Acqua

dellacqua.matteo99@gmail.com

3 Maggio 2022

Indice

Bootstrap

- Verranno utilizzati notebook interattivi utilizzando il linguaggio **Python**

Bootstrap

- Verranno utilizzati notebook interattivi utilizzando il linguaggio **Python**
- Non è un corso di programmazione

Bootstrap

- Verranno utilizzati notebook interattivi utilizzando il linguaggio **Python**
- Non è un corso di programmazione
- Obiettivo: fornire delle basi di analisi dei dati per potersi muovere meglio all'interno del mondo dell'informazione

Importante

Partecipate, intervenite, fate domande e commenti

Perché studiare statistica?

- Essenziale nella modellazione dei problemi

Perché studiare statistica?

- Essenziale nella modellazione dei problemi
- Consente di comprendere meglio la realtà

Perché studiare statistica?

- Essenziale nella modellazione dei problemi
- Consente di comprendere meglio la realtà
- Fornisce un metodo di ragionamento

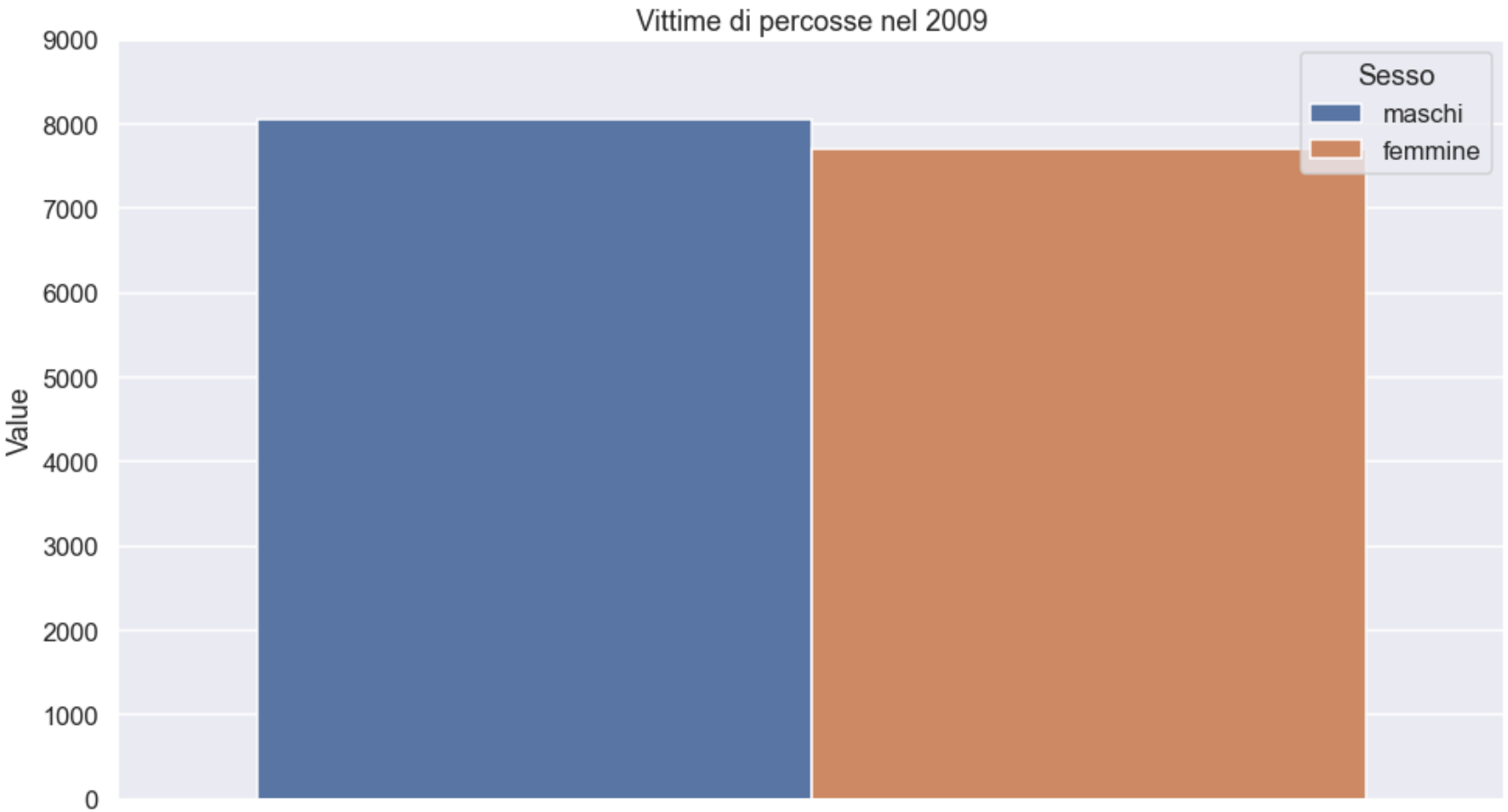
Perché studiare statistica?

- Essenziale nella modellazione dei problemi
- Consente di comprendere meglio la realtà
- Fornisce un metodo di ragionamento

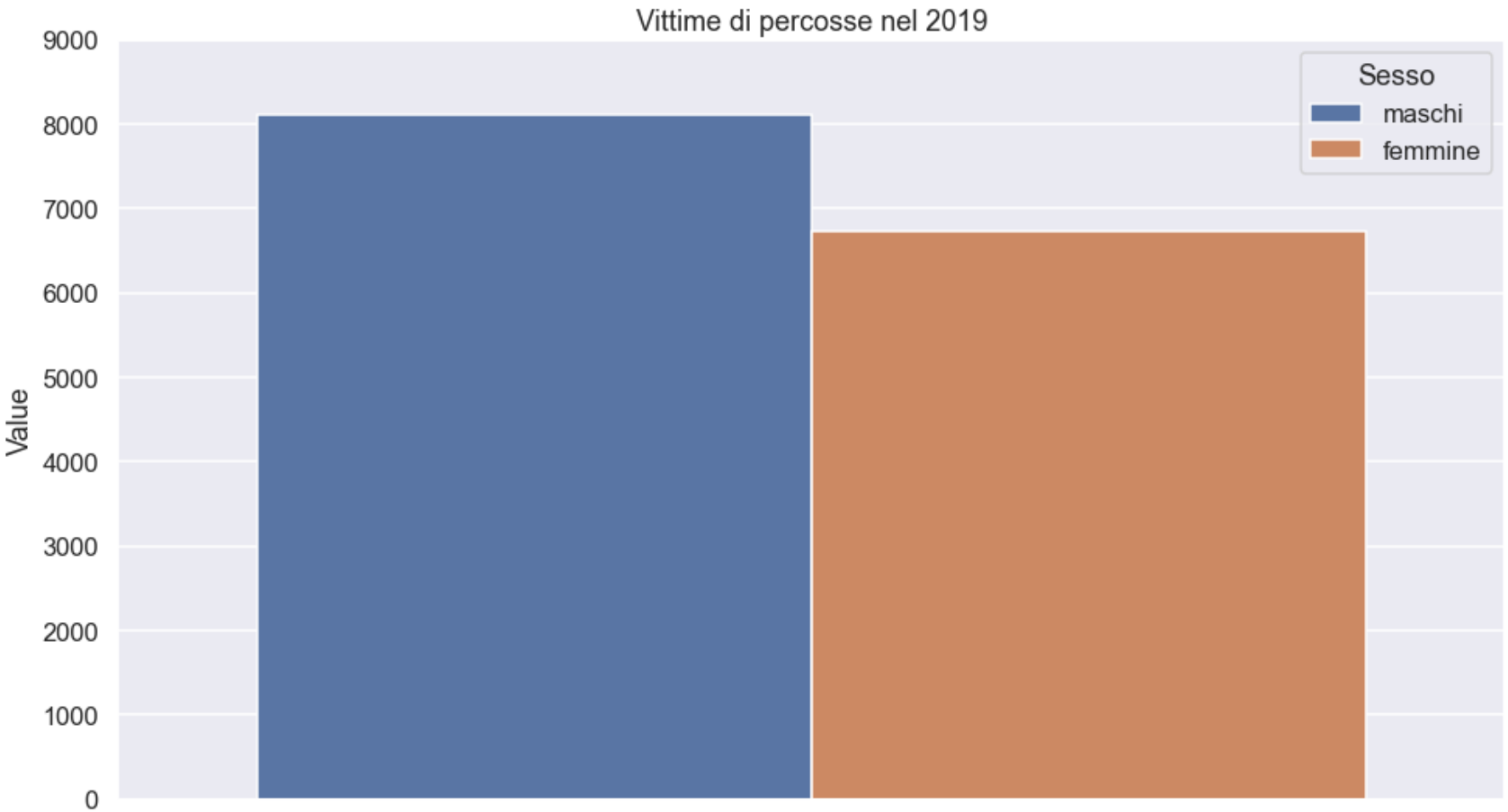
Tutti validi motivi, ma in pratica? A cosa serve?

Vediamo qualche esempio nel mondo attuale

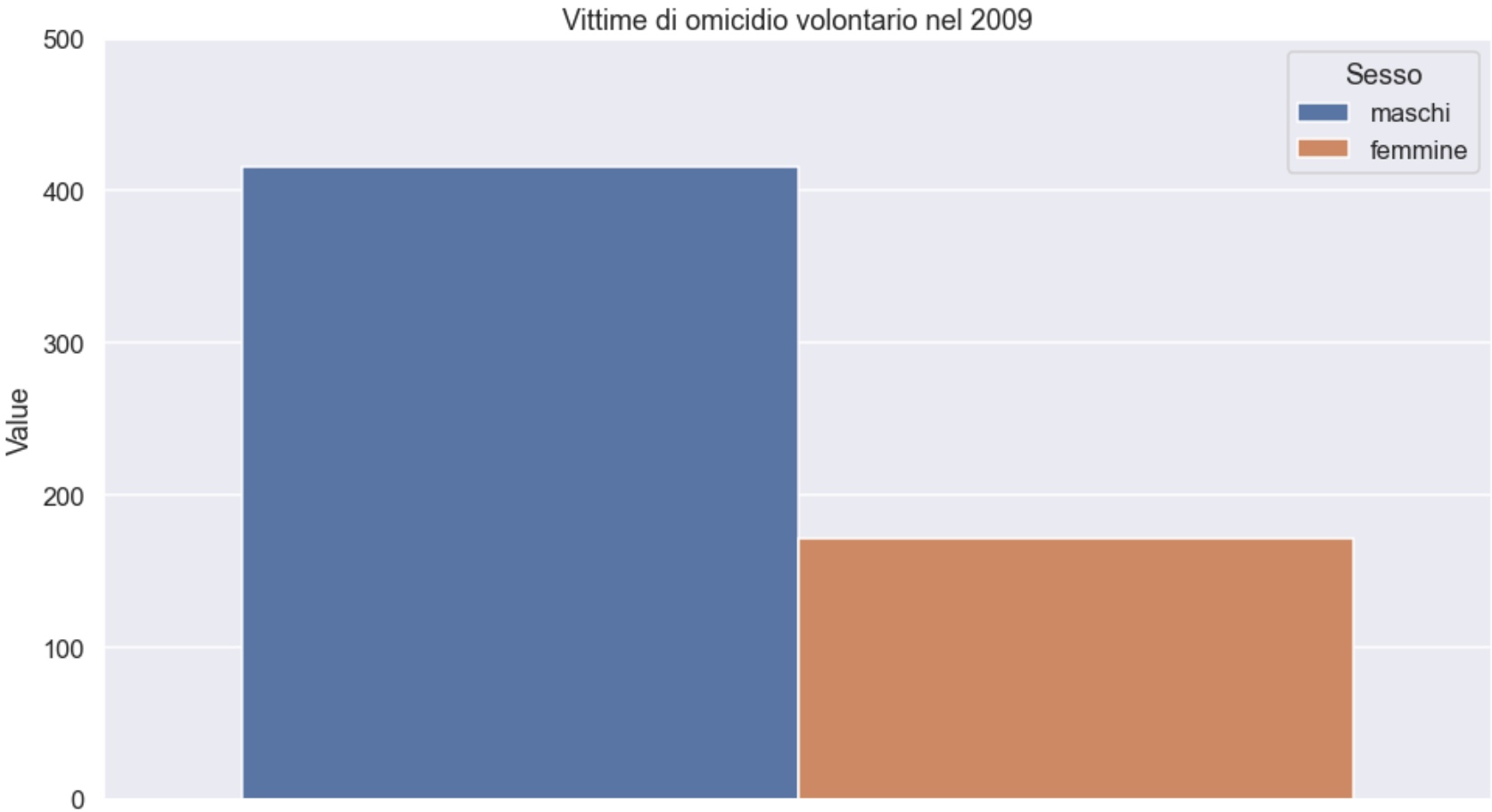
Esempio: vittime di delitti violenti



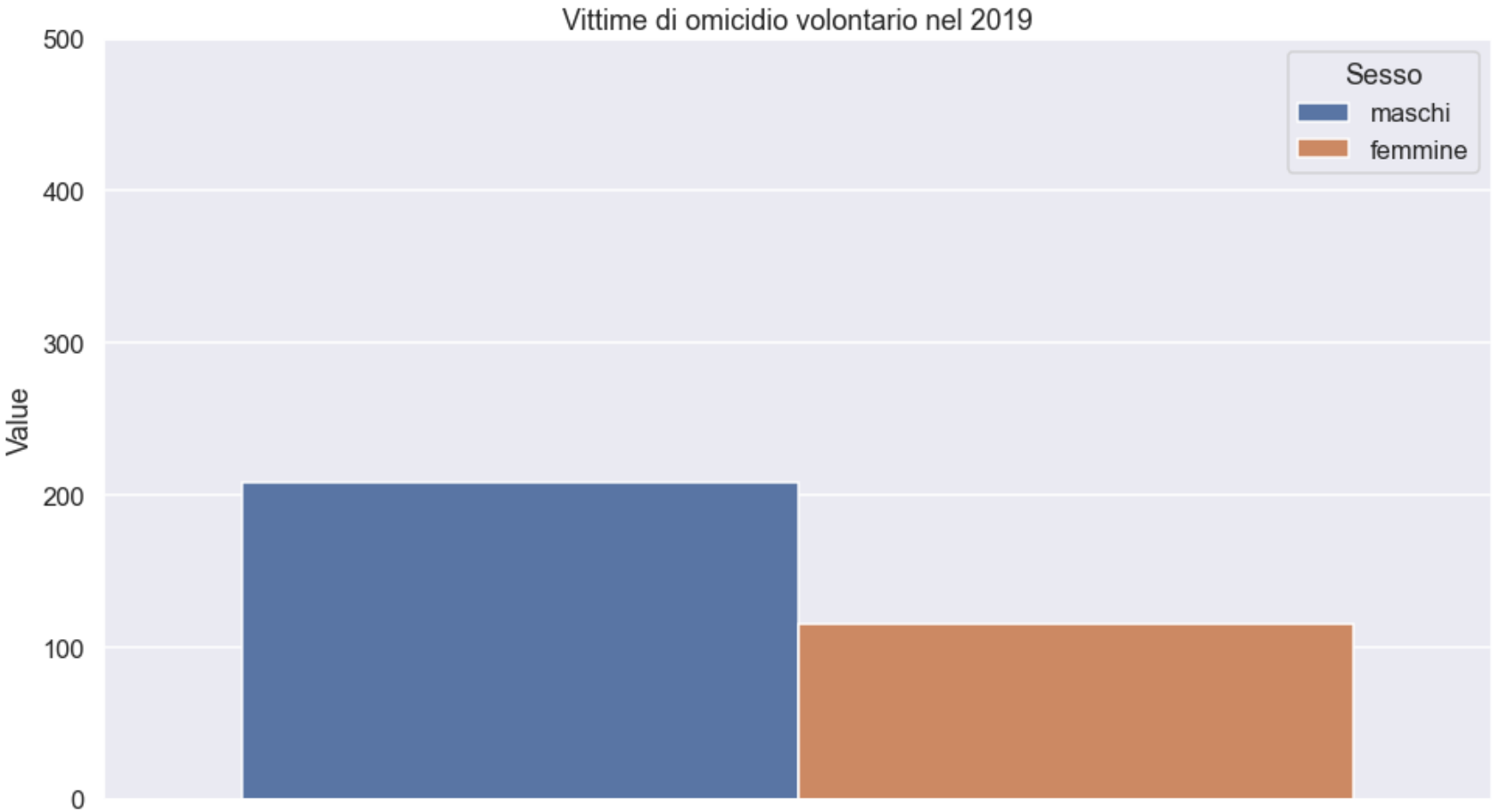
Esempio: vittime di delitti violenti



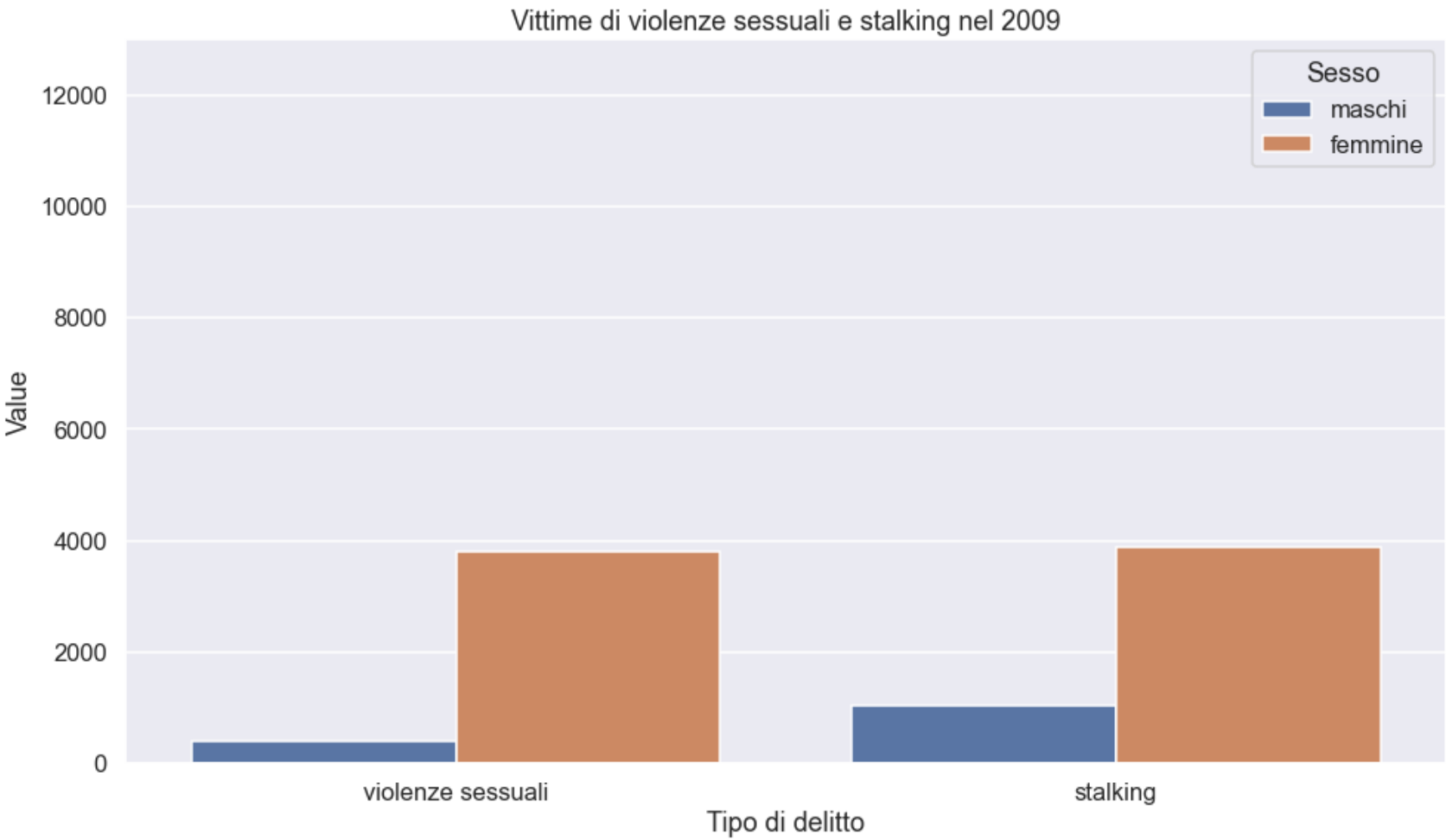
Esempio: vittime di delitti violenti



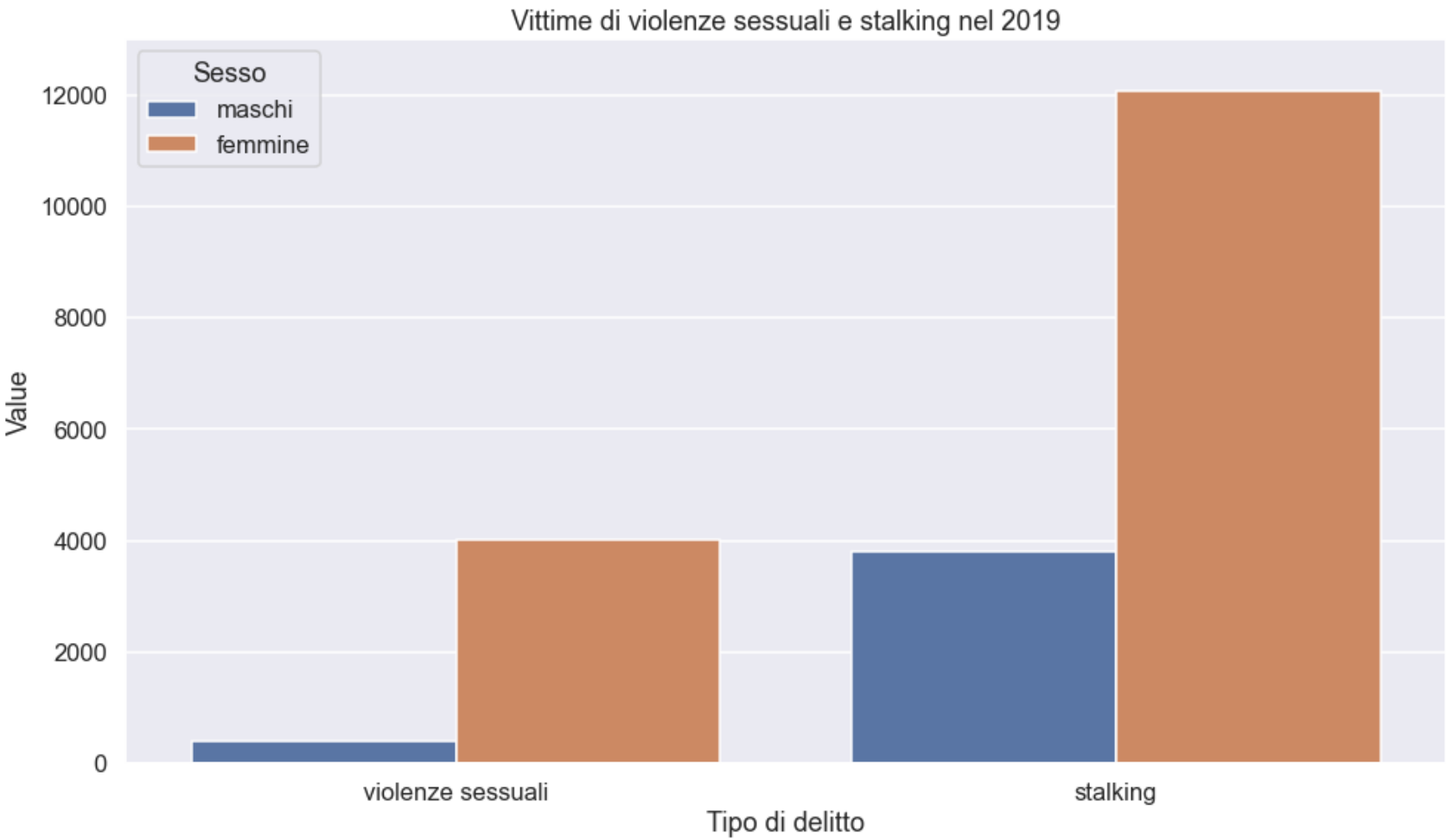
Esempio: vittime di delitti violenti



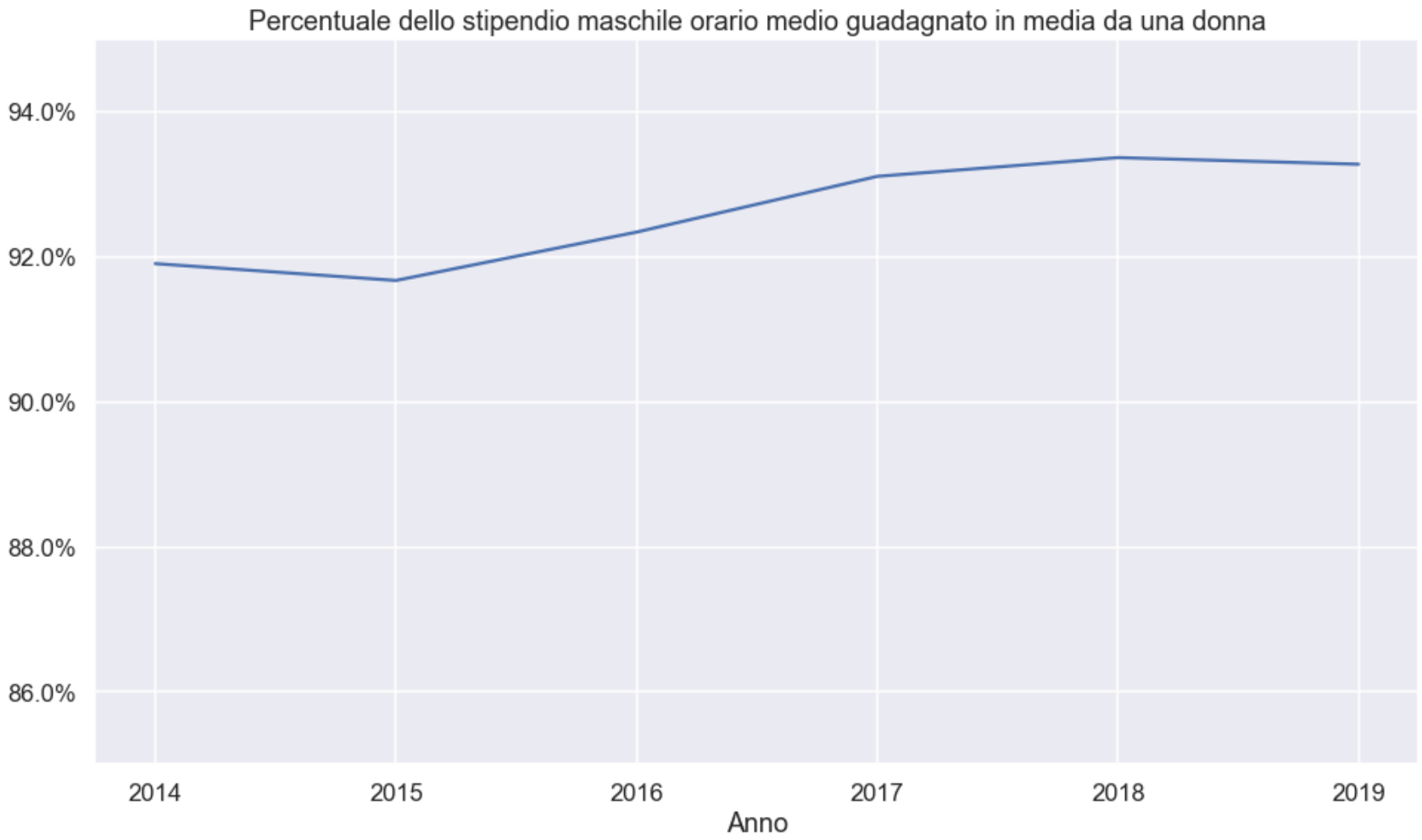
Esempio: vittime di delitti violenti



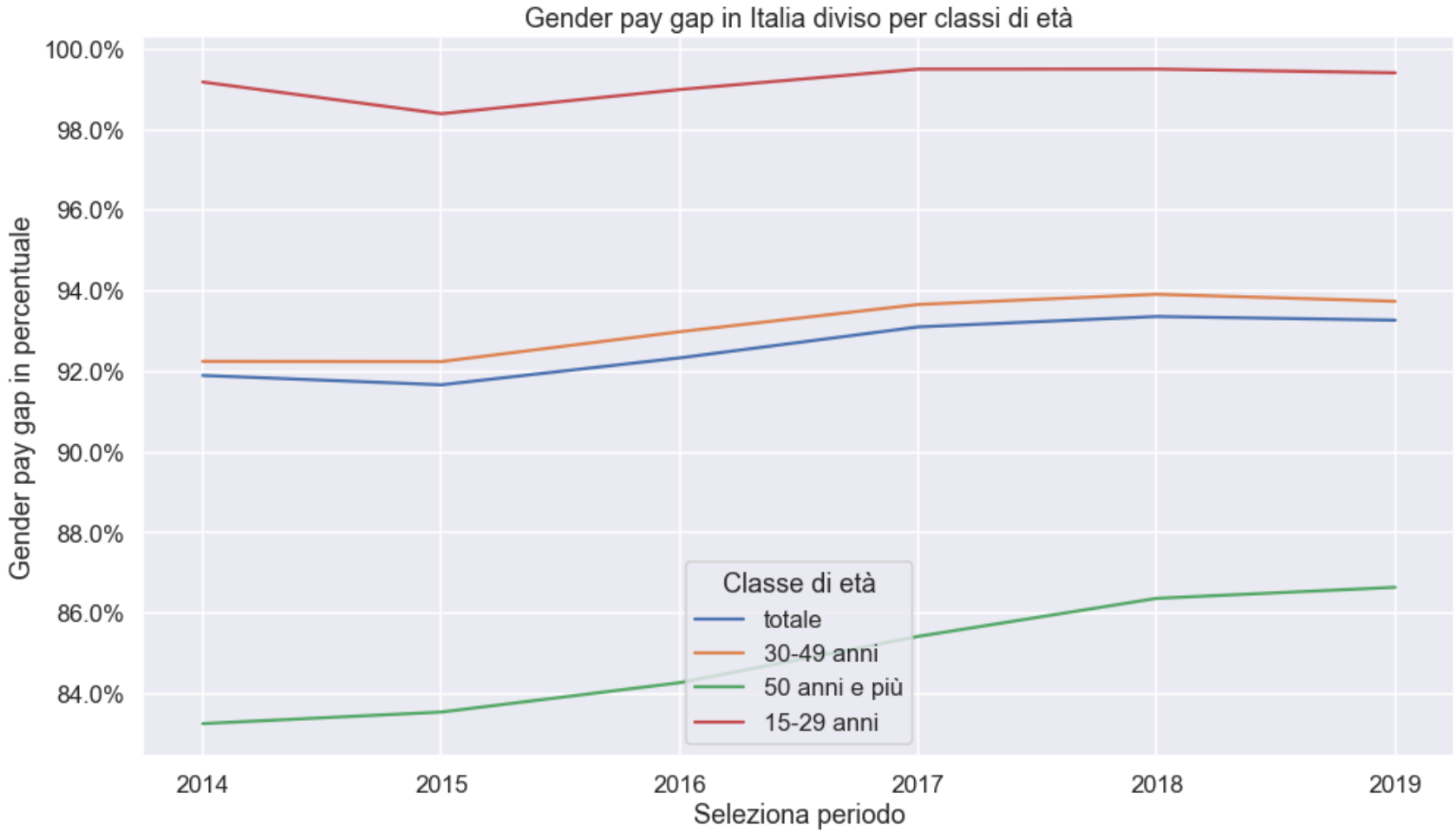
Esempio: vittime di delitti violenti



Esempio: gender pay gap in Italia



Esempio: gender pay gap in Italia



Esempio: gender pay gap in Italia

Quindi?

La statistica riporta il dibattito sul piano razionale evitando il piano emotivo.

Quindi?

La statistica riporta il dibattito sul piano razionale evitando il piano emotivo.

Perché?

Quindi?

La statistica riporta il dibattito sul piano razionale evitando il piano emotivo.

Perché? Supponiamo che alla mia argomentazione venga opposta la seguente obiezione:

Quindi?

La statistica riporta il dibattito sul piano razionale evitando il piano emotivo.

Perché? Supponiamo che alla mia argomentazione venga opposta la seguente obiezione:

> *Questa è la sua opinione™*

Quindi?

La statistica riporta il dibattito sul piano razionale evitando il piano emotivo.

Perché? Supponiamo che alla mia argomentazione venga opposta la seguente obiezione:

> *Questa è la sua opinione*™

Chi dei due fa la figura del peracottaro?

Quindi?

La statistica riporta il dibattito sul piano razionale evitando il piano emotivo.

Perché? Supponiamo che alla mia argomentazione venga opposta la seguente obiezione:

> *Questa è la sua opinione*™

Chi dei due fa la figura del peracottaro?

N.B.: Resta il fatto che qualunque argomentazione debba essere messa in discussione

Quindi?

La statistica riporta il dibattito sul piano razionale evitando il piano emotivo.

Perché? Supponiamo che alla mia argomentazione venga opposta la seguente obiezione:

> *Questa è la sua opinione*™

Chi dei due fa la figura del peracottaro?

N.B.: Resta il fatto che qualunque argomentazione debba essere messa in discussione purché ciò venga fatto in maniera razionale utilizzando esclusivamente fatti e dati oggettivi e discutendo sulle assunzioni.

Dati grezzi e dati aggregati

Cominciamo dando qualche definizione:

Dati grezzi e dati aggregati

Cominciamo dando qualche definizione:

- Dato:

Dati grezzi e dati aggregati

Cominciamo dando qualche definizione:

- Dato: informazione

Dati grezzi e dati aggregati

Cominciamo dando qualche definizione:

- Dato: informazione
- Dato grezzo:

Dati grezzi e dati aggregati

Cominciamo dando qualche definizione:

- Dato: informazione
- Dato grezzo: singola rilevazione effettuata per l'analisi statistica nella forma originale, senza operazioni di *polishing* o di aggregazione

Dati grezzi e dati aggregati

Cominciamo dando qualche definizione:

- Dato: informazione
- Dato grezzo: singola rilevazione effettuata per l'analisi statistica nella forma originale, senza operazioni di *polishing* o di aggregazione
- Dato aggregato:

Dati grezzi e dati aggregati

Cominciamo dando qualche definizione:

- Dato: informazione
- Dato grezzo: singola rilevazione effettuata per l'analisi statistica nella forma originale, senza operazioni di *polishing* o di aggregazione
- Dato aggregato: prodotto a partire dai dati grezzi o da altri dati aggregati. Fornisce informazioni più facilmente interpretabili al costo di una perdita di dettaglio

Dati aggregati

Ci sono due diverse tipologie di dati aggregati:

Dati aggregati

Ci sono due diverse tipologie di dati aggregati:

- Dati quantitativi:

Dati aggregati

Ci sono due diverse tipologie di dati aggregati:

- Dati quantitativi: indici numerici

Dati aggregati

Ci sono due diverse tipologie di dati aggregati:

- Dati quantitativi: indici numerici
- Dati qualitativi:

Dati aggregati

Ci sono due diverse tipologie di dati aggregati:

- Dati quantitativi: indici numerici
- Dati qualitativi: grafici

Esempio di situazione reale

Esempio di situazione reale

Supponiamo di voler analizzare la situazione pandemica in Italia.

Esempio di situazione reale

Supponiamo di voler analizzare la situazione pandemica in Italia.

Per analizzare completamente la popolazione italiana sarebbero richiesti più di 120 milioni di tamponi

Esempio di situazione reale

Supponiamo di voler analizzare la situazione pandemica in Italia.

Per analizzare completamente la popolazione italiana sarebbero richiesti più di 120 milioni di tamponi in un ristretto lasso di tempo (massimo 3-4 giorni, altrimenti l'analisi risulterebbe invalida).

Esempio di situazione reale

Supponiamo di voler analizzare la situazione pandemica in Italia.

Per analizzare completamente la popolazione italiana sarebbero richiesti più di 120 milioni di tamponi in un ristretto lasso di tempo (massimo 3-4 giorni, altrimenti l'analisi risulterebbe invalida). È praticabile?

Esempio di situazione reale

Supponiamo di voler analizzare la situazione pandemica in Italia.

Per analizzare completamente la popolazione italiana sarebbero richiesti più di 120 milioni di tamponi in un ristretto lasso di tempo (massimo 3-4 giorni, altrimenti l'analisi risulterebbe invalida). È praticabile?

La soluzione è prendere un sottoinsieme della popolazione italiana e considerarlo come rappresentativo per l'intera popolazione.

Esempio di situazione reale

Supponiamo di voler analizzare la situazione pandemica in Italia.

Per analizzare completamente la popolazione italiana sarebbero richiesti più di 120 milioni di tamponi in un ristretto lasso di tempo (massimo 3-4 giorni, altrimenti l'analisi risulterebbe invalida). È praticabile?

La soluzione è prendere un sottoinsieme della popolazione italiana e considerarlo come rappresentativo per l'intera popolazione.
Come selezionare il sottoinsieme?

Esempio di situazione reale

Supponiamo di voler analizzare la situazione pandemica in Italia.

Per analizzare completamente la popolazione italiana sarebbero richiesti più di 120 milioni di tamponi in un ristretto lasso di tempo (massimo 3-4 giorni, altrimenti l'analisi risulterebbe invalida). È praticabile?

La soluzione è prendere un sottoinsieme della popolazione italiana e considerarlo come rappresentativo per l'intera popolazione.
Come selezionare il sottoinsieme? Questa è la domanda a cui si proverà a rispondere.

Definizioni

Prima di tutto diamo alcune definizioni:

Definizioni

Prima di tutto diamo alcune definizioni:

- Popolazione:

Definizioni

Prima di tutto diamo alcune definizioni:

- Popolazione: insieme delle entità di interesse

Definizioni

Prima di tutto diamo alcune definizioni:

- Popolazione: insieme delle entità di interesse
- Campione:

Definizioni

Prima di tutto diamo alcune definizioni:

- Popolazione: insieme delle entità di interesse
- Campione: sottoinsieme di dimensioni maneggevoli della popolazione

Parametri del campione

Cosa devo stabilire per scegliere il campione?

Parametri del campione

Cosa devo stabilire per scegliere il campione?

- Dimensioni: quante rilevazioni devo effettuare?

Parametri del campione

Cosa devo stabilire per scegliere il campione?

- Dimensioni: quante rilevazioni devo effettuare?
- Pool: da dove devo prendere le rilevazioni?

Parametri del campione

Cosa devo stabilire per scegliere il campione?

- Dimensioni: quante rilevazioni devo effettuare?
- Pool: da dove devo prendere le rilevazioni?

N.B.: Per un'analisi dettagliata può essere utile utilizzare diversi campioni presi da pool diversi per confrontare come il criterio di selezione influisca sul risultato.

Indici di centralità

Forniscono informazioni sulla posizione dei dati (o, in altre parole, intorno a quali valori numerici si distribuiscono).

Indici di centralità

Forniscono informazioni sulla posizione dei dati (o, in altre parole, intorno a quali valori numerici si distribuiscono).

Tra gli indici di centralità si ricordano:

Indici di centralità

Forniscono informazioni sulla posizione dei dati (o, in altre parole, intorno a quali valori numerici si distribuiscono).

Tra gli indici di centralità si ricordano:

- Media

Indici di centralità

Forniscono informazioni sulla posizione dei dati (o, in altre parole, intorno a quali valori numerici si distribuiscono).

Tra gli indici di centralità si ricordano:

- Media
- Mediana

Indici di centralità

Forniscono informazioni sulla posizione dei dati (o, in altre parole, intorno a quali valori numerici si distribuiscono).

Tra gli indici di centralità si ricordano:

- Media
- Mediana
- Moda

La media

La media è un primo indice utile per descrivere i dati. Essa rappresenta il baricentro dei dati.

La media

La media è un primo indice utile per descrivere i dati. Essa rappresenta il baricentro dei dati.
Dato un campione composto dalle rilevazioni $x_1, x_2, ..., x_n$, la media viene calcolata con la seguente formula

La media

La media è un primo indice utile per descrivere i dati. Essa rappresenta il baricentro dei dati.
Dato un campione composto dalle rilevazioni $x_1, x_2, ..., x_n$, la media viene calcolata con la seguente formula

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

I limiti della media: gli Outliers

Con outlier si intende una rilevazione fuori scala rispetto al resto del campione.

I limiti della media: gli Outliers

Con outlier si intende una rilevazione fuori scala rispetto al resto del campione.

Una rilevazione fuori scala può essere dovuta a un errore nel processo di misurazione, a una corruzione dei dati o può essere semplicemente un valore reale ma fuori scala.

I limiti della media: gli Outliers

La media è molto sensibile agli outliers.

I limiti della media: gli Outliers

La media è molto sensibile agli outliers.

Essendo il baricentro dei dati, più un valore è fuori scala più la media deve spostarsi per mantenere l'equilibrio.

La mediana

La mediana è un indice di centralità robusto rispetto agli outliers.

La mediana

La mediana è un indice di centralità robusto rispetto agli outliers.
Dato un campione composto dalle rilevazioni $x_1, x_2, \ldots, x_n$ **ordinate in ordine crescente**, la mediana viene calcolata solitamente con la seguente formula

La mediana

La mediana è un indice di centralità robusto rispetto agli outliers.
Dato un campione composto dalle rilevazioni $x_1, x_2, ..., x_n$ **ordinate in ordine crescente**, la mediana viene calcolata solitamente con la seguente formula

$$Me = \begin{cases} x_{\frac{n+1}{2}} & \text{se } n \text{ dispari} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{se } n \text{ pari} \end{cases} \quad (2)$$

La mediana

Nella formula precedente, il caso dispari resta uguale per definizione mentre il caso pari può essere modificato a piacere:

La mediana

Nella formula precedente, il caso dispari resta uguale per definizione mentre il caso pari può essere modificato a piacere: di solito si usa la media tra i due valori centrali ma si può anche pensare di utilizzare il valore maggiore tra i due, o il minore.

La mediana

Nella formula precedente, il caso dispari resta uguale per definizione mentre il caso pari può essere modificato a piacere: di solito si usa la media tra i due valori centrali ma si può anche pensare di utilizzare il valore maggiore tra i due, o il minore.

Questo permette alla mediana di ottenere un'interessante proprietà:

La mediana

Nella formula precedente, il caso dispari resta uguale per definizione mentre il caso pari può essere modificato a piacere: di solito si usa la media tra i due valori centrali ma si può anche pensare di utilizzare il valore maggiore tra i due, o il minore.

Questo permette alla mediana di ottenere un'interessante proprietà: a differenza della media, se si evita di utilizzare la media nel caso pari, la mediana garantisce di assumere un valore effettivamente esistente all'interno del campione.

La moda

L'ultimo indice di centralità è la moda.

La moda

L'ultimo indice di centralità è la moda.

Dato un campione composto dalle rilevazioni $x_1, x_2, ..., x_n$, la moda è calcolata come il valore che compare nel campione con la frequenza maggiore.

La moda

L'ultimo indice di centralità è la moda.

Dato un campione composto dalle rilevazioni $x_1, x_2, ..., x_n$, la moda è calcolata come il valore che compare nel campione con la frequenza maggiore.
In caso di "pareggio" si può scegliere come procedere (valore più grande, più piccolo, media tra i candidati, etc.).

La moda

L'ultimo indice di centralità è la moda.

Dato un campione composto dalle rilevazioni $x_1, x_2, ..., x_n$, la moda è calcolata come il valore che compare nel campione con la frequenza maggiore.
In caso di "pareggio" si può scegliere come procedere (valore più grande, più piccolo, media tra i candidati, etc.).

Come per la mediana, anche la moda garantisce di assumere un valore effettivamente esistente all'interno del campione.

Moda di un campione in $\mathbb{R}$

Moda di un campione in $\mathbb{R}$

Cosa succede se abbiamo un campione in $\mathbb{R}$ e nessun valore si ripete?

Moda di un campione in $\mathbb{R}$

Cosa succede se abbiamo un campione in $\mathbb{R}$ e nessun valore si ripete?

In questi casi, si ricorre spesso ad un processo di categorizzazione: ad ogni valore si assegna un'etichetta (di solito sulla base di un intervallo) e poi si calcola la moda sulle etichette.

Link utili

- Repository github: `https://github.com/MatteoH2O1999/iis-alessandrini-lectures`
- Anaconda: `https://www.anaconda.com/products/distribution`
- Istat datasets: `http://dati.istat.it`
- Our World in Data: `https://ourworldindata.org`