

Esercizi lezione 7

Esercizio 1/2

Tra i beginner datasets scaricabili all'indirizzo

<https://www.kaggle.com/datasets/ahmettezcantekin/beginner-datasets>

selezioniamo `amazon.csv`, un dataset contenente una serie di recensioni su Amazon.

- Valutiamo dimensione del dataset, e visualizziamo dieci righe a caso;
- Osserviamo quali sono i nomi di colonna;
- Il dataset è bilanciato? Ovvero, il numero di recensioni positive è uguale a quello delle negative?

Esercizio 2/2

- Quante recensioni parlano di app?
- Quante parlano di giochi?
- In generale quante recensioni raccomandano il prodotto?
- E quante *non* lo raccomandano?
- Stampate qualche esempio dell'una e dell'altra.
- Le recensioni che menzionano il Kindle sono perlopiù positive o negative?

Esercizio

Il dataset `diabetes.csv` raccoglie persone con diabete meno, e il valore di diverse variabili fisiologiche dei pazienti.

- Osserviamone le dimensioni e un'anteprima di cinque righe;
- Prendiamoci un po' di tempo per dare un'occhiata ai metadati delle colonne;
- Stampiamo dei descrittori statistici del dataset;
- Selezioniamo i dati relativi a diverse fasce di età: < 20 , 20-30, 30-40, 40-50, > 50 ;
- Qual è la media della pressione sanguigna diastolica per le diverse fasce di età?
- Qual è la media della pressione per ogni singolo anno di età?

Esercizio

Il dataset `insurance.csv` contiene dati rispetto a caratteristiche e abitudini delle persone, e della zona in cui vivono, rispetto ai costi individuali per le cure mediche come premio per le assicurazioni sulla salute.

Visualizziamone le dimensioni, un'anteprima e osserviamo i nomi di colonna;
Quali sono le medie di `charges` rispetto a `region`? Ci sono differenze significative?

E rispetto a `smoker`? E a `sex`?

Quali sono i descrittori statistici di `bmi`? Quali sono minimo, media e massimo di `charges` rispetto ai diversi quartili dei valori di `bmi`?

Esercizio 1/2

Il dataset `pokemon.csv` contiene un database di Pokémon, con dati quali nome, tipi di attacco, valori di attacco/difesa/eccetera, e se sono o meno leggendari.

- Verifichiamo la dimensione, un'anteprima e osserviamo i nomi di colonna;
- È verosimile che la prima colonna dovrebbe essere un indice?
- Confrontiamolo con l'indice messo automaticamente da Pandas: combaciano?
- Se no, settare la prima colonna come indice.

Esercizio 2/2

- Ci sono ripetizioni, ovvero righe che si riferiscono allo stesso Pokémon?
- Per ogni tipo in `Type 1` quanti ce ne sono?
- Quali sono i Pokémon leggendari?
- E quali sono i leggendari di tipo `1 Grass`?
- E leggendari di tipo `1 Ice` o `Fire`?
- Ordiniamo il dataset per la colonna `Name`;
- Trasformiamo `Name` nell'indice;
- Come accediamo alle statistiche di tutti i Pokémon il cui nome inizia per `D`?
- Quali sono i Pokémon della prima generazione con attacco > 50 e HP < 60 ?
- Ci sono dati nulli?



GRAZIE
Epicode