EPICODE

Esercizi lezione 4

Esercizio 1/3

Estraiamo la tabella `dimproduct` dal database AdventureWorks, e valutiamo quanto segue:

- Quanti dati ci sono in totale?
- Quali sono i metadati?
- Stampiamo il primo elemento
- Stampiamo l'ultimo elemento
- Riusciamo a stampare cinque elementi a caso?
- Quali sono i colori disponibili?

Esercizio 2/3

- In media quanto pesano i prodotti (per questo esercizio, ignoriamo l'unità di misura, ma usiamo solo i valori della colonna relativa)?
- Quanto pesa il più leggero?
- Quanto pesa il più pesante?
- Quanti prodotti pesano più di 100?
- Quanto costano in media i prodotti (colonna `DealerPrice`)?
- Quali sono i quartili dei prezzi?

Esercizio 3/3

- Qual è il prezzo medio per i prodotti di colore blu?
- Qual è il prezzo medio per i prodotti di colore rosso o nero?
- Qual è il prezzo massimo per i prodotti di taglia 42 e peso oltre i 10 Kg?
- Qual è il nome inglese e il costo di produzione (`StandardCost`) di tutti i prodotti di taglia 42, peso oltre i 10 Kg e colore argento?
- Visualizziamo lo `StandardCost` e il `DealerPrice` degli ultimi 20 elementi del dataset: quali sono le differenze? C'è un pattern? C'è qualche elemento che non lo segue?

Esercizio

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 la dimensione del dataset
- 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, oppure no?

Esercizio

Il dataset `diabetes.csv` raccoglie persone con diabete o 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: 20-29, 30-39, 40-50;
- Qual è la media della pressione sanguigna diastolica per le diverse fasce 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

- Quali sono i Pokémon leggendari?
- E quali sono i leggendari di tipo 1 `Grass`?
- E leggendari di tipo 1 `Ice` o `Fire`?
- Trasformiamo la colonna `Name` nell'indice;
- Quali sono i Pokémon della prima generazione con attacco > 50 e HP < 60?

EPICODE

GRAZIE

Epicode