



Esercizi lezione 5

Esercizio 1/2

Andiamo su

http://www.datiopen.it/it/opendata/Mappa_dei_pub_circoli_locali_in_Italia e scarichiamo il file (in formato CSV) della mappa dei pub, circoli e locali in Italia

Nota: il dataset non ha un encoding UTF-8 ma Latin1; inoltre il separatore non è una virgola, ma un punto e virgola.

Dunque per leggerlo dovremo aggiungere i parametri `encoding=` e `sep=`, ad esempio:

```
pd.read_csv(file_path, encoding="latin1", sep=";")
```

dove `file_path` è la posizione del file nel nostro calcolatore.

Esercizio 2/2

Esaminiamo il dataset:

- quanti dati ci sono in totale?
- quali sono i metadati?
- stampiamo il primo elemento
- stampiamo l'ultimo elemento
- riusciamo a stampare un elemento a caso?
- quali sono gli anni di inserimento presenti?
- quante attività ci sono nel quadrato di longitudine 9-10 e latitudine 45-46?
- quante attività ci sono nella provincia di Vicenza?
- quante enoteche ci sono, e come si chiamano?
- quante attività ci sono in Lazio e Abruzzo assieme?

Esercizio

Il dataset `insurance.csv` dei **beginner_datasets** contiene dati su caratteristiche e abitudini delle persone 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`?

Nota: è lo stesso esercizio della volta scorsa, ma stavolta da eseguire con il metodo `.groupby()`

Esercizio

Carichiamo in un DataFrame il dataset `iris.csv` dei **beginner_datasets** e calcoliamo:

- La media della lunghezza dei petali di tutto il dataset
- La media della lunghezza dei petali per ogni specie di Iris, utilizzando il metodo `.groupby()`
- Media, minimo e massimo della larghezza dei sepali per ogni specie, utilizzando `.groupby()` e `.agg()`

Esercizio

Nei **beginner_datasets**, il dataset `wine.csv` contiene delle analisi organolettiche su diverse qualità di vini:

- Qual è la media di concentrazione alcolica per ogni qualità? Ci sono differenze? E rispetto alla media totale?
- C'è differenza nella concentrazione alcolica per vini bianchi e vini rossi?
- Rifacendo le analisi dei due punti precedenti ma per il pH, cambia qualcosa?
- E per i solfati?

Esercizio 1/3

Nei **beginner_datasets**, il file `boston.csv` contiene il Boston Housing Dataset, che deriva dalle informazioni raccolte dal Census Service degli Stati Uniti sulle abitazioni nell'area di Boston. Di seguito vengono descritte le colonne del dataset:

- CRIM - tasso di criminalità pro capite per città
- ZN - proporzione di terreni residenziali suddivisi in zone per lotti superiori a 25.000 piedi quadrati.
- INDUS - percentuale di acri di attività commerciali non al dettaglio per città.
- CHAS - variabile dummy del fiume Charles (1 se il tratto costeggia il fiume; 0 altrimenti)

Esercizio 2/3

- NOX - concentrazione di ossidi di azoto (parti per 10 milioni).
- RM - numero medio di stanze per abitazione
- AGE - proporzione di unità abitative occupate da proprietari costruite prima del 1940
- DIS - distanze ponderate da cinque centri occupazionali di Boston
- RAD - indice di accessibilità alle autostrade radiali
- TAX - aliquota dell'imposta fondiaria sul valore pieno per 10.000 dollari
- PTRATIO - rapporto alunni-insegnanti per città
- BLACK - la percentuale di neri per città
- LSTAT - % di popolazione di condizione più bassa
- MEDV - Valore mediano delle case, espresso in migliaia di dollari

Esercizio 3/3

- La media del prezzo delle case cambia a seconda della distanza dal fiume Charles?
- Si nota una correlazione tra il tasso di criminalità e il valore delle abitazioni? Come si può spiegare il risultato?
- Qual è la media del numero di stanze rispetto al rapporto alunni-insegnanti? E del valore delle case? Appare esserci una qualche correlazione? Come si può spiegare il risultato?
- Rispetto all'accessibilità alle autostrade, cambia qualcosa la media delle età delle abitazioni? E del numero di stanze? E delle tasse?

Esercizio 1/3

Abbiamo un DataFrame di dipendenti:

```
employees_df = pd.DataFrame({  
    'employee_id': [101, 102, 103, 104, 105],  
    'name': ['Alice', 'Bob', 'Charlie', 'David', 'Emma'],  
    'department_id': [1, 2, 1, 2, 3]  
})
```

Esercizio 2/3

E un DataFrame di dipartimenti:

```
departments_df = pd.DataFrame({  
    'department_id': [1, 2, 3],  
    'department_name': ['HR', 'IT', 'Finance'],  
    'location': ['New York', 'San Francisco', 'Chicago']  
})
```

Esercizio 3/3

- Unire questi DataFrame in base alla colonna comune department_id, in modo da avere nel risultato informazioni sia sui dipendenti che sui dipartimenti, usando la funzione `.merge()`
- Per ogni DataFrame, trasformare la colonna department_id nell'indice, facendo in modo che la modifica sia permanente; poi unire i due dataset mediante il metodo `.join()`
- Ci sono differenze nel risultato? Quali? Perché?

Esercizio

- Dal database **AdventureWorksDW** importiamo le tabelle `dimemployee` e `dimemployeesalesterritory` come DataFrame
- Effettuiamo un join tra i due DataFrame usando le colonne `EmployeeKey`
- Controlliamo la dimensione del DataFrame risultante: è quella attesa?
- Importiamo ora la tabella `dimsalesterritory` ed effettuiamo un join tra questa e il DataFrame risultante della join precedente, usando le colonne `SalesTerritoryKey`
- Su questo DataFrame contare quanti dipendenti ci sono per ogni paese (`country`) e per ogni regione (`region`)
- Valutiamo la media del `BaseRate` per ogni paese: ci sono differenze?



GRAZIE
EPCODE