DATA SCIENCE

LOGI IT X
Logically better

I dati in python

Come rappresentare i dati in python?

- Le stringhe
- I vettori le matrici
  - Liste
  - Tuple
  - Dictionary

Le stringhe

La stringa è una delle strutture dati utilizzata per comunicare.

X = ‘ Salve, Buongiorno!’

Le funzioni che possono operare sulle stringhe

lower(), upper(), capitalize() decode(), encode()

La funzione strip() elimina gli spazi

La funzione split(“, “) suddivide la stringa in un insieme (array) di sottostringhe.

La funzione join() unisce le stringhe a formare una unica stringa

Le liste

Le liste contengono dati (elemento) di tipo diverso.

elem = [‘uno’, ‘due’, tre’, 34]

Esistono diverse funzioni per operare sulle liste

len() numero di elementi ,

+ concatena due liste

Porzioni di liste ***nome[inizio: fine: incremento]***

L’operatore ***in*** permette di individuare un elemento in una lista

Il metodo ***append*** aggiunge elementi ad una lista

Altri metodi insert(), sort(), remove(), reverse()

Le funzioni min(nome_lista) e max(nome_lista)

Alcune operazioni sulle liste

- Funzione range per generare liste ordinate di interi
  - A=list(range(0,10,3)) -> [0,3,6,9]

- Operatore **in** verifica se il valore presente a sinistra è contenuto negli elementi presenti a destra risultato true o false
- Nel caso del ciclo for invece effettua una scanzione di tutti gli elementi

Le tuple

- Le tuple sono liste immutabili
  - My_tuple= (1, 2, 3, 4)
- Le tuple sono più veloci e supportano le stesse operazioni delle liste a parte le operazioni che variano il contenuto
- Posso convertire liste in tuple
  - My_list = list(My_tuple)

Dictionary

- In python un dizionario è una collezione di dati costituito da una chiave ed un valore

Telefoni={‘mario’: ‘0761303030’, ‘paolo’: ‘06404040’}

Accesso al dato Telefono[‘mario’] → 0761303030

Operatori: in, not in

    if ‘anna’ in Telefoni: (vero o falso)

Aggiunta o cancellazione

    Telefoni[‘giorgio’]=‘02090909’

    Telefoni.update(‘giorgio’ = ‘34499’)

    del Telefono[‘mario’]

Ciclo

    for key in Telefoni:

    print(key, Telefoni[key])

Metodi dictionary

- **clear** ripulisce il dizionario
- **get** restituisce il valore associato alla chiave

  rubrica.get('anna', 'telefono non trovato')
- **setdefault** assegna un valore se non presente
- **itemes** restituisce tutti i valori
- **keys** restituisce tutte le chiavi
- **pop(chiave, valore)** restituisce il valore associato ed elimina la coppia associata
- **values** restituisce tutti i valori del dizionario

Metodi dictionary

- popitem() Rimuove e restituisce un elemento arbitrario da dizionario
- update(d2) Aggiunge gli elementi del dizionario d2 a quelli del dizionario
- copy() Crea e restituisce una copia di dizionario

I valori del dictionary posso essere liste o altri dizionari

Dic = {'jack': 4098, 'joe': [4,1,2,7]}

Dic['joe'][0] -> 4

Esempio dictonary


```
def main():
  rubrica={'mario':'063333', 'paolo':'028888', 'giovanni':'07666666'}
  print(rubrica)
  if 'anna' not in rubrica:
     rubrica['anna']='098888771’
  #print(rubrica)
  #print(len(rubrica))
  rubrica.update(anna=['0761','343434','Roma'])
  print(rubrica['anna'][2])
  for key in rubrica:
     print(key, " : ", rubrica[key])
  #print(rubrica.values())
  #print(rubrica.keys())

main()
```

Set

E' un insieme che memorizza una collezione di dati, tutti gli elementi sono unici, non sono ordinati, possono essere di tipo diverso.

myset =set()

myset = set(['uno', 'due', 'tre', 'due'])

myset.add('sei')

myset.update(['nove', 'otto'])

myset.remove('uno')

if 'uno' in myset: (vero, o falso)

Unione *set1.union(set2)* oppure *set3 =set1 | set2*

Intersezione *set1.intersection(set2)*

Differenza set1.difference(set2)

Leggere un file csv


```
import csv
with open("incassi.csv", newline="\n") as filecsv:
    lettore = csv.reader(filecsv, delimiter=";")
    header= next(lettore)
    print(header)
    for linea in lettore:
        print(linea)
```

Basi di statistica

La statistica serve a modellare i concetti tramite la modellazione dei dati di interesse (popolazione).

La scienza dei dati si occupa di standardizzare i dati con lo scopo di rappresentarli ed interpretarli.

- Come ottenere e campionare i dati
- Le misurazioni del centro, della varianza e della posizione relativa
- La correlazione tra i dati

Statistica

Paramentro

- Misura numerica che descrive una caratteristica della popolazione (set di dati)

Campione

- Sottoinsieme rappresentativo della popolazione

Statistica

- Misura numerica che descrive una caratteristica di un campione di una popolazione

Campionamento dei dati

I dati si ottengono tramite osservazioni o sperimentazioni

Osservazioni: misure di specifiche caratteristiche senza alcuna manipolazione. Si tratta solo di osservare e raccogliere i dati.

Sperimentazioni: le misure (osservazioni) si ottengono dopo uno specifico trattamento che comporta elaborazioni ed analisi.

Campionamento casuale: ogni membro di una popolazione ha la stessa probabilità di essere scelto. Riduzione dell'impatto dei fattori di confusione.

Misurazione del centro

Dipende dalle caratteristiche del dataset.

Serve a generalizzare il valore contenuto in una grande massa di dati.

- E' la misura del valore centrale di un dataset.
  - **Media aritmetica**

$$media = \frac{\sum_{i=0}^{n} x^i}{n}$$

  - **Mediana**

    La mediana è il numero che si trova nel mezzo del dataset dopo aver disposto gli elementi in ordine

Misurazioni della variabilità

Misura la dispersione dei dati

E' un modo utile per verificare se i dati contengono molti valori anomali

Misura della variabilità dei dati.

- **Intervallo** definito come il valore massimo meno il valore minimo.

$$Int = V_{max} - V_{min}$$

Ci indica quanto distano i valori estremi.

E' condizionato dai valori anomali.

Deviazione standard

- **Deviazione standard**: misura quanto i valori deviano, rispetto alla media aritmetica, i valori contenuti nei dati

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n}}$$

Calcola la distanza media fra i valori contenuti nei dati e la media artmetica, (scarto quadratico medio).

L'unità di misura è la stessa del dato

Coefficiente di variazione

Se dobbiamo valutare la dispersione dei dati di 2 differenti dataset che adottano unita di misura completamente diverse dobbiamo utilizzare il **coefficiente di variazione**.

E' definito come il rapporto tra deviazione standard e la loro media.

E' privo di unità di misura

Posizione relativa Z-score

- Z-score indica quanto un singolo valore è lontano dalla media.

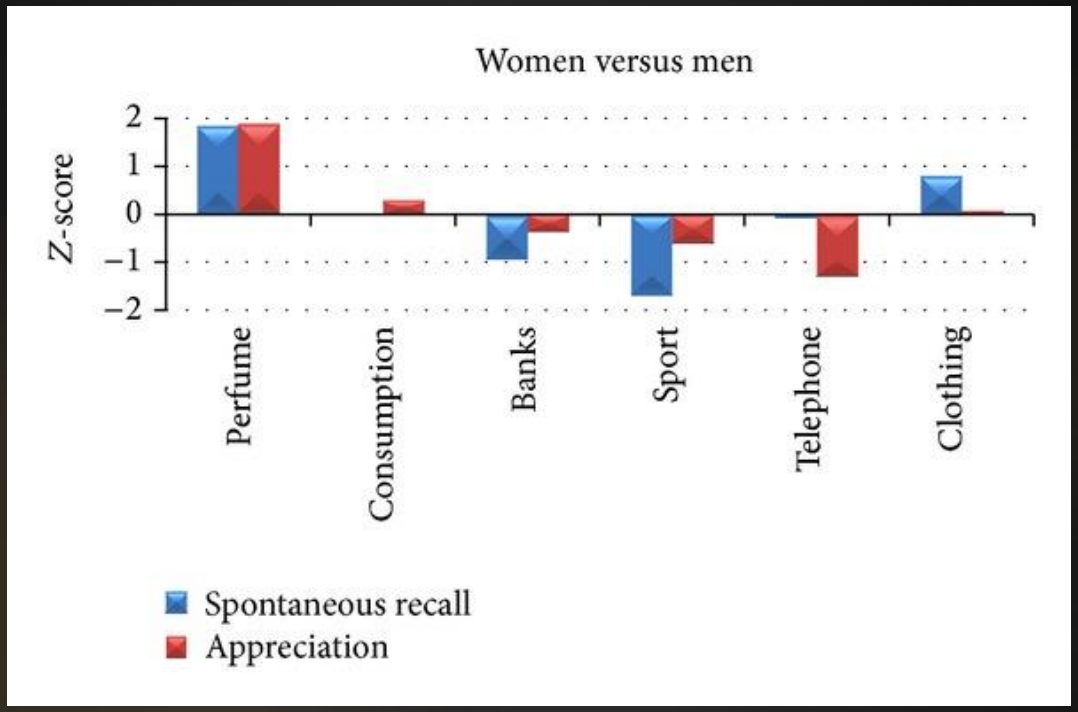
$$z = \frac{X - \overline{X}}{S}$$

where z is the standard score,
S = the standard deviation of a sample,
X = each value in the data set,
$\overline{X}$ = mean of all values in the data set.

Si tratta di un metodo per normalizzare i dati che si trovano su scale differenti.

I valori rappresentano quanto ogni singolo valore differisce dalla media

Z-score



Sono un modo efficacie per standardizzare i dati che appartengono a dataset differenti

Correlazione tra dati 1/2

- La capacità di ottenere, ripulire e tracciare dati aiuta a raccontare la storia che i dati possono offrire.
- Per comprendere come sono correlati i dati è necessario introdurre il **coefficiente di correlazione**.
- Il coefficiente di correlazione indica l'intensità con cui due variabili sono correlate. Quanto 2 variabili si muovono insieme. Alterando una delle variabili come varierà l'altra.
- I coefficienti di correlazione vanno da -1 a 1.
  - La minima correlazione è 0.
  - -1 indica che se aumenta una variabile l'altra diminuisce

Correlazione tra dati 2/2

- La correlazione potrebbe non essere lineare.
- Esistono funzioni predefinite in alcune librerie che permettono di stabilire dei coeffienti di correlazione all'interno di matrici contenenti dati (libreria Pandas).
- In realtà per comprendere le correlazioni sono necessari metodi di machine learning più sofisticati.