Statistica descrittiva

Esercitazione in laboratorio n. 0

Conservare il programma `R` ordinato e ben commentato

I dati da esaminare sono tratti dalla libreria Machine Learning Repository `https://rpubs.com/Ashutosh_kr/352670`. Sono estratti dal database dei censimenti statunitensi del 1996. Riguardano caratterisitiche economiche e sociali di 16281 persone adulte degli Stati Uniti.

I dati sono contenuti nel file `adult.txt`. *Sono separati da vigola. I valori mancanti sono indicati con il punto interrogativo (?). La prima riga non deve essere letta. Non ci sono i nomi delle variabili in testa al file.*

Le variabili considerate sono, nell'ordine in cui sono presenti nel file, le seguenti:

1. `age`: continuous.
2. `workclass`: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
3. `fnlwgt`: continuous.
4. `education`: Preschool, 1st-4th, 5th-6th, 7th-8th, 9th, 10th, 11th, 12th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, Some-college, Bachelors, Masters, Doctorate.
5. `education-num`: continuous.
6. `marital-status`: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
7. `occupation`: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
8. `relationship`: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried. Ruolo nella famiglia.
9. `race`: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
10. `sex`: Female, Male.
11. `capital-gain`: continuous. Plusvalenza: aumento di valore entro un determinato periodo di tempo di beni immobili e di valori mobiliari
12. `capital-loss`: continuous. Minusvalenza: differenza negativa, in un determinato periodo di tempo, del valore di un'attività finanziaria o di un'attività reale, come valori mobiliari o beni immobili.
13. `hours-per-week`: continuous.
14. `native-country`: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
15. `income`: <=50K., >50K.

---

1. Assegnare una cartella di lavoro (istruzione `setwd`). Creare un data frame contenente i dati.
2. Assegnare un nome alle colonne del data frame (il nome deve essere quello indicato nell'elenco sopra (istruzione `colnames`).
3. Visualizzare le dimensioni del data frame.

4. Visualizzare la struttura del data frame. Se variabili qualitative fossero lette come `chr` (in qualche versione di R), trasformarle in `Factor`; ad esempio `dati$workclass=as.factor(dati$workclass)`.

5. Visualizzare le prime 10 righe del data frame sulla finestra `console`. Visualizzare il data frame (istruzione `View`).

6. Visualizzare i livelli di `education`. Controllare se coincide o meno con l'ordine indicato nella descrizione sopra.

7. Costruire un nuovo fattore di nome `education_rec` ricodificando come fattore ordinale la variabile `education`, utilizzando come ordine dei livelli quello indicato nella descrizione sopra.

8. Costruire un data frame di nome `fattori` con solo le variabili qualitative omettendo le righe con valori mancanti (istruzione `na.omit`);visualizzare le prime 10 righe.

9. Considerare il tipo di lavoro (`workclass`). Visualizzare i livelli della variabile. Costruire una tabella di contingenza con le frequenze assolute, una tabella con le frequenze percentuali e un diagramma a barre. Procedere in modo analogo per altre variabili qualitative.

10. Costruire una tabella di contingenza "a due vie" (congiunta - bivariata) con le frequenze assolute e una tabella con le frequenze percentuali per `workclass` e `education_rec`.

11. Costruire un nuovo data frame ordinando le unità sperimentali secondo `workclass` in ordine discendente. Visualizzare le prime 5 righe e le ultime 5 righe del data frame.

12. Usando il data frame del punto precedente costruire due data frame per i due livelli del genere, e successivamente costruire un data frame che li concateni (cioè li scriva in sequenza). Per ciascun data frame visualizzare il numero di righe e di colonne.