Analisi dei dati con R

Docente: Maria Chiara Debernardi

Lingua

Italiano

Descrizione del corso e obiettivi

R è un linguaggio di programmazione open source che è stato sviluppato specificatamente per l'analisi dei dati. Grazie ai pacchetti gratuiti che ne espandono le capacità, consente di risolvere in modo semplice qualunque tipo di problema statistico, dalle analisi descrittive e inferenziali, ai problemi di data mining, machine learning e text analysis. Per questi motivi è sempre più adottato non solo in ambito accademico, ma anche lavorativo.
Il corso intende fornire un'introduzione all'utilizzo del linguaggio R, presentando le caratteristiche e i principali pacchetti che R mette a disposizione per l'analisi dei dati all'interno dell'ambiente di sviluppo RStudio.
Durante il corso verrà presentata l'architettura del linguaggio e la filosofia dell'analisi dati in esso implementata (con particolare riferimento al *tidyverse*), senza trascurare riferimenti alle sue ampie potenzialità di applicazione in ambito economico e aziendale attraverso la presentazione di specifici pacchetti di analisi dedicati al Data Mining.

Al termine del corso i partecipanti saranno in grado di:

- Manipolare i principali oggetti disponibili in R
- Effettuare analisi descrittive usando tabelle e grafici
- Costruire e leggere i risultati di intervalli di confidenza, verifiche d'ipotesi, modelli di regressione lineare e Anova
- Creare semplici funzioni ad hoc in linguaggio R
- Utilizzare le funzionalità dei principali package del *tidyverse*

Attenzione: lo scopo del corso è presentare il linguaggio R e le sue caratteristiche principali. Non può essere considerato sostitutivo di un corso formale in Statistica, dato che i dettagli delle specifiche metodologie utilizzate non saranno trattati.

Destinatari

Il corso è aperto a tutti gli studenti Bocconi. In particolare si rivolge agli iscritti dal secondo anno di triennale in avanti, interessati a realizzare analisi statistiche sia per la redazione del progetto di tesi, sia nella propria futura attività lavorativa.

Prerequisiti

È fondamentale che i partecipanti abbiano frequentato e superato positivamente il corso di informatica previsto dal proprio piano di studi, o possiedano competenze equivalenti.
È inoltre auspicabile una buona conoscenza dei fondamenti di statistica descrittiva e inferenziale, corrispondenti al primo esame di Statistica del proprio piano di studi.

Durata

16 ore

Modalità didattica

Sarà possibile partecipare al corso esclusivamente in maniera presenziale.

Calendario

Lezione	Data	Ora	Aula
1	ven 03/11/2023	14.45 – 16.15	N31 (Velodromo)
2	ven 03/11/2023	16.30 – 18.00	N31 (Velodromo)
3	ven 10/11/2023	14.45 – 16.15	N31 (Velodromo)
4	ven 10/11/2023	16.30 – 18.00	N31 (Velodromo)
5	mer 15/11/2023	18.15 – 19.45	N31 (Velodromo)
6	ven 17/11/2023	14.45 – 16.15	N31 (Velodromo)
7	ven 17/11/2023	16.30 – 18.00	N31 (Velodromo)
8	mer 22/11/2023	18.15 – 19.45	N31 (Velodromo)

Nota: le lezioni saranno tenute in aula tradizionale ed **è previsto che ciascuno studente disponga del proprio computer portatile**.

Programma delle lezioni

Lezione	Argomenti	Riferimenti bibliografici
1	**Introduzione a R e RStudio** - R e l'ambiente di sviluppo RStudio - I package e i siti CRAN - L'help - Primi passi con R - Gli script ***Esercizi***	**Capp. 1, 4, 6, 8**
2	**I dati in R** - I tipi di dati elementari: numeri e stringhe - La struttura base in R: i vettori - Le strutture complesse: matrici e array, liste, data frame - Gestione dei formati (le conversioni) - I factors in R ***Esercizi***	**Capp. 7, 10, 14, 15**
3	**Prima lettura dei dati** - Il *tidyverse* - Importazione (ed esportazione) di dati - Distribuzioni di frequenze - Statistiche di sintesi uni e bi-variate - Il package *DataExplorer* ***Esercizi***	**Capp. 7, 11, 12, 20**
4	**Rappresentare e manipolare i dati** - Il package *dplyr* per manipolare i dataset - La logica del chaining mediante le *pipes* - Rappresentare i dati attraverso i grafici - I principali pacchetti grafici di R ***Esercizi***	**Capp. 3, 5, 7, 18, 23, 28**
5	**L'inferenza in R** - Come si realizza un'analisi statistica - Intervalli di confidenza e verifiche d'ipotesi - La regressione lineare (e l'Anova) - Preparazione delle variabili: le trasformazioni - La gestione dei missing - Analisi e trattamento degli outlier ***Esercizi***	**Capp. 22, 23, 24, 25**

Lezione	Argomenti	Riferimenti bibliografici
6	**Programmare con R** - Il linguaggio R: le strutture di programmazione - Creare le proprie funzioni ad hoc in R ***Esercizi***	**Capp. 19, 21**
7	**Analisi delle serie temporali** - Il tempo in R - Il package *lubridate* - Analisi esplorativa e modelli autoregressivi: i pacchetti specifici di R ***Esercizi***	**Capp. 16, 24**
8	**Data Mining in R** - Alcuni problemi tipici e come affrontarli - Il package *rattle* per il DM - Pacchetti per le analisi più avanzate (hints): *caret* e *lime* ***Esercizi***	**V. slide della lezione**

Bibliografia consigliata

H. Wickham, M. Çetinkaya-Rundel, G. Grolemund, *R for Data Science, 2nd Edition*, O'Reilly Media, 2023
consultabile gratuitamente online: r4ds.hadley.nz

Della prima edizione del testo (2017) esiste anche una traduzione in italiano, disponibile gratuitamente online: it.r4ds_1ed.hadley.nz

Software di riferimento

1) Linguaggio R (r-project.org): ultima release disponibile (4.3.1 o superiore)
   download, in base al proprio Sistema Operativo, da cran.stat.unipd.it
2) Ambiente di sviluppo RStudio (rstudio.com): ultima release disponibile della versione Desktop – Open Source License/Free (2023.09.0+463 o superiore)
   download la versione specifica per il proprio SO da rstudio/download

Posti disponibili

Questa attività è a numero chiuso quindi l'iscrizione non sarà possibile oltre **110 posti** o dopo la chiusura del periodo di iscrizione.
È possibile annullare l'iscrizione esclusivamente tramite agenda yoU@B entro e non oltre il termine del periodo di iscrizione al corso stesso.