Analisi dei dati per l'impresa

06/06/25 - Punti max: 24

Nome Cognome:	matricola

1. Dato l'oggetto `df <- data.frame(col_1=10:13, col_2=c("a", "b", "c", "d"))`, quale operazione è corretta per aggiungere la variabile `col_3 = c(5, 1, 18, 16)`?

   - ☐ `df <- dplyr::mutate(col_3 = c(5, 1, 18, 16))`
   - ☐ `df[,3] <- col_3 =c(5, 1, 18, 16)`
   - ☐ `df$col_3 <- c(5, 1, 18, 16)`

2. Quale dei seguenti 'nomi' è corretto assegnare ad una variabile?

   - ☐ `qui quo qua`
   - ☐ `.quiquoqua`
   - ☐ `qui@quo@qua`

3. Quale operazione è corretta per estrarre da `df <- data.frame(a=0:4, b=10:14, c=20:24)` le righe in posizione pari della sottomatrice costituita da a e b?

   - ☐ `df_new <- df %>% select(a, b) %>% filter(a %% 2 == 0)`
   - ☐ `df_new <- df %>% slice(c(2,4)) %>% select(a, b)`
   - ☐ `df_new <- df[df$a %in% c(0,2,4), -3]`

4. Dopo aver trasformato il vettore `v = c("a", "b", "a", "c", "b")` in fattore, quale comando è corretto per cambiare il primo livello del fattore in "e"?

   - ☐ `v[which(v=="a")] <- "e"`
   - ☐ `levels(v)[1] = "e"`
   - ☐ `v <- factor(v, levels = c("e", "b", "c"))`

5. Dato il vettore `x=c(-1.2,0,10,6.2,3,-3,9,12,sqrt(2))`, per estrarre i valori positivi e minori di 10, possiamo usare il comando

   - ☐ `x[0<x<10]`
   - ☐ `x>0 & x<10`
   - ☐ `x[x>0 & x<10]`
   - ☐ `x[x>0 | x<10]`

6. Dato il fattore `x <- factor(letters[1:4])`, e definiti i livelli con `levels(x)<-rev(levels(x))`, se si trasforma `x` nel vettore dei codici (interi) identificativi dei livelli si ottiene

   - ☐ `1 2 3 4`
   - ☐ `"d" "c" "b" "a"`
   - ☐ `4 3 2 1`

7. Se creo il `vettorex<-c(1,"ok")` ottengo l'oggetto

   - ☐ `c("1", "ok")`
   - ☐ `c(1, "ok")`
   - ☐ `c(1, NA)`

8. Si considerino le variabili `X1=grade` (grado di rischiosità, con livelli A, B, C e D) e `X2=homeownership` (proprietà della casa, le cui categorie sono "own", "mortgage" e "rent") del dataframe `loans`. Per visualizzare la distribuzione in frequenze relative di `X1` condizionatamente a `X2` si può usare il comando di ggplot2

   - ☐ `ggplot(loans, aes(x = homeownership, y = grade)) + geom_boxplot()`
   - ☐ `ggplot(loans, aes(x = homeownership, fill = grade)) + geom_bar()`
   - ☐ `ggplot(loans, aes(x = homeownership, fill = grade)) + geom_bar(position = "fill")`

9. La distanza di Minkowskysi utilizza per variabili quantitative.
   V F

10. Ordina le seguenti fasi del metodo k-means (indica nelle caselle i numeri da 1 a 4):

    - ☐ Aggiornamento dei centroidi: si calcolano le medie dei gruppi
    - ☐ Aggiornamento dei clusters: ogni unità è assegnata al cluster più vicino
    - ☐ Si iterano i due passi precedenti fino a convergenza
    - ☐ Le unità vengono assegnate casualmente a k gruppi e si calcolano le medie dei gruppi

11. Ss considerino le seguenti istruzioni in R che caricano i dati "iris" e creano una matrice con le variabili numeriche:

```
library(cluster)
data(iris)
iris<-as.matrix(iris[,1:4])
```

    Quale delle seguenti istruzioni calcola i centroidi dei gruppi che si ottengono dall'algoritmo k-means con 4 clusters?

    - ☐ `kmeans(iris, centers=4)$means`
    - ☐ `kmeans(iris, centers=4)$centers`
    - ☐ `kmeans(iris, k=4)$centers`
    - ☐ `kmeans(iris, centers=4)[[centers]]`

12. Si consideri il data set "diamond" del pacchetto `UsingR` contenente il prezzo di 48 anelli di diamanti (in dollari di Singapore) e la dimensione del diamante in carati. Si stima un modello di regressione lineare con Y=price e X=carat.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -259.63      17.32  -14.99   <2e-16 ***
carat        3721.02      81.79   45.50   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In base alle stime riportate nella figura, si può affermare che

- ☐ la variazione attesa stimata del prezzo dei diamanti è pari a circa -259.63 dollari SIN per ogni aumento del numero di carati
- ☐ l'aumento del numero di carati non è significativo per un aumento di prezzo
- ☐ il coefficiente di `carat` rappresenta la variazione percentuale del prezzo per ogni aumento unitario di carati
- ☐ l'aumento atteso stimato del prezzo dei diamanti è pari a circa 3721 dollari SIN per ogni aumento unitario di carati

13. Si considerino i dati "Insurance" nel pacchetto `MASS` che si ottengono mediante l'istruzione `data(Insurance)`, relativi atitolari di polizze di una compagnia assicurativa e il numero di richieste di risarcimento per auto (terzo trimestre 1973). Le variabili sono: Claims, District, Group, Age, Holders. Quale misura di distanza/dissimilarità è più appropriata per il dataset "Insurance"?

    - ☐ distanza euclidea
    - ☐ distanza di Manhattan
    - ☐ indice di Jaccard
    - ☐ distanza di Gower

14. Tra i diversi modi per misurare la distanza fra due gruppi nei metodi gerarchici vi è

    - ☐ il legame di Ward, che massimizza la varianza delle variabili entro ciascun gruppo
    - ☐ il legame completo, che calcola la distanza tra due gruppi come il minimo delle distanze tra le osservazioni dei due gruppi
    - ☐ il legame singolo, che calcola la distanza tra due gruppi come il minimo delle distanze tra le osservazioni dei due gruppi
    - ☐ il legame medio, che calcola la distanza tra due gruppi utilizzando la distanze tra i centroidi

15. L'intercetta della retta di regressione stima il cambiamento di Y per un'unità di incremento di X.
    V F

16. Un venditore di moto usate vuole studiare la relazione tra i Km effettuati e il prezzo delle moto. Viene stimato un modello di regressione lineare sulla base di un campione casuale di 100 moto; la stima dell'intercetta risulta pari a 4533 e coefficiente angolare pari a -0.0212. Qual è il valore della previsione per il prezzo di una una moto con 25000 Km?

    - ☐ 4003 euro
    - ☐ 5063 euro
    - ☐ 530 euro
    - ☐ 4533 euro