**כריית נתונים ב- R**

***מיני תרגיל מספר 2***

**שאלות:**

בתרגיל זה נתשמש במסד traindata.csv.

1. קרא את הקובץ לתוך מסד ב- R
2. הצג את סטטיסטיקת המסד (summary)
3. השלם ערכים חסרים במשתנה income בעזרת ממוצע ההכנסות
4. חשב את ממוצע ההכנסות לפי מגדר
5. דגום 20% מהנתונים (השתמש ב seed = 1) וחשב שנית ממוצע ההכנסות לפי מגדר במדגם

**פתרון:**

traindata <- read.csv("traindata.csv", stringsAsFactors = TRUE)

summary(traindata)

traindata$income[is.na(traindata$income)] <- mean(traindata$income, na.rm = TRUE)

aggregate(income ~ gender, data = traindata, FUN = mean)

set.seed(1)

aggregate(income ~ gender, data = traindata[sample(traindata$id, 20),], FUN = mean)