Data Frames

Practice

Before starting to solve the exercises below, please extract the CSV files from the csv.zip archive and put them in a separate folder on your hard disk. Then make this folder your working directory.

You can find the download link for the csv.zip archive in the last section of the course.

1. Create three vectors of the same length, two of them with discrete random values and one with continuous random values. Next, create a data frame with these vectors.

2. Save the data frame created at #1 in your working directory, without row names.

3. Create two data frames with the read.csv() function, using the files education.csv and phone.csv.

4. Create a data frame using the file phone.csv and name it phone. Then perform the following operations on this data frame:

   - access the value in the fifth row for the variable income
   - access the values in the rows 7 to 12, all the variables
   - access the values in the rows 7 to 12, variables age, income and churn
   - access the variable tenure (all the entries)
   - access the variables tenure, educ and churn (all the entries)

   Use all the possible indexing methods you know (as a list or as a matrix).

5. Select a random sample of 250 entries from the data frame created at #4. Store them in a new data frame phone2.

6. Select a random sample of 250 entries from the data frame created at #4, keeping only the variables tenure, income and members. Store them in a new data frame phone3.

7. Create a data frame using the file directmail.csv and name it mail. Then perform the following operations on this data frame:

    - select the entries where the age is greater than 30
    - select the entries where the age is greater than 30 and the region is West
    - select the entries where the age is lower than 50 and the variable children is equal to 1
    - select the entries where the education is college
    - select the entries where the education is college and the variable reside is greater than or equal to 7, keeping only the variables age, reside, gender and region
    - select the entries where the age is greater than 40 or the income is less than 25
    - select the entries where the gender is male or the variable reside is smaller than 10
    - select the entries where the gender is male or the variable reside is smaller than 10, keeping only the variables previous, income and children

8. In the data frame created at #7 perform the following operations:
    - in the second row, change the value of the variable previous to Yes
    - in the row 20, change the value of the variable reside to 5
    - change all the values in row 6 with the following values, respectively: Yes, 33, 75+, College, 10, Male, 1, East
    - change the values in row 10 as follows: income to 50-74, reside to 9, children to 1 and region to South
    - change all the values of the variable previous to Unknown

9. In the data frame phone created at #4 add a new entry with the following values, respectively: 38, 23, 116, 4, 3, 1.

10. In the data frame phone created at #4 add a new column containing normally distributed random values.

11. In the data frame phone created at #4 add a new column containing the ratio income/members. Afterwards delete this variable.

12. In the data frame phone created at #4 do the following operations:

    - name the third row Jack

- name the fifth, the ninth and the fourteenth row Mary, Paul and Christine, respectively
- name the rows 20 to 25 with the letters from a to f
- rename the variable educ into education.

13. In the data frame phone created at #4 compute the sum, the mean and the standard deviation for all the variables, using the appropriate functions in the apply() family.

14. In the data frame mail created at #7 compute the sum, the mean and the standard deviation for the numeric variables, using the appropriate functions in the apply() family.

15. Sort the data frame phone by age, ascending and descending.

16. Sort the data frame phone by tenure, ascending and descending.

17. Sort the data frame phone by age ascending and by tenure descending.

18. Sort the data frame phone by the variable churn ascending and by education descending.

19. Sort the data frame mail by the variable reside, ascending and descending.

20. Sort the data frame mail by education, ascending and descending.

21. Shuffle the data frames phone and mail.

22. Create four vectors as follows:

    - x = (1, 2, 3, 4, 5, 6, 7)
    - y = (100, 200, 300, 400, 500, 600, 700)
    - z = (2, 4, 1, 6, 5, 7, 3)
    - w = (TRUE, TRUE, FALSE, FALSE, TRUE, TRUE, FALSE)

    Next, create a data frame dt1 with the vectors x and y, and a data frame dt2 with the vectors z and w. Merge the two data frames into a new one called dt.

    What variable(s) have you used for merging?