“Processing large dataset with R”

Workshop 1

19/10/2021

1. Introduction, advanced R and programming basis
2. Handling massive data (dplyr)

Exercise 1

1a)create a vector v going from 1 to 10 with increments of 0.5.
1b)define a function "fahrenheit_to_celsius()" that converts temperatures from Fahrenheit to Celsius.
Following the Eq. : °C = (°F − 32) × ⅝
1c)from vector v create a sample,t, of size 100 with replacement where the probabilities are given by a random uniform distribution.
Which number occur the most?
Create an histogram from t with yellow cells that has as main title = "Sampling with replacements".
1d) Create a 7×8 matrix filled by row of Gaussian random numbers, with mean 1 and standard deviation 2.
Add 1 to the elements lower than the mean.
1e)Using the iris datasets make a boxplot of the variables in the dataset, last column excluded.
Make a matrix of scatterplots of the previous variables where color and form of the points depend on the species
1f)Plot with different colors the estimated density functions of four random gamma distribution
with n =1000 and scale parameter respectively: 1, 2 , 3, 4.

Exercise 2

install.packages("nycflights13") ; install.packages("tidyverse"); library(dplyr)

we'll use nycflights13::flights. This data frame contains all 336,776 flights that departed from New York City in 2013. The data comes from the US Bureau of Transportation Statistics, and is documented in ?flights.

2a)Create a new data frame that contains only the flights on 8 April 2013.

Find the flight with the lowest departure delay.

2b) How many flights were delayed (on arrival or departure) by more than four hours?

Create a new data frame containing the dep_delay variable for the flights with the highest departure delay for each month,

the month, the day and the departure delay monthly average.

2c) Considering only the flights that landed in LAX airport, show the departure delay and the arrival delay, ordered according to the departure delay in a descendent order. Then compute the column mean.

Exercise 3

3a) Create a dataframe of 100 people where:

- the first column represents the age, distribuited as a random uniform,from 20 to 40 years old,

- the second column represents the Weight, distributed as a random uniform, from 50 to 90 kg with one decimal,
- the third column if they are graduated or not, respectively with a proportion of 60% and 40%.

3c) Insert 5 missing values in each column at random locations, with a for loop.

3d) Change column name "Graduated" to "Driving_License".
Count the number of missing values in the dataframe. Now remove them.

3e) Make a Min-Max Normalization: $(X - \min(X)) / (\max(X) - \min(X))$ of the first two columns

3f) Make a a z-score standardization: $(X - \mu) / \sigma$ of the first two columns.