Faculty of Computers and Artificial Intelligence
Bioinformatics Program
2020/2021
Biology Statistics Course

Assignment 1
R Practice+Exploratory Data Analysis

Medical Appointments Dataset

This dataset is drawn from 300,000 primary physician visits in Brazil across 2014 and 2015. The information about the appointment was automatically coded when the patient scheduled the appointment and then the patient was marked as having either attended or not. The information about the appointment included demographic data, time data, and conditions concerning the reason for the visit.

Variables:

1. **Age**: integer age of patient
2. **Gender**: M or F
3. **ScheduleDay** : date and time appointment was made
4. **AppointmentDay**: date of appointment without time
5. **Neighborhood**: Hospital area
6. **Diabetes**: 0 or 1 for condition (1 means patient was scheduled to treat condition)
7. **Alcoholism**: 0 or 1 for condition
8. **Handcap**: 0 or 1 for condition
9. **Hypertension**: 0 or 1 for condition
10. **Scholarship**: 0 or 1: indicating whether the family of the patient takes part in the initiative that provides families with small cash transfers in exchange for keeping children in school and completing health care visits
11. **SMSReminder**: 0 ,1 ,2 for number of text message reminders sent to patient about appointment
12. **No-show**: Yes or No: Did the patient showed up at the appointment

Tasks:

Using R, you are requested to read correctly the provided dataset and perform the following tasks:

1. Decide on the correct class of each column (which should be numeric/factor/character, …)
2. Write a code to find out:
   2.1. Number of appointments in every neighborhood ➔ show result as a table
   2.2. Number of appointments for every patient ➔ show by histogram
   2.3. The distribution of ages ➔ density plot
   2.4. Are there any columns with missing values? print the number of rows and which columns
   2.5. Are there any columns with wrong values (For example a column with unexpected value with respect to column description)? print the values /their number and which columns
   2.6. Compare the distribution of ages among males and females using line plots
   2.7. Average number of visits for every patient ➔ print as table
   2.8. Which age range has the higher number of visits? (Use code only to identify different groups: children/teenagers/adults/elderlies)
   2.9. Which age ranges showing higher absence rates (didn't show up at vists)?
   2.10. plot how the number of visits changes on monthly basis. (You will need to treat the character dates as date object)

Rules:

- Deadline is Tuesday 12:00 AM 16 Nov
- Discussion would be announced later by each TA
- Solve this assignment in groups of 2