Beadandó feladatok R-ből
2013. tavasza

Összesen három függvényt és egy kódot kell írnia. A beadáskor a kódot ill. a függvényeket kell elküldenie emailben (csima@cs.bme.hu), beadási határidő május 3., péntek 12 óra, a beadott kódokat az utolsó laboron kell elmagyaráznia. Csak olyan függvényt illetve kódot küldjön be, ami hiba nélkül lefordult. Ha nem sikerül teljesen megoldani egy feladatot (nem tud mindent a függvény, amit kellene neki), de ér el részeredményt, azt is érdemes beküldeni, részpontokat is lehet szerezni.

Töltse le a specdata.zip file-t a weboldalról a saját working directoryjába. Ez a tömörített file 332 CSV file-t tartalmaz, melyek mindegyike egy-egy USA-beli mérőállomás adatait tartalmazza (a file-ok egyike volt a 2. laboron használt 001.csv is). Tömörítse ki a specdata.zip file-t a specdata directoryba, az első három beadandó feladatnál ezekkel a file-okkal kell majd dolgoznia.

1. (2 pont) Írjon egy darab.R nevű függvényt, melynek egy argumentuma van, ennek neve `also`, és a függvény azt számolja ki, hogy hány olyan sora van a 017.csv file-nak, melyben a nitrát értéke nagyobb, mint `also`.

   Néhány minta input-output pár, amit reprodukálnia kell a függvénynek:

   ```
   > darab(-3)
   [1] 962

   > darab(1.3)
   [1] 217
   ```

2. (5 pont) Írjon egy maximum.R nevű függvényt, aminek két argumentuma van: az első neve `sorszam`, ez egy mérőállomás azonosító (háromjegyű számként megadva, 001 és 332 között), a második neve pedig `korlat`, ez egy egész szám. A függvény először ellenőrizze, hogy a megadott `sorszam` érték a kívánt 1-332 tartományból kerül-e ki, és ha nem, akkor írja ki, hogy `hibas sorszam`. Aztán nézze meg, hogy van-e legalább `korlat` darab teljes sor (ahol minden érték adott) a `sorszam`.csv file-ban: ha nincs, akkor írja ki, hogy `tul keves adat`. Ha mindkét argumentum rendben van, akkor pedig írja ki az adott `sorszam`.csv file-ban szereplő maximális szulfát értéket és azt, hogy melyik napon volt ez (nem csak a teljes sorok számítanak itt). Ha több napon is felvevődött a maximális érték, akkor az összeset listázza ki.

   Segítség:
   1) nézze meg a `paste` függvényt
   2) nézze meg a match függvény leírásában szereplő %in% bináris operátort

   Néhány minta input-output pár, amit reprodukálnia kell a függvénynek:

   ```
   > maximum(1000, 2)
   Error in maximum(1000, 2) :  hibas sorszam

   > maximum(100, 2000)
   Error in maximum(100, 2000) :  tul keves adat

   > maximum(100, 2)
           Date     sulfate
    247  2008-09-03    11.1

   > maximum(001, 2)
           Date     sulfate
    987  2005-09-13    19.1

   > maximum(055, 10)
           Date     sulfate
    304  2000-10-30    15.1
    310  2000-11-05    15.1
   ```

3. (2 pont) Módosítsa az első feladatban megadott függvényt úgy, hogy ne csak a háromjegyű számként megadott `sorszam` értéket ismerje fel, hanem az egy- illetve kétjegyű mérőállomás-azonosítókat is tudja kezelni.

   Néhány minta input-output pár, amit reprodukálnia kell az új függvénynek:

```
> maximum(10, 2)
         Date     sulfate
 497  2003-05-12    2.27

> maximum(1, 2)
         Date     sulfate
 987  2005-09-13    19.1
```

4. (6 pont) Töltse le a korhaz.zip file-t a working directoryjába és tömörítse ki. Két file lesz benne,

   - az outcome-of-care-measures.csv (sok minden más adat mellett) az összes amerikai kórházra megadja a kórházi felvételt követő 30 napon belüli halálozási adatokat heart attack, heart failure és pneumonia kategóriákban,
   - a Hospital_Revised_Flatfiles.pdf leírja több, a kórházak összehasonlítására alkalmas adatbázis szerkezetét (változók neve, típusa, stb.), egy ilyen adatbázis az előbbi outcome-of-care-measures.csv, ennek leírása a 19. pontban található (17-20. oldal).

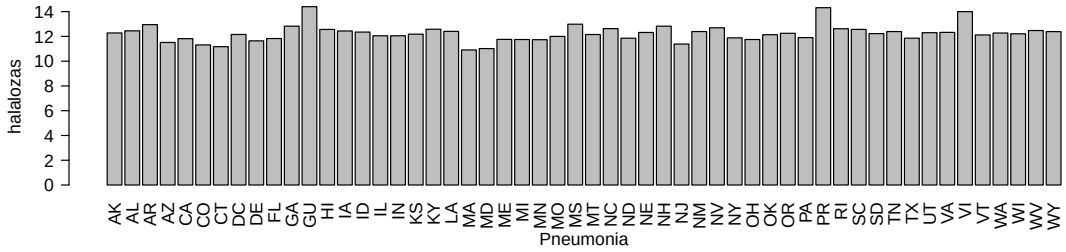
   Az adatbázist a data.medicare.gov oldalról szedtem, itt található a pdf file-ban leírt további rengeteg adatbázis is, de ezekkel most semmi dolgunk nem lesz. (Sajnos magyarul semmi hasonlót nem találtam, ezért dolgozunk az amerikai adatokkal.)

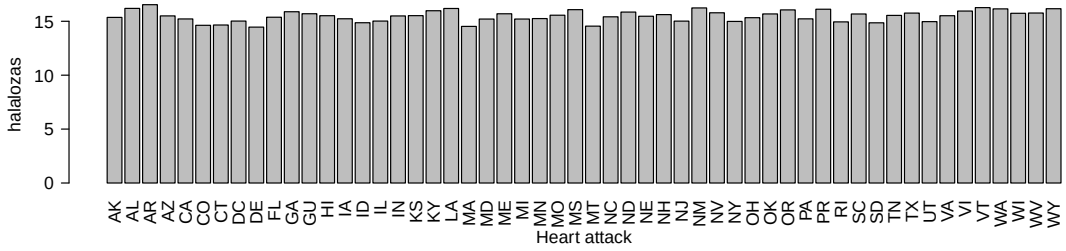
   Az outcome-of-care-measures.csv file adatait felhasználva készítsen három oszlopdiagrammot a három kategóriára (heart attack, heart failure és pneumonia), melyeken azt ábrázolja, hogy az egyes államokban mekkora a 30 napon belüli halálozás átlaga. (11., 17. és 23. mező adatai.)
   A három diagramm egymás alatt legyen, az y tengely felirata legyen `halalozas`, az x tengelyek feliratai legyenek a megfelelő betegségek.
   Az x tengelyeken szerepeljen minden állam rövidítése a megfelelő oszlop alatt, merőlegesen a tengelyre.

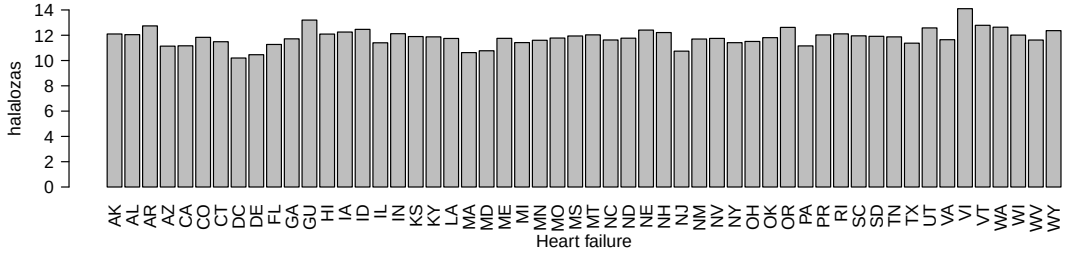
   Segítség:

   - a read.csv hívásakor célszerű beállítani, hogy colClasses =”character” és utána még gondoskodni róla, hogy amikor az átlagot számoljuk, akkor már numeric típusú legyen a megfelelő oszlop
   - a barplot függvény help-jét illetve a par függvény help-jét érdemes nézni

   A kódnak, amit beküld, a következő oldalon található ábrát kell előállítania:


halalozas
0
2
4
6
8
10
12
14
AK AL AR AZ CA CO CT DC DE FL GA GU HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS MT NC ND NE NH NJ NM NV NY OH OK OR PA PR RI SC SD TN TX UT VA VI VT WA WI WV WY
Pneumonia


halalozas
0
5
10
15
AK AL AR AZ CA CO CT DC DE FL GA GU HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS MT NC ND NE NH NJ NM NV NY OH OK OR PA PR RI SC SD TN TX UT VA VI VT WA WI WV WY
Heart attack


halalozas
0
2
4
6
8
10
12
14
AK AL AR AZ CA CO CT DC DE FL GA GU HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS MT NC ND NE NH NJ NM NV NY OH OK OR PA PR RI SC SD TN TX UT VA VI VT WA WI WV WY
Heart failure