Beadandó feladatok R-ből
2014. tavasz

Tudnivalók

Összesen három függvényt és egy kódot kell írnia, ezekben két letöltött és kitömörített adatbázissal kell dolgoznia. A beadáskor a kódot ill. a függvényeket kell elküldenie emailben a csima@cs.bme.hu címre.

A beadás előtt ellenőrizze a következőket (különben a beadás nem lesz sikeres):

1. A három függvényt és az ábrát előállító kódot négy külön R scriptben küldje, ezek nevei legyenek `darab.R`, `sorszam.R` (ha két részben oldja meg a második és harmadik faladatot akkor `sorszam1.R` és `sorszam2.R`) és `abra.R`.

2. A feladatmegoldás során csak az `rstudio` alap package-eit használja és esetleg a `lattice`-t vagy a `ggplot2`-t, ha a grafikai feladatot ezzel oldja meg.

3. A scriptekben az adatfile-ok elérése relatív path-szal történjen. Ehhez az kell, hogy a letöltött `specdata.zip` és `korhaz.zip` file-okat a working directory-jába tömörítse ki, így ott létrehozva egy `specdata` directory-t és egy `outcome-of-care-measures.csv` file-t. (Ha simán kitömöríti a letöltött zip-file-okat, akkor ez így lesz.) Ezeket relatív hivatkozással érje el a scriptekből, amit ír, ez azt jelenti, hogy például az első feladatban a `read.csv` függvényt `read.csv("./specdata/017.csv")` alakban hívja meg, a `read.csv("C:/specdata/017.csv")` nem jó.

4. Ne használjon `for`-ciklust! Minden feladat megoldható enélkül, pusztán az órán tanult ismeretek (és egy-két új függvény) segítsével. Ha mégis `for`-ciklust használ, akkor az elért pontjainak 60%-a lesz a kapott pontszáma.

5. Csak olyan függvényt illetve kódot küldjön be, ami hiba nélkül lefordult ill. generálta a kívánt ábrát.

Beadási határidő a páratlan heti csoportnak **április 29., kedd, 22 óra**, a páros heti csoportnak **május 6., kedd, 22 óra**. A beadott kódokat az utolsó laboron kell elmagyaráznia, ennek célja az, hogy meggyőződjünk arról, hogy a kód saját munka. Ha ennek ellenkezője derül ki bármilyen módon, az elégtelen osztályzatot jelent. 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.
Ha bármi kérdése van a beadással kapcsolatban, akkor emailben keressen nyugodtan (csima@cs.bme.hu).

Feladatok

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 (a working directory-n belül), 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. Ha több napon is felvevődött a maximális érték, akkor az összeset listázza ki.

   Segítség (egy lehetséges megoldáshoz):
   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 (egy lehetséges megoldáshoz):

- 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 ábrához hasonlót kell előállítania. Ez az ábra a base grafikai csomaggal készült, ha Ön is ezt használja, akkor ezt kell reprodukálnia, ha másik grafikai csomagot használ, akkor egy olyan ábrát kell előállítania, ami ugyanezeket az adatokat ugyanígy ábrázolja, ugyanilyen elosztásban.