Tilastotiede ja R tutuksi I, syksy 2019
Viikon 5 RTMC-tehtävät

Lataa viidennen viikon tehtäväsetit Week5-Exercises0 ja Week5-ExercisesB. Jos jostain syytä et saa TMC-liitännäistä käyttöön, voit ladata tehtäväpohjat kurssin TMC-nettisivuilta ja palauttaa tehtävät serverille zip-palautuksena.

Setti 0. Tästä setistä ei saa pisteitä, mutta tehtäviä voi tehdä oman oppimisen tukemiseksi. Käytä täsmälleen samoja muuttujanimiä tehtävien ratkaisuissa!

1. Luo taulukko `dat` (tehtäväpohjassa valmiina olevalla) komennolla:

```
dat <- data.frame(x1=c(2, 1, 3, 1, 1, 1, 1, 3, 1, 1, 1, 3, 3, 2, 1),
                  x2=c(-2.19, 3.78, -0.03, 0.25, 0.39, -3.18, -0.77,
                       -0.52, -1.41, 1.03, -0.24, -5.11, 3.35, -5.49,
                       -1.78))
```

   (a) Määritä funktion `cut` avulla faktorivektori `x3`, jossa on kaksi tasoa `"neg"` ja `"pos"`, ja jonka kunkin elementin taso määräytyy sen mukaan, saako taulukon `dat` muuttuja `x2` negatiivisen vai positiivisen arvon (nolla ajatellaan tässä positiiviseksi). Voit nimetä tasot `cut`-funktion argumentin `labels` avulla. Vasemmanpuoleiset suljetut välit saa argumentilla `right=FALSE`.

   (b) Liitä vektori `x3` taulukkoon `dat` ”x3” nimiseksi sarakkeeksi.

   (c) Laske muuttujan `x2` keskiarvo erikseen niille riveille, joiden `x3`-taso on `"neg"`, ja niille joiden taso on `"pos"`. Tallenna keskiarvot kaksipaikkaiseen vektoriin `a1` järjestyksessä ”neg”, ”pos”. Vihje: `aggregate`

   (d) Ristiintaulukoi muuttujat `x1` ja `x3` funktiolla `table` siten, että ensimmäisellä rivillä on `x3`:n tason ”neg” frekvenssit kullekin muuttujan `x1` arvolle (järjestyksessä 1, 2, 3), ja toisella rivillä on tason ”pos” frekvenssit. Tallenna tuloksena oleva taulukko muuttujaan `tab1`.

Setti B. Käytä täsmälleen samoja muuttujanimiä tehtävien ratkaisuissa!

1. Tarkastellaan R:n mukana tulevan `datasets` paketin aineistoa `iris`. Kuvauksen aineistosta saat näkyviin komennolla `help(iris, datasets)`.

   (a) Laske muuttujien `Sepal.Length` ja `Sepal.Width` välinen sekä muuttujien `Petal.Length` ja `Petal.Width` välinen (Pearson) korrelaatiokerroin, ja tallenna tulokset muuttujiin `corsepal` ja `corpetal`. Vihje: `cor`.

   (b) Laske a-kohdan korrelaatiokertoimet kullekin lajille (`Species`) erikseen, ja tallenna tulokset kolmipaikkaisiin vektoreihin `corsepal2` ja `corpetal2` siten, että korrelaatiokertoimet ovat järjestyksessä: setosa, versicolor, virginica. Vihje: tehtävän voi ratkaista monella eri tavalla. Esimerkiksi valmiilla funktiolla `by` voit laskea korrelaatiokertoimet osa-aineistoittain (Kannattaa lukea `by`:n dokumentaatio huolellisesti. Funktion dokumentaatioon pääsee komennolla `help(by)`.) Eräs toinen vaihtoehto on käydä lajit läpi vaikkapa `for`-silmukalla siten, että kunkin lajin kohdalla muodostat osa-aineiston, joka sisältää vain kyseistä lajia koskevat havainnot, ja sitten lasket halutut korrelaatiot osa-aineiston muuttujille. Tehtävä voi aluksi tuntua hankalalta, mutta erityisesti jälkimmäisen esimerkkiratkaisutavan voi jakaa pieniin osiin.

2. Tarkastellaan R:stä valmiiksi löytyvän `datasets` paketin aineistoa `state.x77`, jossa on tietoja Yhdysvaltain osavaltioista 1970-luvulta. Kuvauksen aineistosta saat näkyviin komennolla `help(state.x77, datasets)`. **Huomaa**, että `state.x77` on luokaltaan matriisi eikä taulukko.

   (a) Kopioi `state.x77` uuteen muuttujaan `mystate.x77` siten, että muutat sen samalla taulukoksi (`data.frame`), jonka rivi- ja sarakenimet ovat samat kuin matriisilla `state.x77`. Vihje: `as.data.frame`

   (b) Määritä luomaasi `mystate.x77` -aineistoon `cut`-funktion avulla (”Murder2” nimiseksi) sarakkeeksi uusi muuttuja `Murder2`, jossa osavaltiot on luokiteltu murha-asteen (`Murder`, murhia per 100000 asukasta) mukaan ryhmiin $[0, 5)$, $[5, 10)$ ja $[10, 100000)$. Anna muuttujan tasoille nimet `"low"`, `"medium"` ja `"high"`. Vihje: Vasemmalta puolelta suljetut välit saa `cut`-funktion argumentilla `right = FALSE`, ja funktiossa tasojen nimet saa määriteltyä `labels`-argumentin avulla.

   (c) Määritä luomaasi `mystate.x77` -aineistoon `cut`-funktion avulla (”LifeExp2” nimiseksi) sarakkeeksi uusi muuttuja `LifeExp2`, jossa osavaltiot on luokiteltu elinajanodotteen mukaan ryhmiin $(67, 70]$, $(70, 72]$ ja $(72, 74]$. Anna muuttujan tasoille nimet `"low"`, `"medium"` ja `"high"`.

   (d) Ristiintaulukoi muuttujat `Murder2` ja `LifeExp2` funktion `table` avulla siten, että ylimmällä rivillä on elienajanodoteryhmän ”low” frekvenssit kullekin murharyhmälle (järjestyksessä ”low”, ”medium”, ”high”), keskimmäisellä rivillä ryhmän ”medium” frekvenssit ja alimmalla ”high”:n. Tallenna tuloksena oleva taulukko muuttujaan `tab1`. **Huomaa**, että `tab1` ei ole luokan `data.frame` olio, vaan luokan `table` olio (tämä ei sinänsä vaikuta tehtävän ratkaisemiseen, mutta suomennoksen ”taulukko” kaksoismerkitys voi aiheuttaa hämmennystä).

3. **Huom!** Tätä tehtävää ei tarkasteta RTMC:llä, vaan piirretyt kuvat tarkastetaan manuaalisesti. Liitä pyydetyt kuvat kurssin MOOC-sivuille palautettavan pdf-tiedoston loppuun. Kuvat piirtävä koodi tulee kuitenkin kirjoittaa RTMC:n tehtäväpohjalle ja submitoida serverille. Kuvat voi piirtää painamalla ”Source”-painiketta.

   Tarkastellaan jälleen `datasets` paketin `state.x77` aineistoa.

   (a) Piirrä hajontakuvamatriisi, jossa on aineiston neljän ensimmäisen sarakkeen (Population, Income, Illiteracy ja Life Exp) väliset hajontakuvat. Vihje: `pairs`.

   (b) Piirrä tarkempi hajontakuva (eli hajontakuva joka ei ole osana hajontakuvamatriisia) muuttujien `Illiteracy` ja `Life Exp` välisestä suhteesta siten, että kukin havainto on pisteen sijasta merkitty sitä vastaavan osavaltion nimellä. Osavaltioiden nimet löytyvät komennolla `rownames(state.x77)` tai `datasets`-paketin muuttujasta `state.name`. Otsikoi vaaka- ja pystyakseli sopivalla tavalla. Vihje: eräs ratkaisu on piirtää ensin tyhjä kuva käyttämällä argumenttia `type = "n"`, ja lisätä sitten osavaltioiden nimet oikeille koordinaateille funktiolla `text`. Muuttujaan Life Exp (joka sisältää välilyönnin) voi tarvittaessa viitata ympäröimällä sen nimen backtickeillä: `` `Life Exp` ``.

   (c) Paranna kuvaa siten, että osavaltioiden nimet on kirjoitettu eri väreillä sen mukaan, mihin alueeseen (Northeast, South, North Central, West) osavaltio kuuluu. Älä käytä väreinä mustaa tai valkoista. Osavaltioita vastaavat alueet löytyvät `datasets`:n muuttujasta `state.region`. Lisää vielä kuvan oikeaan yläkulmaan selitys sille, minkä värinen teksti viittaa mihinkin alueeseen. Vihje: tavoitteena on antaa argumentiksi `col` vektori, joka on kuin muuttuja `state.region`, mutta alueiden sijasta arvoina on eri värit (esim. ”red”). Toinen vihje: Funktio `legend`.