Harjoitustyö

Monimuuttujamenetelmät, kevät 2022

Maria Seppänen

Aineiston alkuperä

Harjoitustyössä käytetty aineisto kuvastaa Tampereella olevia vuokra-asuntoja. Aineisto löytyy Tampereen Yliopiston Moodlesivulta "Tilastomenetelmien työkalupakki". Sieltä löytyy "Tilastotieteen peruskurssien hajoitusaineistoja 2003-2018", josta lopulta työssä käytetty aineisto "Lumilaudat.xls".

Aineiston havaintoyksiköt ovat yksittäisiä lumilautoja ja muuttujat ovat näiden lumilautojen piirteitä. Aineiston muuttujista on näissä analyyseissä käytetty vain viittä muuttujaa "MERKKI", "MALLI", "KANTTI", "HINTA" ja "PITUUS". Muuttujat "MERKKI" ja "MALLI" ovat kategorisia muuttujia, jotka sisältävät kunkin lumilaudan merkin ja mallin. Koska "MALLI" on kategorinen muuttuja, jossa on 35 eri vaihtoehtoa, muutan sen numeeriseksi muuttujaksi, joka saa arvoja 1 ja 35 väliltä. Muuttujat "HINTA" ja "PITUUS" ovat jatkuvia numeerisia muuttujia, jotka kuvastavat kyseisen laudan hintaa ja pituutta. Muuttuja "KANTTI" on myös numeetinen jatkuva muuttuja. Havaintoja on yhteensä 80.

Ensin tarkastelen aineiston muuttujia sekä numeerisesti, että graafisesti. Sen jälkeen etsin pääkomponenttianalyysin avulla kuinka moneen komponenttiin aineiston muuttujat "MALLI", "KANTTI", "HINTA" ja "PITUUS" ja tarkastelen graafisesti, jakautuvatko nämä komponentit merkeittäin. Teen myös ryhmittelyanalyysin k-means-menetelmällä kolmelle jatkuvalle numeeriselle muuttujalle "KANTTI", "HINTA" ja "PITUUS". Viimeisenä luon näille samoille muuttujille mixture-mallin. Jotta r-koodin kirjoittaminen ja lukeminen olisi mukavampaa, muutan muuttujien nimet suurista kirjaimista pieniksi kirjaimiksi.

Aineiston numeerinen ja graafinen tarkastelu

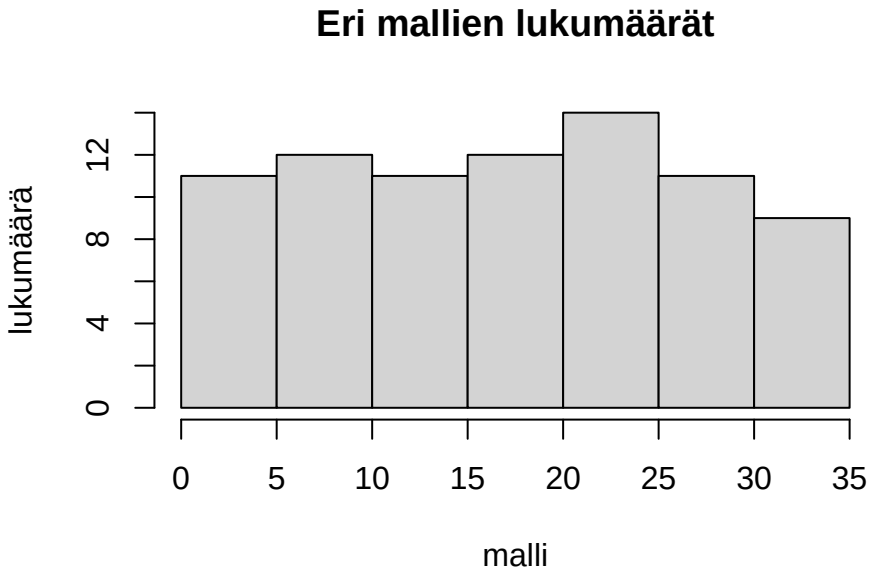
Tarkastellaan lumilautojen kantin, hinnan sekä pituuden perustunnuslukuja. Alla on taulukko, josta löytyvät muuttujien keskiarvot, mediaanit, keskihajonnat, sekä minimi- ja maksimiarvot. Keskiarvot vaihtelevat 116.3025 (kantti) ja 486.5500 (hinta) välillä. Keskiarvot ja mediaanit ovat hyvin lähellä toisiaan, mikä viittaisi siihen, että havainnot ovat suhteellisen tasaisesti jakautuneet. Jos nämä eroaisivat toisistaan, olisi datassa oltava taipumusta johonkin suuntaan tai jotain poikkeuksellisia havaintoja. Minimi- ja maksimiarvot eroavat toisistaan myös, eli muuttujat ovat jakautuneet hyvin erilaisille asteikoille. Tämä tulee myöhemmin ottaa huomioon pääkomponenttianalyysiä tehdessä.

```
##        keskiarvo mediaani keskihajonta min   max
## kantti  116.3025   117.65     11.04937  79 136.5
## hinta   486.5500   490.00    145.58254 250 750.0
## pituus  151.1625   154.00     12.05729 105 173.0
```

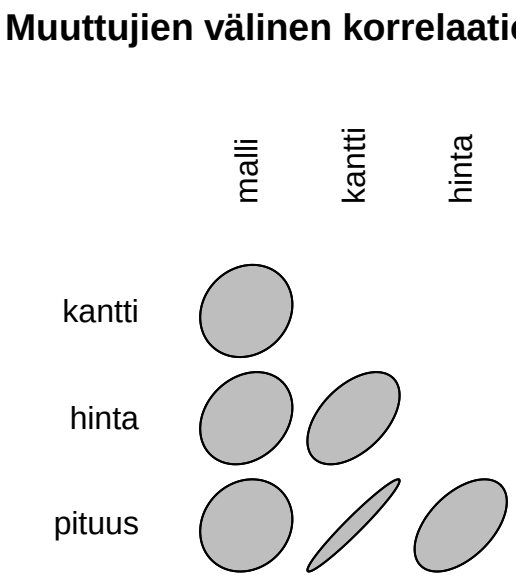
Tarkastellaan kategoristen muuttujien jakautumista. Merkit ovat jakautuneet tasaisesti, jokaista merkkiä on 20 havaintoa. Malleja on hyvin monta ja ovat jakautuneet hyvin tasaisesti, jokaista mallia ollen 1-3. Alla olevasta histogrammista, jossa mallit on jaettu viiden ryhmiin, voi myös visuaalisesti hahmottaa suhteellisen tasaista jakautumista.

```
## $merkki
##
## Burton  Forum  Lamar   Ride
##     20     20     20     20
```

```
##
## $malli
##
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
##  3  2  2  2  2  2  3  3  2  2  2  2  2  2  3  4  2  2  2  2  5  2  3  2  2  2
## 27 28 29 30 31 32 33 34 35
##  3  2  2  2  2  3  1  1  2
```



Tarkastellaan myös muuttujen välisiä korrelaatioita graafisesti. Kaikilla muuttujilla on keskenään pieni positiivinen korrelaatio, mutta muuttujien kantti ja hinta välillä on hyvin vahva positiivinen korrelaatio. Tämä viittaa siihen, että pääkomponenttianalyysi voisi olla hyödyllinen.



Pääkomponenttianalyysi

Pääkomponenttianalyysi on apuna aineistoissa, joissa muuttujia on hyvin paljon ja/tai ne korreloivat keskenään. Vaikka juuri tässä aineistossa kyse ei välttämättä muuttujien määrästä, tämän menetelmä on hyödyllinen muuttujien korrelaation vuoksi. Tässä analyysimenetelmässä aineiston havaituista muuttujista luodaan ns. pääkomponentit, joita voi kutsua myös latenteiksi muuttujiksi. Näillä pääkomponenteilla pyritään selittämään havaittujien muuttujien korrelaatio pienemmällä uusilla, korreloimattomilla pääkomponenteilla. Pääkomponentit voidaan määritellä niin, että j:s pääkomponentti on

$$y_j = a_{j1}x_1 + a_{j2}x_2 + ... + a_{jp}x_p = \mathbf{a_j'x},$$

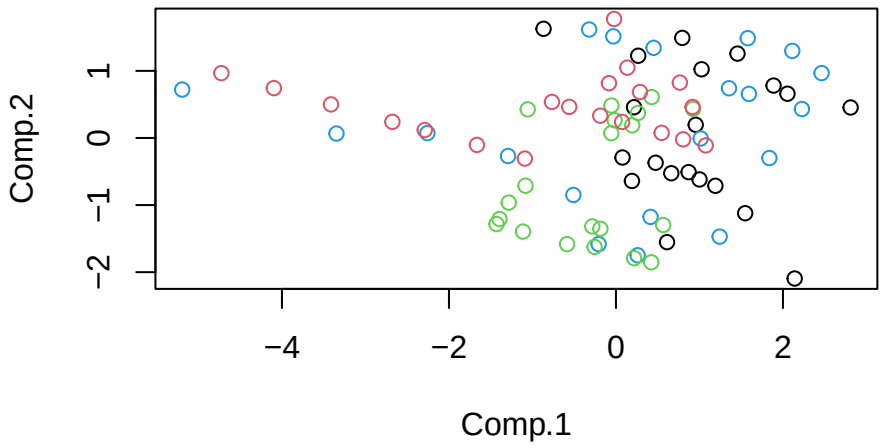
joka maksimoi y_j:n varianssin ehdoilla $\mathbf{a_j'a_j} = 1$ ja $\mathbf{a_j'a_i} = 0 (i < j)$.

Itse analyysissä on otettava huomioon muuttujien eri vaihteluvälit, joten käytän scale()-funktiota muuttujien standardoimiseen. Analyysiin käytän muuttujia “malli”, “kantti”, “hinta” ja “pituus”. Pääkomponenttialanyysin tulos alla.

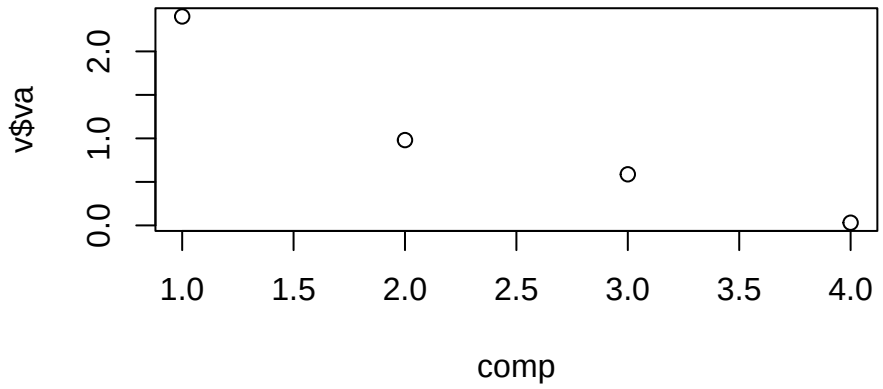
```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3      Comp.4
## Standard deviation     1.5398867 0.9839538 0.7615119 0.175166958
## Proportion of Variance 0.6003167 0.2451051 0.1468102 0.007767965
## Cumulative Proportion  0.6003167 0.8454218 0.9922320 1.000000000
##
## Loadings:
##        Comp.1 Comp.2 Comp.3 Comp.4
## malli   0.194  0.936  0.292
## kantti  0.607 -0.213  0.305 -0.702
## hinta   0.470  0.173 -0.865
## pituus  0.610 -0.218  0.270  0.712
```

Neljästä komponentista jo kahden ensimmäisen kumulatiivinen selitysosuus on 0.8454, eli noin 85 prosenttia.

Kolmella pääkomponentilla päästään jo yli 99 prosentin lukemiin. Käytänössä 70-90 prosentin selitysosuus on hyvä, joten tässä mallissa sopiva pääkomponenttien määrä olisi kaksi. Kun tarkastelee erillisiä latauksia, voi huomata, että ensimmäinen pääkomponentti sisältää enimmäkseen muuttujien "kantti" ja "pituus" vaihtelun, toinen pääkomponentti muuttujan "malli" ja kolmas pääkomponentti muuttujan "hinta". Neljännen pääkomponentin lataukset ovat hyvin heikkoja. Alla visuaalinen mallinnus kahdesta ensimmäisestä pääkomponentista.



Pääkomponenttien selitysosuudet voi myös mallintaa graafisesti. Alla kuva, josta voi myös huomata, että kaksi pääkomponenttia olisi tässä tilanteessa tarpeeksi. Kuvan alta löytyy myös neljän pääkomponentin korrelaatio alkuperäisten muuttujien kanssa. Suurin korrelaatio löytyy ensimmäisestä kahdesta pääkomponentista, kuten oli odotettavissakin.

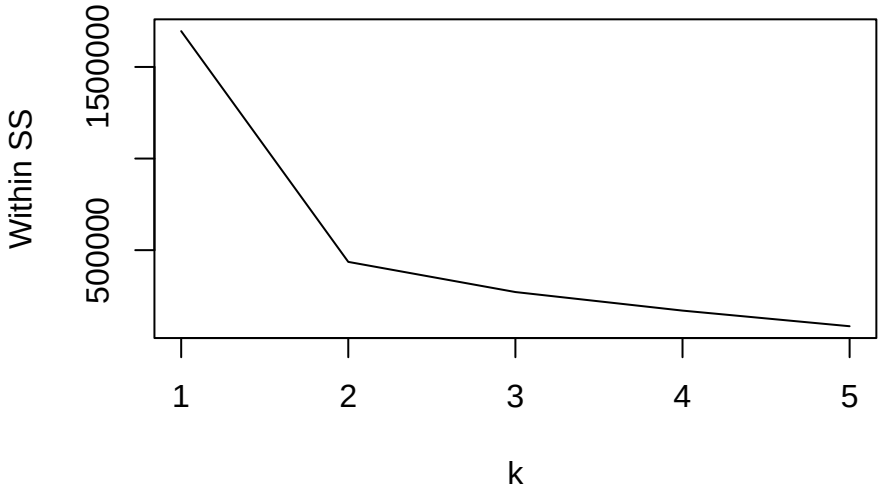



```
##           Comp.1     Comp.2     Comp.3       Comp.4
## malli  0.3000775  0.9272323  0.2240327  0.001782160
## kantti 0.9410269 -0.2110919  0.2336585 -0.123742852
## hinta  0.7287874  0.1713521 -0.6629427 -0.003784158
```

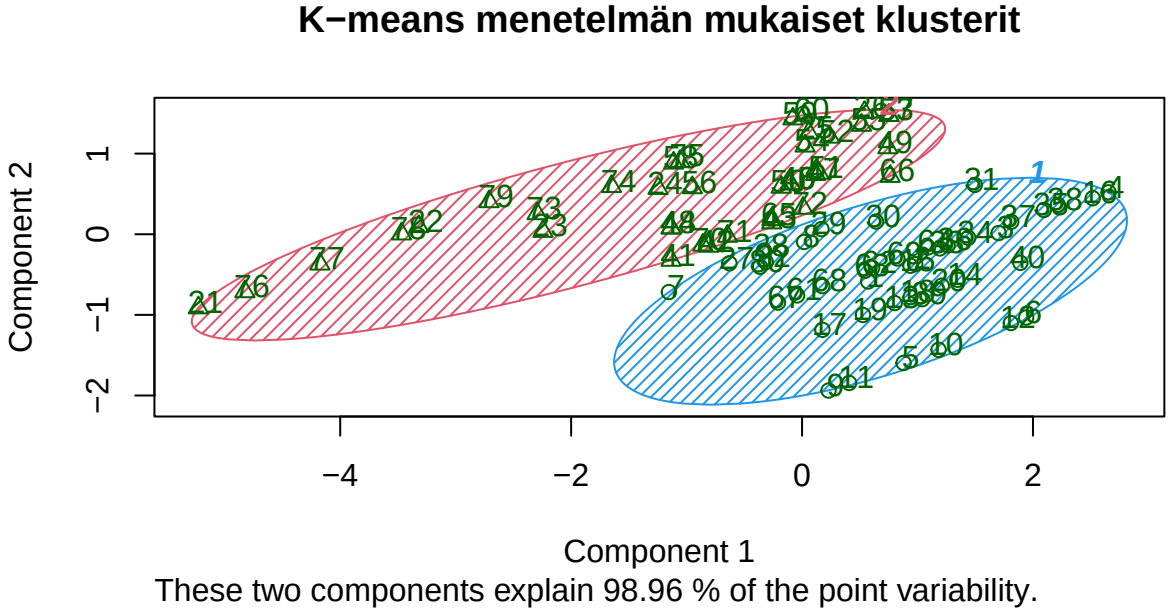
```
## pituus 0.9458106 -0.2161930  0.2072700  0.125467415
```

Ryhmittelyanalyysi

Ryhmittelyanalyysin tarkoitus on ryhmitellä aineisto sopiviin klustereihin niin, että yhden klusterin sisältämät havainnot ovat jollain tapaa samankaltaisia, mutta havainnot eri klustereissa eroavat toisistaan. Käytän ryhmittelyanalyysiin tällä kertaa kmeans-menetelmää. Analyysiin käytän muuttujia “hinta”, “kantti” ja “pituus”. Tarkastellaan tällä menetelmällä luotua kuvaajaa. Suurin hyppy tapahtuu yhden ja kahden välillä, jonka jälkeen erot ovat hyvin pieniä. Siis kaksi klusteria olisi tämän perusteella sopivin ratkaisu.

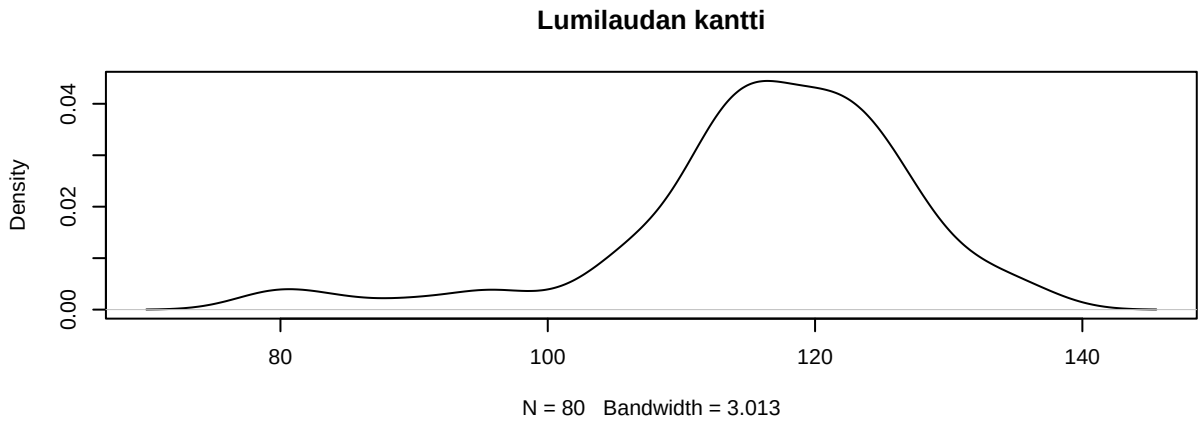


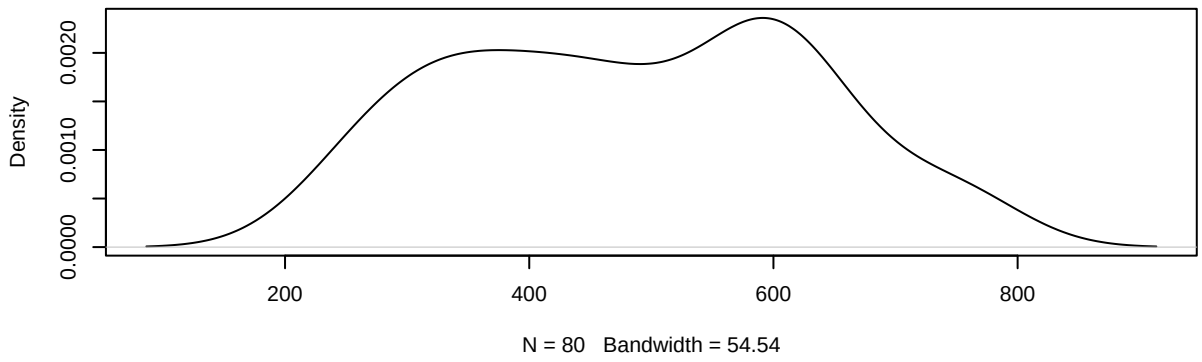
Nämä kaksi klusteria voi myös mallintaa graafisesti, kuten on alla. Siinä nähdään, että klusterit ovat täysin erillisiä, eli ei ole ryhmien välisiä päällekkäisyyksiä, mikä on toivottua.

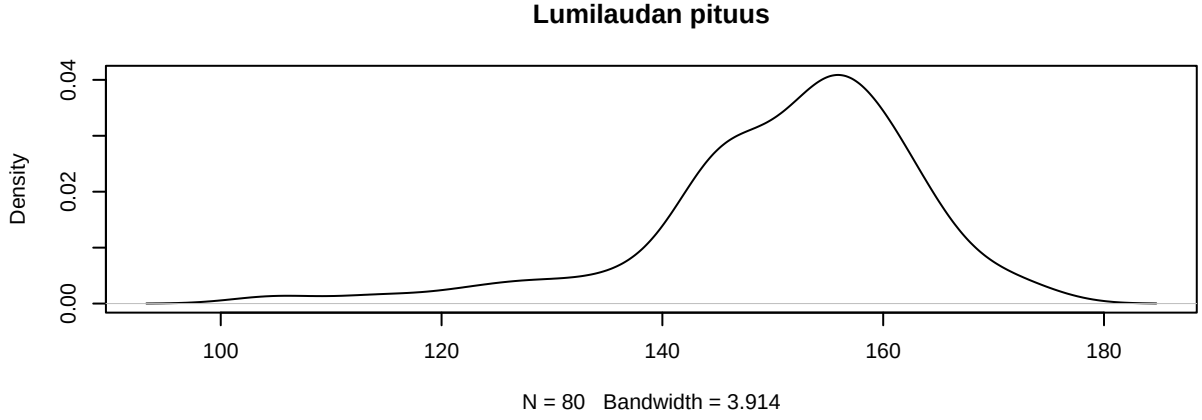


Mixture-malli

Mixture-malli liittyy läheisesti klusterianalyysiin, mutta edellisestä poiketen luodaan tilastollinen malli. Mixture-mallia tehdessä oletetaan, että aineisto on jakautunut ryhmiin niin, että jokaisella on oma tiheysfunktionsa. Tarkastellaan siis muuttujien "kantti", "hinta" sekä "pituus" tiheysfunktiota. Näiden muuttujien tiheysfunktiot sisältävät monta huippua ja epätasaisuuksia, jotka viittaavat mahdolliseen moniuloitteiseen normaalijakaumaan. Mixture-mallit sopivat tällaiseen aineistoon mainiosti.







Luodaan siis näille kolmelle muuttujalle mixture-malli. Alta voi nähdä, että mallissa aineisto on jaettu kolmeen komponenttiin. BIC-informaatiokriteerin perusteella kovarianssirakenteen oletetaan olevan $\Sigma_k = c\mathbf{T}_k\mathbf{\Lambda}_k\mathbf{T}_k'$.

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust EVV (ellipsoidal, equal volume) model with 3 components:
##
##  log-likelihood  n df       BIC       ICL
##       -933.2402 80 27 -1984.795 -1987.963
##
## Clustering table:
##  1  2  3
## 46 26  8
##
## Mixing probabilities:
##         1         2         3
## 0.5783570 0.3172719 0.1043711
##
## Means:
##            [,1]     [,2]      [,3]
## kantti 120.0274 116.8923  93.86883
## hinta  590.5390 370.1931 264.01758
## pituus 154.9503 152.5156 126.05993
##
## Variances:
## [,,1]
##           kantti     hinta    pituus
## kantti  45.85844  117.0578  50.01811
## hinta  117.05779 7127.5688 168.18480
## pituus  50.01811  168.1848  56.32352
## [,,2]
##            kantti      hinta   pituus
## kantti   57.34856 -118.49027 55.81495
## hinta  -118.49027 2585.24466 32.68973
## pituus   55.81495   32.68973 67.24810
## [,,3]
##          kantti    hinta   pituus
## kantti 157.8013 248.7171 138.9746
## hinta  248.7171 542.8093 175.8559
## pituus 138.9746 175.8559 154.8734
```

Alla taulukko klusterien jakautumisesta merkeittäin. Ne eivät jakaudu tarkalleen merkeittäin, mutta esimerkiksi “Burton” on täysin ensimmäisessä klusterissa ja “Lamar” täysin toisessa. “Ride” on enimmäkseen ensimmäisessä klusterissa ja “Forum” on kaikkein eniten jakautunut eri klustereihin.

```
##    merkki
##     Burton Forum Lamar Ride
##   1     20    12     0   14
##   2      0     2    20    4
##   3      0     6     0    2
```

Alla myös kaavio mallin klusterien jakautumisesta. Tekemäni mixture-malli sopii huonoiten muuttujien "malli" ja "kantti" kuvastamiseen, sillä niissä klusterit ovat osittain myös täysin päällekäin, jolloin ryhmien välillä ei ole eroa ja ne voisi todennäköisesti jakaa vain kahteen klusteriin. Muuten klusterit ovat erillään ja näyttävät toimivan aineistoon suhteellisen hyvin.