TILM 3558 Harjoitustyö, Tehtävä 3
Lasse Rintakumpu, 63555

Monimuuttujamenetelmät

Tarkastellaan pankin asiakasaineistoa (`pankkiotos.sav`). Havaintoaineisto sisältää n_{kaikki} = 2453 tilastoyksikköä, joista tähän tarkasteluun on valittu satunnaisesti n = 1000 tilastoyksikköä. [1] Muodostetaan datan 41 luokitellusta muuttujasta (`autom_lainan_perinta_luok–kulutusluotot1_luok`) pääkomponentit, joiden perusteella muodostetaan klusterianalyysillä datan tilastoyksiköistä asiakasryhmiä.

1. Pääkomponenttianalyysin edellytysten tarkastelu

Lähdetään liikkeelle pääkomponenttianalyysin edellytysten tarkastelusta. Testataan ensin muuttujien välisiä korrelaatioita tasoilla 0.05 ja 0.1. Tasolla 0.05 merkitseviksi saadaan 41*40=1640 korrelaatiosta 816 eli likimain 50%. Tasolla 0.1 merkitseviksi saadaan 914 eli noin 56%. Suurin osa korrelaatioista on tilastollisesti merkitseviä, mutta parannetaan datan sopivuutta pääkomponenttianalyysiin pudottamalla muuttujat, joiden korrelaatiot kaikkien muiden muuttujien kanssa ovat <0.3. Päädytään pudottamaan seuraavat kahdeksan muuttujaa:

```
asuntolaina_b_kpl_luok
vakuutus_b_luok
vakuutus_c_luok
kayttotili_vel_luok
asuntolaina_d_kpl_luok
asuntolaina_e_kpl_luok
toimeksianto_a_kpl_luok
toimeksianto_b_kpl_luok.
```

Pyritään seuraavaksi testaamaan Bartlettin sfäärisyystestillä muuttujien välisiä korrelaatioita. Testi on herkkä poikkeamille normaalijakaumasta, jota datan luokitellut muuttujat tuskin noudattavat. Testataan kuitenkin satunnaisesti muutaman muuttujan normaalisuutta:

```
shapiro.test(pankkiotos_pca_edit[,sample(1:33,1)])
# W = 0.6501, p-value < 2.2e-16
shapiro.test(pankkiotos_pca_edit[,sample(1:33,1)])
# W = 0.309, p-value < 2.2e-16
shapiro.test(pankkiotos_pca_edit[,sample(1:33,1)])
# W = 0.6581, p-value < 2.2e-16.
```

[1] Tarkasteltavat tilastoyksiköt on poimittu SPSS-ohjelmalla. Itse analyysissa on käytetty R-ohjelmistoa. Poimintaan käytetty SPSS-syntaksi sekä analyysiin käytetty R-koodi tehtävän liitteenä.

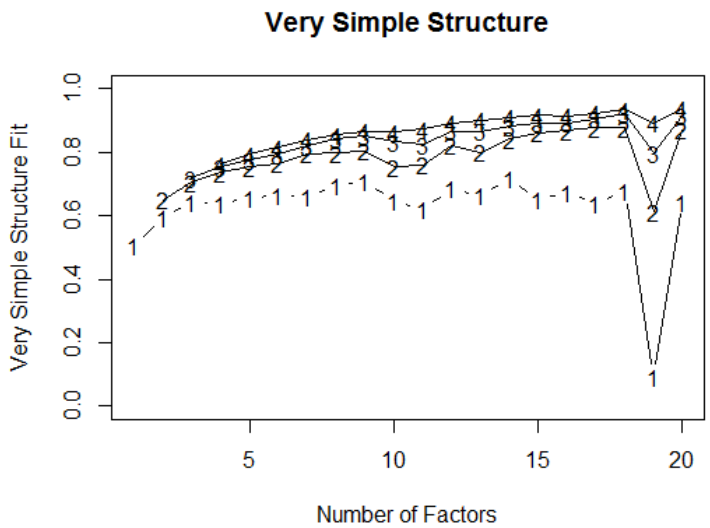
Mikään testatuista muuttujista ei noudata normaalijakaumaa. Jätetään Bartlettin sfäärisyystesti tekemättä. Luodaan datasta vielä anti image -korrelaatiomatriisi. Matriisissa kaikki osittaiskorrelaatiot ovat pienempiä kuin lävistäjän Measure of Sampling Adequacy -arvot. Alla tulostus kuuden ensimmäisen muuttujan (`autom_lainan_perinta_luok`, `lainojen_lukumaara_luok`, `asuntoluotot1_luok`, `automaattinostoja_luok`, `vakuutus_a_luok` ja `asuntolaina_a_kpl_luok`) osalta:

```
      [,1]       [,2]        [,3]       [,4]       [,5]        [,6]
[1,]  0.8658202 -0.3465899 -0.18074289  0.0321246 -0.0036200 -0.03898437
[2,] -0.3465899  0.8489369 -0.02921377  0.0255790 -0.0359194 -0.06068759
[3,] -0.1807429 -0.0292138  0.70529886 -0.0113844  0.0700977 -0.64620256
[4,]  0.0321246  0.0255790 -0.01138445  0.8908980 -0.0524762  0.01414464
[5,] -0.0036200 -0.0359194  0.07009772 -0.0524762  0.6574366 -0.14474710
[6,] -0.0389844 -0.0606876 -0.64620256  0.0141446 -0.1447471  0.58503483.
```

Testataan vielä muuttujien kokojen sopivuutta pääkomponenttianalyysiin koko datan osalta Kaiser-Meyer-Olkin -testillä. Saadaan KMO-suhdeluvuksi 0.72>0.6. Otoskokoa voidaan pitää riittävän suurena pääkomponenttianalyysiin. Todetaan pääkomponenttianalyysin edellytysten olevan voimassa ja siirrytään itse analyysiin.

2. Pääkomponenttianalyysi

Pyritään määrittämään Promax-rotatoitujen pääkomponenttien lukumäärä Very Simple Structure -menetelmällä. Saadaan tulostukseksi:



Pyritään valitsemaan pääkomponenttien lukumäärä niin, että sopivuus on vähintään 80%. Kuvaajasta huomataan, että kompleksisuuksilla 1 ja 2 (joissa huomioidaan ainoastaan 1 tai 2

komponenttiin vahvimmin latautunutta muuttujaa) sopivuus saavutetaan vasta yli kymmenellä pääkomponentilla. Velicer MAP -kriteeri ehdottaa neljää pääkomponenttia, joten yli kymmenen komponenttia on todennäköisesti liikaa. Tarkastellaan suurempia kompleksisuuksia, joista kompleksisuudella 4 saavutetaan 80% sopivuus seitsemällä pääkomponentilla. Aloitetaan tämän perusteella pääkomponenttianalyysi seitsemällä pääkomponentilla. Saadaan tulokseksi seuraavat lataukset:

	PC1	PC4	PC2	PC3	PC7	PC5	PC6
autom_lainan_perinta_luok		0.629			0.475		
lainojen_lukumaara_luok		0.548			0.513		
asuntoluotot1_luok	-0.103	0.786			0.255		0.194
automaattinostoja_luok	0.877		-0.134				
vakuutus_a_luok	-0.140					0.101	0.656
asuntolaina_a_kpl_luok	-0.106	0.351		-0.138	0.273		0.456
korkeakork_kpl_luok		0.137	0.403		-0.196	0.222	
rahasto_a1_luok		-0.152		0.243		0.214	
pankkikorttilkm_luok	0.637	-0.145			-0.215	-0.129	0.261
luottokortteja_yhteens...					-0.188		0.583
maaraaikaistileja_luok	-0.136		0.632		-0.129	-0.109	
maksuautomaattitapahtu…	0.810	-0.163	-0.244	-0.100			-0.254
kayttotili_tal_luok			0.649				
asuntolaina_c_kpl_luok		0.667		0.149	0.122	-0.124	-0.145
osakkeet_euroa_1_luok						0.988	
eri_osakesarjoja_luok				0.112		0.933	
rahasto_b1_luok				0.675			
ottoja_luok	0.815	0.173					
pkorttimaksuja_luok	0.730	0.248			-0.186		
panoja_luok	0.674		0.287		0.180		
palveluja_kpl_luok	0.386	0.465	0.101		0.107		
rahastolajeja_luok				0.960			
lainarastit_luok		0.111			0.538		-0.139
saastotililla_luok	-0.131	0.110	0.767	-0.100	-0.178		
suoraveloituksia_luok	0.259		0.540			0.128	
netissa_maksut_luok	0.267	0.577	-0.103		-0.172		0.160
maksupalvelussa_maksut…		-0.149	0.668		0.456		
tiskilla_maksut_luok	0.105	-0.201	0.529		0.469		
tilinylityspaivat_luok	0.378		-0.225		0.165		0.101
kv_maksukortit_luok	0.154	0.504					-0.267
rahasto_c1_luok				0.786			
korttiluotot1_luok		0.199					0.607
kulutusluotot1_luok		0.221			0.662.		

Muuttujista `rahasto_a1_luok` ei ole latautunut selkeästi yhteenkään pääkomponenttiin. Poistetaan muuttuja analyysista.

Tarkastellaan seuraavaksi muuttujien kommunaliteetteja:

autom_lainan_perinta_luok	0.6664377	lainojen_lukumaara_luok	0.6362016
asuntoluotot1_luok	0.6821972	automaattinostoja_luok	0.7530206
vakuutus_a_luok	0.4015140	asuntolaina_a_kpl_luok	0.4365984
korkeakork_kpl_luok	0.3418248	pankkikorttilkm_luok	0.5176078
luottokortteja_yhteensa_luok	0.3368983	maaraaikaistileja_luok	0.4335168
maksuautomaattitapahtumia_luok	0.5793514	kayttotili_tal_luok	0.4101920
asuntolaina_c_kpl_luok	0.4496636	osakkeet_euroa_1_luok	0.9127898
eri_osakesarjoja_luok	0.9160190	rahasto_b1_luok	0.4540376
ottoja_luok	0.8046949	pkorttimaksuja_luok	0.7447466
panoja_luok	0.6699779	palveluja_kpl_luok	0.5821450
rahastolajeja_luok	0.9180276	lainarastit_luok	0.2966631
saastotililla_luok	0.6677582	suoraveloituksia_luok	0.4672776
netissa_maksut_luok	0.5806915	maksupalvelussa_maks...	0.6719322
tiskilla_maksut_luok	0.5685945	tilinylityspaivat_luok	0.2795146
kv_maksukortit_luok	0.3975283	rahasto_c1_luok	0.6425869
korttiluotot1_luok	0.4164324	kulutusluotot1_luok	0.4594335

Muuttujien kommunaliteetit ovat yleisesti riittävän korkeita. Ainoastaan muuttujilla `tilinylityspaivat_luok, lainarastit_luok` ja `luottokortteja_yhteensa_luok` havaitaan alhaiset kommunaliteetit (0.28, 0.30 ja 0.34). Pidetään muuttujat kuitenkin mukana analyysissa, koska ne kaikki ovat selkeästi latautuneet yhdelle pääkomponentille (PC1, PC7 ja PC6). Päädytään analyysilla seitsemään pääkomponenttiin, joille ovat selvimmin latautuneet seuraavat muuttujat:

PC1:

- Automaattinostot (`automaattinostoja_luok, 0.876`),
- ottojen lukumäärä (`ottoja_luok, 0.820`),
- maksuautomaattitapahtumien lukumäärä (`maksuautomaattitapahtum..., 0.801`),
- pankkikorttimaksujen lukumäärä (`pkorttimaksuja_luok, 0.743`),
- panojen lukumäärä (`panoja_luok, 0.678`),
- pankkikorttien lukumäärä (`pankkikorttilkm_luok, 0.638`),
- pankkipalvelujen lukumäärä (`palveluja_kpl_luok, 0.395`),
- tilinylityspäivien lukumäärä (`tilinylityspaivat_luok, 0.370`).

PC2:

- Säästötilin saldo (`saastotililla_luok, 0.850`),
- määräaikaistilien lukumäärä (`maaraaikaistileja_luok, 0.683`),
- käyttötilin saldo positiivinen (`kayttotili_tal_luok, 0.656`),
- Korkeakorkoisia tilejä (`korkeakork_kpl_luok, 0.477`),
- suoraveloitusten lukumäärä (`suoraveloituksia_luok, 0.453`),
- maksupalvelussa maksut (`maksupalvelussa_maksut_luok, 0.406`).

PC3:

- Erilaisten rahastolajien lukumäärä (rahastolajeja_luok, 0.978),
- c-tyypin rahastojen markkina-arvo, 1. mittaus (rahasto_c1_luok, 0.830),
- B-tyypin rahastojen markkina-arvo, 1. mittaus (rahasto_b1_luok, 0.683).

PC4:

- Asuntoluottoja euroa, 1. mittaus (asuntoluotot1_luok, 0.862),
- automaattisia lainan perintöjä lukumäärä (autom_lainan_perinta_luok, 0.747),
- kaikkien lainojen lukumäärä (lainojen_lukumaara_luok, 0.679),
- c-tyypin asuntolainojen lukumäärä (asuntolaina_c_kpl_luok, 0.668),
- netissä maksut lukumäärä (netissa_maksut_luok, 0.531),
- pankkipalvelujen lukumäärä (palveluja_kpl_luok, 0.472)
- a-tyypin asuntolainojen lukumäärä (asuntolaina_a_kpl_luok, 0.470),
- kansainvälisten maksukorttien lukumäärä (kv_maksukortit_luok, 0.434),
- kulutusluottoja yhteensä, 1. mittaus (kulutusluotot1_luok, 0.397).

PC5:

- Osakkeiden arvo euroa, 1. mittaus (osakkeet_euroa_1_luok, 0.979),
- erilaisten osakesarjojen lukumäärä (eri_osakesarjoja_luok, 0.926).

PC6:

- A-tyypin vakuutus (vakuutus_a_luok, 0.642),
- korttiluotot euroa, 1. mittaus (korttiluotot1_luok, 0.590),.
- luottokorttien lukumäärä (luottokortteja_yhteensa_luok, 0.572),
- a-tyypin asuntolainojen lukumäärä (asuntolaina_a_kpl_luok, 0.436).

PC7:

- Maksupalvelussa maksut lukumäärä (maksupalvelussa_maksut_luok, 0.745),
- pankin tiskillä maksut lukumäärä (tiskilla_maksut_luok, 0.725),
- kulutusluottoja yhteensä euroa, 1. mittaus (kulutusluotot1_luok, 0.399),
- lainarästit (lainarastit_luok, 0,314).

Lisäksi seuraavilla pääkomponenteilla on vahvoja negatiivisia latauksia:

PC2: Tilinylityspäivien lukumäärä (tilinylityspaivat_luok, -0.310).
PC7: Netissä maksut lukumäärä (netissa_maksut_luok, -0.359).

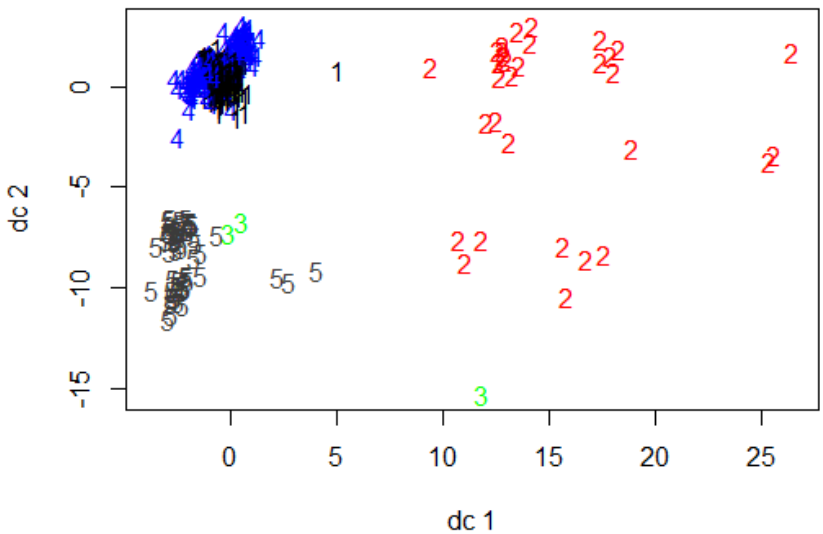
Pääkomponenttien suurimpia latauksia tarkastelemalla havaitaan, että analyysi seitsemällä pääkomponentilla tuottaa selkeät, suhteellisen luonnollisesti pankin eri liiketoiminta-alueita

tai palveluja vastaavat komponentit. Pääkomponenttiin yksi on latautunut erityisesti käyttötilejä, pääkomponenttiin kaksi säästötilejä, pääkomponenttiin kolme rahastoja, pääkomponenttiin neljä asunto- ja muita lainoja, pääkomponenttiin viisi osakkeita, pääkomponenttiin kuusi vakuutuksia ja pääkomponenttiin seitsemän konttoriasiointia koskevia muuttujia.

Lasketaan näiden seitsemän pääkomponentin pohjalta pääkomponenttipistemäärät asiakasdatalle. Tallennetaan pistemäärät omaan matriisiinsa `pca_pistemaarat`, jonka riveiltä löytyvät tilastoyksiköt ja sarakkeista pääkomponentit PC1–7 nimettynä uudelleen `a1_kayttotili`, `a2_saastotili`, `a3_rahasto`, `a4_laina`, `a5_osake`, `a6_vakuutus` ja `a7_tiski`.

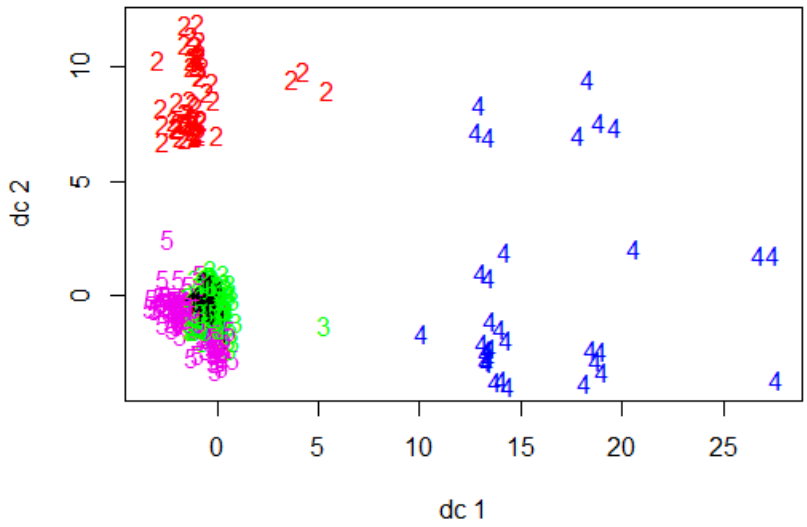
3. Klusterianalyysi

Siirrytään hyödyntämään pääkomponenttipistemäärämatriisia klusterianalyysissä, jonka tavoitteena on asiakassegmenttien muodostaminen. Pyritään etsimään sopiva klusterien määrä sekä keskukset hierarkkisin klusterointimenetelmin ja hyödyntämään näitä ei-hierarkkisen menetelmän alkuarvoina. Etsitään alkuarvo klustereiden määrälle malliperustaisella menetelmällä. Päädytään viiteen klusteriin, joiden keskukset etsitään keskimääräisen linkityksen avulla. Käytetään löydettyjä keskuksia k-means -klusteroinnin alkuarvoina, jolloin päädytään seuraavaan klusterointiin (esitettynä graafisesti):

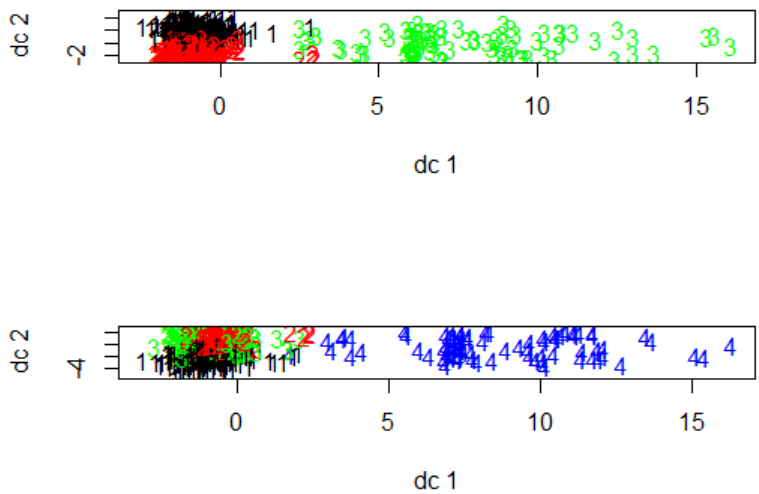


Kaaviosta havaitaan, että klusteriin numero kolme päätyy ainoastaan kolme havaintoa. Vaikka klusterien määrä laskettaisiin neljään, tuottaa k-means-klusterointi edelleen saman, kolmen

havainnon klusterin. Poistetaan kyseiset havainnot datasta ja jatketaan klusterointia viidellä klusterilla. Päädytään seuraavaan klusterointiin:



Kaaviosta huomataan k-means klusteroinnin muodostavan kaksi selkeää klusteria (klusterit #2 ja #4) sekä kolme toisistaan huonosti erottuvaa klusteria (#1, #3 ja #5). Erityisesti klusterien #3 ja #5 alkiot ovat kaaviossa selkeästi päällekkäisiä. Tarkastellaan vaihtoehtoina klustereiden vähentämistä kolmeen tai neljään:



Klustereiden vähentäminen ei näytä tuottavan selkeämpi klustereita. Palataan alkuperäiseen viiteen klusteriin ja tarkastellaan klusteriprofiileita:

```
   a1_kayttotili a2_saastotili a3_rahasto a4_laina a5_osake a6_vakuutus a7_tiski
1 -0.80         -0.45         -0.27      -0.19    -0.28    -0.33       -0.20
2  0.23          0.18          0.60      -0.04     3.41     0.00        0.09
3  0.75         -0.30          0.58      -0.08    -0.34     0.00        0.30
4  0.25          0.40          0.77       5.03     0.93    -0.26       -0.10
5  0.68          1.59         -0.69      -0.36    -0.08     0.92       -0.21.
```

Profiilien mukaan klustereita profiloivat eri pääkomponenttipistemuuttujat seuraavasti:

- Klusteri 1: `ei profiloivaa muuttujaa,`
- klusteri 2: `a5_osake,`
- klusteri 5: `a2_saastotili, a6_vakuutus,`
- klusteri 3: `a1_kayttotili, a7_tiski`
- klusteri 4: `a4_laina, a3_rahasto.`

Klusterien frekvenssit ovat puolestaan seuraavat:

```
       1   2   3   4   5
[1] 445  60 294  33 165.
```

Käytetään klusteriprofiileita asiakassegmentoinnin pohjana. Pienimmät klusterit #2 ja #4 profiloituvat selkeästi osake- ja (asunto)laina-asiakkaiksi. Myös käyttötiliasiakkaat (klusteri #3) erottuvat säästötiliasiakkaista (klusteri #5). Suurin klusteri on kuitenkin #1, jota ei profiloi selkeästi mikään muuttuja ja jossa kaikki keskiarvot ovat myös negatiivisia, joten klusterin voi ajatella edustavan passiivista asiakassegmenttiä. Tiliasiakkaat (erityisesti käyttötiliasiakkaat) ja passiiviset asiakkaat eivät kuitenkaan klustereina eroa suuresti toisistaan, mikä antaisi aiheen jatkoanalyysille.

Liitteet

Liitteenä tilastoyksikköjen valinnassa käytetty SPSS-syntaksi sekä analyysissa käytetty R-koodi.

Liite 1. SPSS-syntaksi

```
SET SEED=63555.
DATASET ACTIVATE DataSet1.
USE ALL.
```

```
do if $casenum=1.
compute #s_$_1=1000.
compute #s_$_2=2453.
end if.
do if #s_$_2 > 0.
compute filter_$=uniform(1)* #s_$_2 < #s_$_1.
compute #s_$_1=#s_$_1 - filter_$.
compute #s_$_2=#s_$_2 - 1.
else.
compute filter_$=0.
end if.
VARIABLE LABELS filter_$ '1000 from the first 2453 cases (SAMPLE)'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
```

Liite 2. R-koodi

```
########################################
# TILM3558 Harjoitustyö, Osa 3, R-koodi #
# Lasse Rintakumpu, 63555               #
# 31.8.2015                             #
########################################

# Asetetaan työhakemisto
wd <- "D:/Dropbox/Edu/Statistics/TILM 3558 Harjoitustyö" #Hipatlaptop
setwd(wd)

# Funktio kirjastojen asentamiselle / lataamiselle
lataa_kirjasto <- function(kirjasto) {
  if(kirjasto %in% rownames(installed.packages()) == FALSE)
  {install.packages(kirjasto)}
  library(kirjasto, character.only = TRUE)
}

# Ladataan/asennetaan käytetyt kirjastot
lapply(c("psych", "corpcor", "corrplot", "nFactors", "GPArotation",
"mclust", "modeltools", "fpc", "cluster"), lataa_kirjasto)

# Funktio poikkeavien havaintojen poistamiseen klusterista
poikkeavatHavainnot <- function(data, klusteri_kmeans, klusteri_numero,
havaintojenMaara=1, poisto=FALSE) {
  keskustat <-
klusteri_kmeans$centers[klusteri_kmeans$cluster[klusteri_kmeans$cluster==kl
usteri_numero], ]
  etaisyydet <-
sqrt(rowSums((data[klusteri_kmeans$cluster==klusteri_numero] -
keskustat)^2))
```

```
  ph<-order(etaisyydet, decreasing=T)[1:havaintojenMaara]
  if(poisto==FALSE) {
    # Palauta poistettavat IDt
    return(ph)
  } else {
    # Palauta data, josta poistetu arvot
    return(data[-ph,])
  }
}

# Ladataan havaintoaineisto ja valitaan tiedot satunnaisesti valituilta
1000 riviltä
pankkiotos <-
read.csv("https://raw.githubusercontent.com/rintakumpu/tilm3558/master/pank
kiotos_filtered.csv", sep=",", dec=".", header=TRUE, row.names=NULL,
fileEncoding = "UTF-8-BOM")
pankkiotos <- as.data.frame(pankkiotos[pankkiotos$filter==1,])

# Tallennetaan pääkomponenttianalyysissa käytettävät muutujat omaan
# matriisiinsa
pankkiotos_pca <- pankkiotos[,69:109]

########################################################
# 1. Pääkomponenttianalyysin edellytysten tarkastelu #
########################################################

# Tarkastellaan muuttujien välisten korrelaatioiden merkitsevyyttä
# Funktio korrelaatioiden läpikäyntiin
cor.mtest <- function(mat, conf.level = 0.95) {
  mat <- as.matrix(mat)
  n <- ncol(mat)
  p.mat <- lowCI.mat <- uppCI.mat <- sig.mat <- matrix(NA, n, n)
  diag(p.mat) <- 0
  diag(lowCI.mat) <- diag(uppCI.mat) <- 1
  for (i in 1:(n - 1)) {
    for (j in (i + 1):n) {
      tmp <- cor.test(mat[, i], mat[, j], conf.level = conf.level)
      p.mat[i, j] <- p.mat[j, i] <- tmp$p.value
      lowCI.mat[i, j] <- lowCI.2mat[j, i] <- tmp$conf.int[1]
      uppCI.mat[i, j] <- uppCI.mat[j, i] <- tmp$conf.int[2]

      # Kuuluuko nolla luottamusvälille?
      if(tmp$conf.int[1]<=0 & tmp$conf.int[2]>=0) {
        sig.mat[i,j] <- sig.mat[j,i] <- FALSE
      } else { sig.mat[i,j] <- sig.mat[j,i] <- TRUE }
    }
  }

  # Palauttaa listan, jossa merkitsevät korrelaatiot TRUE
```

```
  return(sig.mat)
}

korrelaatiot_p005<-cor.mtest(pankkiotos_pca, 0.95)
sum(unlist(korrelaatiot_p005), na.rm=TRUE) # 816 / 1640 (41*40, eli
korrelaatiot muuttujien itsensä kanssa korrelaatiot poistettu)
# => 50% merkitseviä

korrelaatiot_p01<-cor.mtest(pankkiotos_pca, 0.90)
sum(unlist(korrelaatiot_p01), na.rm=TRUE) # 914 / 1640 => 56% merkitseviä

# Muuttujat juuri ja juuri sopivia pääkomponenttianalyysiin
# Tarkastellaan vielä korrelaatiota

korrelaatiot<-round(cor(pankkiotos_pca),2)
# Ainoastaan hyvin heikkoja korrelaatiot (kaikki korrelaatiot <0.2)
# sisältävät muuttujat
korrelaatiot[7,] # asuntolaina_b_kpl_luok
korrelaatiot[8,] # vakuutus_b_luok
korrelaatiot[9,] # vakuutus_c_luok
korrelaatiot[17,] # kayttotili_vel_luok
korrelaatiot[25,] # asuntolaina_d_kpl_luok
korrelaatiot[30,] # asuntolaina_e_kpl_luok
korrelaatiot[36,] # toimeksianto_a_kpl_luok
korrelaatiot[37,] # toimeksianto_b_kpl_luok
pudotettavat <- c("asuntolaina_b_kpl_luok", "vakuutus_b_luok",
"vakuutus_c_luok", "kayttotili_vel_luok",  "asuntolaina_d_kpl_luok",
"asuntolaina_e_kpl_luok", "toimeksianto_a_kpl_luok",
"toimeksianto_b_kpl_luok")

# Pudotetaan nämä
pankkiotos_pca_edit <- pankkiotos_pca[,!(names(pankkiotos_pca) %in%
pudotettavat)]

# Luodaan anti image -korrelaatiomatriisi
antiImage <- cor2pcor(cor(pankkiotos_pca_edit))*-1 #
Osittaiskorrelaatiomatriisi * -1
# Korvataan lävistäjä MSA-arvoilla
diag(antiImage) <- as.vector(KMO(pankkiotos_pca_edit)$MSAi) # MSA-arvot
vektorina

# Anti image -matriisi yhdeksän ensimmäisen muuttujan osalta
#             [,1]       [,2]        [,3]       [,4]       [,5]        [,6]
[,7]        [,8]       [,9]
# [1,]  0.8658202 -0.3465899 -0.18074289  0.0321246 -0.0036200 -0.03898437
0.0184956 -0.0069339  0.1221809
# [2,] -0.3465899  0.8489369 -0.02921377  0.0255790 -0.0359194 -0.06068759
0.0919892  0.0318555 -0.1080796
# [3,] -0.1807429 -0.0292138  0.70529886 -0.0113844  0.0700977 -0.64620256
```

```
0.0145237 -0.0121349  0.0417501
# [4,]  0.0321246  0.0255790 -0.01138445  0.8908980 -0.0524762  0.01414464
-0.0032462 -0.0014667 -0.0701634
# [5,] -0.0036200 -0.0359194  0.07009772 -0.0524762  0.6574366 -0.14474710
-0.0197016 -0.0233843  0.0109619
# [6,] -0.0389844 -0.0606876 -0.64620256  0.0141446 -0.1447471  0.58503483
-0.0166150  0.0129481 -0.0275446
# [7,]  0.0184956  0.0919892  0.01452367 -0.0032462 -0.0197016 -0.01661504
0.6347396  0.0729614  0.0267469
# [8,] -0.0069339  0.0318555 -0.01213485 -0.0014667 -0.0233843  0.01294808
0.0729614  0.2853626 -0.0142156
# [9,]  0.1221809 -0.1080796  0.04175006 -0.0701634  0.0109619 -0.02754464
0.0267469 -0.0142156  0.6234076

# Matriisin luvut suurempia diagonaalilla kuin sen ulkopuolella,
# myös muuttujilla 10-33.

# Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy
KMO(pankkiotos_pca_edit)[[1]] # 0.718899 > 0.6
# KMO OK

# Bartlettin sfäärisyystesti on herkkä poikkeamille normaalijakaumista,
# normaalisuus datan luokitelluilla muuttujilla tuskin pätee,
# testataan kuitenkin satunnaisesti pari muuttujaa:

shapiro.test(pankkiotos_pca_edit[,sample(1:33,1)]) # W = 0.6501, p-value <
2.2e-16
shapiro.test(pankkiotos_pca_edit[,sample(1:33,1)]) # W = 0.309, p-value <
2.2e-16
shapiro.test(pankkiotos_pca_edit[,sample(1:33,1)]) # W = 0.6581, p-value <
2.2e-16

# Normaalisuusoletus ei päde, unohdetaan Bartlettin testi

##############################
# 2. Pääkomponenttianalyysi #
##############################

# Pyritään määrittelemään pääkomponenttien lukumäärä automaattisesti
komponenttien_maara <- VSS(pankkiotos_pca_edit, n=20, rotate="promax",
fm="pc", diagonal=FALSE, plot=TRUE)
pdf('pankkiotos_vss.pdf')
dev.off()

# Kaavion perusteella n. 80% sopivuus saavutetaan n. 8 pääkomponentille
# The Velicer MAP achieves a minimum of 0.02  with  4  factors

#     vss1 vss2   map dof chisq prob sqresid  fit
# 1  0.506 0.00 0.021   0    NA   NA    37.9 0.51
```

```
# 2  0.596 0.65 0.020   0    NA   NA    26.9 0.65
# 3  0.641 0.70 0.020   0    NA   NA    21.6 0.72
# 4  0.639 0.74 0.019   0    NA   NA    18.2 0.76
# 5  0.655 0.75 0.020   0    NA   NA    15.7 0.80
# 6  0.665 0.77 0.021   0    NA   NA    13.5 0.82
# 7  0.660 0.79 0.019   0    NA   NA    11.5 0.85
# 8  0.696 0.80 0.020   0    NA   NA    10.0 0.87
# 9  0.709 0.80 0.021   0    NA   NA     8.7 0.89
# 10 0.646 0.75 0.023   0    NA   NA     7.6 0.90

# => Lähdetään liikkeelle 7 pääkomponentista (fit 85%)

malli_pca <-principal(pankkiotos_pca_edit, nfactors=7, rotate="promax")
malli_pca$loadings

# Pääkomponenteille latautuneet muuttujat:
#                                 PC1    PC4    PC2    PC3    PC7    PC5
PC6
# autom_lainan_perinta_luok             0.629                 0.475
# lainojen_lukumaara_luok               0.548                 0.513
# asuntoluotot1_luok             -0.103  0.786                 0.255
0.194
# automaattinostoja_luok          0.877        -0.134
# vakuutus_a_luok                -0.140                               0.101
0.656
# asuntolaina_a_kpl_luok         -0.106  0.351        -0.138  0.273
0.456
# korkeakork_kpl_luok                    0.137  0.403        -0.196  0.222
# rahasto_a1_luok                       -0.152         0.243         0.214
! FLAG
# pankkikorttilkm_luok            0.637 -0.145                -0.215 -0.129
0.261
# luottokortteja_yhteensa_luok                                -0.188
0.583
# maaraaikaistileja_luok         -0.136         0.632         -0.129 -0.109
# maksuautomaattitapahtumia_luok  0.810 -0.163 -0.244 -0.100
-0.254
# kayttotili_tal_luok                           0.649
# asuntolaina_c_kpl_luok                 0.667         0.149  0.122 -0.124
-0.145
# osakkeet_euroa_1_luok                                              0.988
# eri_osakesarjoja_luok                                 0.112         0.933
# rahasto_b1_luok                                       0.675
# ottoja_luok                     0.815  0.173
# pkorttimaksuja_luok             0.730  0.248                 -0.186
# panoja_luok                     0.674         0.287          0.180
# palveluja_kpl_luok              0.386  0.465  0.101          0.107
# rahastolajeja_luok                                    0.960
# lainarastit_luok                       0.111                 0.538
```

```
-0.139
# saastotililla_luok             -0.131  0.110  0.767 -0.100 -0.178
# suoraveloituksia_luok           0.259         0.540                0.128
# netissa_maksut_luok             0.267  0.577 -0.103        -0.172
0.160
# maksupalvelussa_maksut_luok           -0.149  0.668         0.456
# tiskilla_maksut_luok            0.105 -0.201  0.529         0.469
# tilinylityspaivat_luok          0.378        -0.225         0.165
0.101
# kv_maksukortit_luok             0.154  0.504
-0.267
# rahasto_c1_luok                                    0.786
# korttiluotot1_luok                    0.199
0.607
# kulutusluotot1_luok                   0.221                0.662

# rahasto_a1_luok ei ole latautunut selkeästi yhteenkään pääkomponenttiin
# poistetaan ko. muuttuja
pankkiotos_pca_edit2 <- pankkiotos_pca_edit[,!(names(pankkiotos_pca_edit)
%in% c("rahasto_a1_luok"))]

malli_pca2 <-principal(pankkiotos_pca_edit2, nfactors=7, rotate="promax")
malli_pca2$loadings

# Nyt kaikki muuttujat ovat latautuneet vähintään yhteen pääkomponenttiin
# (lataukset >0.3)

# Tarkastellaan vielä kommunaliteetteja
as.matrix(malli_pca2$communality)

# autom_lainan_perinta_luok      0.6664377
# lainojen_lukumaara_luok        0.6362016
# asuntoluotot1_luok             0.6821972
# automaattinostoja_luok         0.7530206
# vakuutus_a_luok                0.4015140
# asuntolaina_a_kpl_luok         0.4365984
# korkeakork_kpl_luok            0.3418248
# pankkikorttilkm_luok           0.5176078
# luottokortteja_yhteensa_luok   0.3368983
# maaraaikaistileja_luok         0.4335168
# maksuautomaattitapahtumia_luok 0.5793514
# kayttotili_tal_luok            0.4101920
# asuntolaina_c_kpl_luok         0.4496636
# osakkeet_euroa_1_luok          0.9127898
# eri_osakesarjoja_luok          0.9160190
# rahasto_b1_luok                0.4540376
# ottoja_luok                    0.8046949
# pkorttimaksuja_luok            0.7447466
# panoja_luok                    0.6699779
```

```
# palveluja_kpl_luok             0.5821450
# rahastolajeja_luok             0.9180276
# lainarastit_luok               0.2966631
# saastotililla_luok             0.6677582
# suoraveloituksia_luok          0.4672776
# netissa_maksut_luok            0.5806915
# maksupalvelussa_maksut_luok    0.6719322
# tiskilla_maksut_luok           0.5685945
# tilinylityspaivat_luok         0.2795146
# kv_maksukortit_luok            0.3975283
# rahasto_c1_luok                0.6425869
# korttiluotot1_luok             0.4164324
# kulutusluotot1_luok            0.4594335

# Alhaiset kommunaliteetit havaitaan seuraavissa muuttujissa:
# tilinylityspaivat_luok         0.2795146
# lainarastit_luok               0.2966631
# luottokortteja_yhteensa_luok   0.3368983
# Pidetään muuttujat kuitenkin mukana mallissa

# Lataukset muuttujittain
malli_pca2$loadings

# PC1: automaattinostoja_luok, pankkikorttilkm_luok,
maksuautomaattitapahtumia_luok,
# ottoja_luok, pkorttimaksuja_luok, panoja_luok, palveluja_kpl_luok,
# tilinylityspaivat_luok
# PC2: korkeakork_kpl_luok, maaraaikaistileja_luok, kayttotili_tal_luok,
# saastotililla_luok, suoraveloituksia_luok, maksupalvelussa_maksut_luok
# PC3: rahasto_b1_luok, rahastolajeja_luok, rahasto_c1_luok
# PC4: autom_lainan_perinta_luok, lainojen_lukumaara_luok,
asuntoluotot1_luok,
# netissa_maksut_luok, kv_maksukortit_luok, kulutusluotot1_luok,
# asuntolaina_a_kpl_luok, asuntolaina_c_kpl_luok, palveluja_kpl_luok
# PC5: osakkeet_euroa_1_luok, eri_osakesarjoja_luok
# PC6: vakuutus_a_luok, asuntolaina_a_kpl_luok,
luottokortteja_yhteensa_luok,
# korttiluotot1_luok
# PC7: lainarastit_luok, maksupalvelussa_maksut_luok, tiskilla_maksut_luok
# kulutusluotot1_luok

# Ja negatiivisesti
# PC2: maksuautomaattitapahtumia_luok, lainarastit_luok,
tilinylityspaivat_luok
# PC6: kv_maksukortit_luok
# PC7: netissa_maksut_luok

# Tallennetaan pääkomponenttipistemäärät
pca_pistemaarat <- as.matrix(malli_pca2$scores)
```

```
# Nimetään uudet muuttujat niille latautuneiden muuttujien /
palvelutyyppien mukaan
colnames(pca_pistemaarat) <- c("a1_kayttotili","a2_saastotili",
"a3_rahasto", "a4_laina", "a5_osake", "a6_vakuutus", "a7_tiski")

#######################
# 3. Klusterianalyysi #
#######################

# Valitaan klusterianalyysiin kaikki a-muuttujat, a1-a7
malli_clust <- Mclust(pca_pistemaarat)
# best model: diagonal, varying volume and shape (VVI) with 5 components
klusterit_maara <- 5
klusterit_hclust <- hclust(dist(pca_pistemaarat), method="average")
# Etsitään keskukset
klusterit_keskukset <- as.matrix(tapply(pca_pistemaarat,
list(rep(cutree(klusterit_hclust, klusterit_maara), ncol(pca_pistemaarat)),
col(pca_pistemaarat)), mean))
colnames(klusterit_keskukset) <- as.list(dimnames(pca_pistemaarat)[[2]])

# Siirytään k-means klusterointiin, käytetään hierarkkisella
# klusteroinnilla haettuja keskuksia
klusterit_kmeans <- kmeans(pca_pistemaarat, centers=klusterit_keskukset)
# cluster means
# aggregate(pca_pistemaarat,by=list(klusterit_kmeans$cluster),FUN=mean)
plotcluster(pca_pistemaarat, klusterit_kmeans$cluster)
pdf('kmeans_5_klusteria.pdf')
dev.off()

# Graafista havaitaan, että klusteriin kolme päätyy vain kolme
# havaintoa, lasketaan klusterien määrää neljään
klusterit_maara <- 4
klusterit_hclust <- hclust(dist(pca_pistemaarat), method="average")
# Etsitään keskukset
klusterit_keskukset <- as.matrix(tapply(pca_pistemaarat,
list(rep(cutree(klusterit_hclust, klusterit_maara), ncol(pca_pistemaarat)),
col(pca_pistemaarat)), mean))
colnames(klusterit_keskukset) <- as.list(dimnames(pca_pistemaarat)[[2]])
klusterit_kmeans2 <- kmeans(pca_pistemaarat, centers=klusterit_keskukset)
plotcluster(pca_pistemaarat, klusterit_kmeans2$cluster)
pdf('kmeans_4_klusteria.pdf')
dev.off()

pca_pistemaarat[klusterit_kmeans2$cluster==3,] #130, 296, 2047
pca_pistemaarat[klusterit_kmeans$cluster==3,] #130, 296, 2047


# Kyseiset kolme havaintoa muodostavat edelleen oman klusterinsa
```

```
# poistetaan havainnot ja palataan viiteen klusteriin
# (kolmen havainnon perusteella ei järkevää rakentaa asiakassegmenttiä)
pca_pistemaarat_klusterit <- pca_pistemaarat[klusterit_kmeans2$cluster!=3,]

klusterit_maara <- 5
klusterit_hclust <- hclust(dist(pca_pistemaarat_klusterit),
method="average")
# Etsitään keskukset
klusterit_keskukset <- as.matrix(tapply(pca_pistemaarat_klusterit,
list(rep(cutree(klusterit_hclust, klusterit_maara),
ncol(pca_pistemaarat_klusterit)), col(pca_pistemaarat_klusterit)), mean))
colnames(klusterit_keskukset) <-
as.list(dimnames(pca_pistemaarat_klusterit)[[2]])
klusterit_kmeans3 <- kmeans(pca_pistemaarat_klusterit,
centers=klusterit_keskukset, iter.max=50)
plotcluster(pca_pistemaarat_klusterit, klusterit_kmeans3$cluster)
pdf('kmeans_5_klusteria_b.pdf')
dev.off()

# Saadaan selkeät kolme klusteria, joista yksi koostuu
# klustereista 1, 3 ja 5

# Tulostetaan klusteriprofiilit
klusterit_kmeans3$centers

#           a1_kayttotili a2_saastotili a3_rahasto    a4_laina    a5_osake
a6_vakuutus
# klusteri1    -0.8047753    -0.4519664 -0.2739878 -0.18701888 -0.28499522
-0.333005011
# klusteri2     0.2313708     0.1827550  0.6026877 -0.04420253  3.40812615
-0.004559621
# klusteri3     0.7497662    -0.2982691  0.5846725 -0.08088073 -0.33915974
0.001166909
# klusteri4     0.2453241     0.3970253  0.7726770  5.02865062  0.93435144
-0.260605398
# klusteri5     0.6824509     1.5869745 -0.6884828 -0.36437771 -0.08396267
0.923484347

#               a7_tiski
# klusteri1   -0.19861906
# klusteri2    0.08569227
# klusteri3    0.30040246
# klusteri4   -0.10219167
# klusteri5   -0.21107184

# Klusteri1: Ei profiloivaa muuttujaa / pienet keskiarvot
# Klusteri5: a2_saastotili, a6_vakuutus
# Klusteri3: a1_kayttotili, a7_tiski
```

```
# Klusteri2: a5_osake
# Klusteri4: a4_laina, a3_rahasto

klusterit_kmeans3$size
#      1   2   3   4   5
#[1] 445  60 294  33 165


par(mfrow=c(2,1))
klusterit_maara <- 3
klusterit_hclust <- hclust(dist(pca_pistemaarat_klusterit), method="ward")
# Etsitään keskukset
klusterit_keskukset <- as.matrix(tapply(pca_pistemaarat_klusterit,
list(rep(cutree(klusterit_hclust, klusterit_maara),
ncol(pca_pistemaarat_klusterit)), col(pca_pistemaarat_klusterit)), mean))
colnames(klusterit_keskukset) <-
as.list(dimnames(pca_pistemaarat_klusterit)[[2]])
klusterit_kmeans4 <- kmeans(pca_pistemaarat_klusterit,
centers=klusterit_keskukset, iter.max=50)
plotcluster(pca_pistemaarat_klusterit, klusterit_kmeans4$cluster)

klusterit_maara <- 4
klusterit_hclust <- hclust(dist(pca_pistemaarat_klusterit), method="ward")
# Etsitään keskukset
klusterit_keskukset <- as.matrix(tapply(pca_pistemaarat_klusterit,
list(rep(cutree(klusterit_hclust, klusterit_maara),
ncol(pca_pistemaarat_klusterit)), col(pca_pistemaarat_klusterit)), mean))
colnames(klusterit_keskukset) <-
as.list(dimnames(pca_pistemaarat_klusterit)[[2]])
klusterit_kmeans4 <- kmeans(pca_pistemaarat_klusterit,
centers=klusterit_keskukset, iter.max=50)
plotcluster(pca_pistemaarat_klusterit, klusterit_kmeans4$cluster)

pdf('kmeans_3-4_klusteria.pdf')
dev.off()

# Mahdolliset asiakassegmentit =>
# 1: Tiliasiakkaat (käyttö- ja säästö)
# 2: Osakeasiakkaat 3: Laina-asiakkaat

# Suurin klusteri no 1 ... passiiviset asiakkaat
# jos ei uusien palvelujen ostohalukkuutta, kannattaa klusterilta
# todennäköisesti alkaa periä suurempia "tilinhoito"- ja "palvelu"maksuja,
# mikä on helppo perustella esim. kasvaneilla kustannuksilla (vaikka
tosiasiassa
# kustannukset ovat erityisesti irtisanomisten myötä pienentyneet)
```