Satinitigan_Karl_HW7

Karl Satinitigan

3/15/2020

k-means clustering by hand

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidymodels)
```

```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo
```

```
## -- Attaching packages --------------------------------------------------- tidymodels 0.0.3 --
```

```
## v broom     0.5.4     v recipes   0.1.9
## v dials     0.0.4     v rsample   0.0.5
## v infer     0.5.1     v yardstick 0.0.4
## v parsnip   0.0.5
```

```
## -- Conflicts ------------------------------------------------------ tidymodels_conflicts() --
## x scales::discard()   masks purrr::discard()
## x dplyr::filter()     masks stats::filter()
## x recipes::fixed()    masks stringr::fixed()
## x dplyr::lag()        masks stats::lag()
## x dials::margin()     masks ggplot2::margin()
## x yardstick::spec()   masks readr::spec()
## x recipes::step()     masks stats::step()
## x recipes::yj_trans() masks scales::yj_trans()
```

```
library(patchwork)
library(here)
```

```
## here() starts at /Users/karl/Documents/UChicago/0 Computational Modeling/Problem Sets/Satinitigan_Ka
```

```
library(tictoc)
library(ggdendro)
library(gganimate)
```

```
## No renderer backend detected. gganimate will default to writing frames to separate files
## Consider installing:
## - the `gifski` package for gif output
## - the `av` package for video output
```

```
## and restarting the R session
```

```
library(cluster)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(skimr)
library(e1071)
library(dplyr)
library(dbscan)

set.seed(1234)
theme_set(theme_minimal())


input_1 <- c(5,8,7,8,3,4,2,3,4,5)
input_2 <- c(8,6,5,4,3,2,2,8,9,8)

inputs <- data.frame(x1 = input_1, x2 = input_2)

krandom <- inputs %>% mutate(k = as.factor(sample.int(3, 10, replace = TRUE)), id = 1:10)

edistance <- function(x1, x2, cen1, cen2){
  return ((x1 - cen1)^2 + (x2-cen2)^2)
}

newkrandom <- krandom
for (i in 1:25){
  centroid <- newkrandom %>%
    mutate(centroidk = k) %>%
    group_by(centroidk) %>%
    summarise(x1_c = mean(x1), x2_c = mean(x2))
  newerkrandom <- merge(centroid, krandom) %>%
    mutate(distance = edistance(x1, x2, x1_c, x2_c)) %>%
    group_by(id) %>%
    slice(which.min(distance)) %>%
    mutate(k = centroidk) %>%
    select(x1, x2, k)
  if (nrow(full_join(newkrandom, newerkrandom)) == 10) {
    break()
  }
  newkrandom <- newerkrandom
}
```

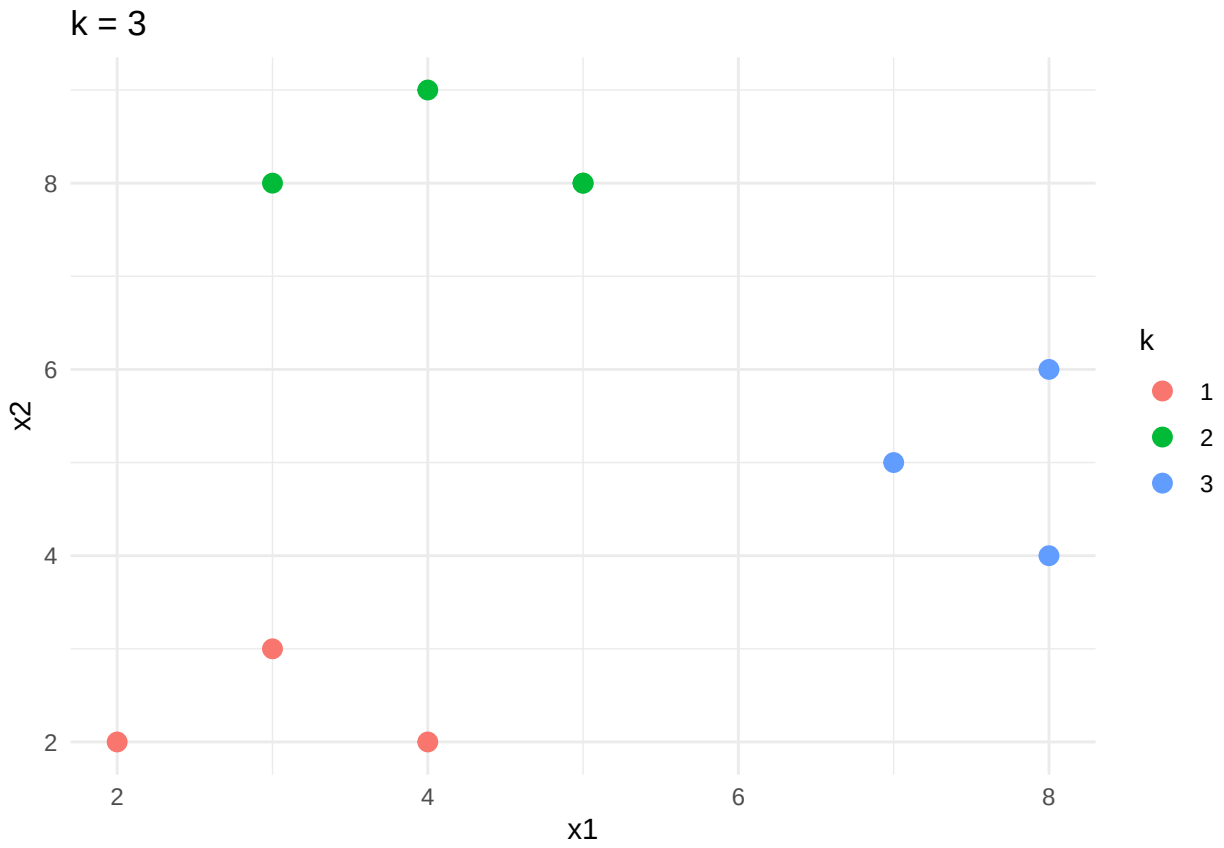
```
## Adding missing grouping variables: `id`
```

```
## Joining, by = c("x1", "x2", "k", "id")
```

```
## Adding missing grouping variables: `id`
```

```
## Joining, by = c("id", "x1", "x2", "k")
```

```
finalplot <- newkrandom %>% ggplot(aes(x1, x2, color = k)) + geom_point(size = 3) + ggtitle("k = 3")
finalplot
```

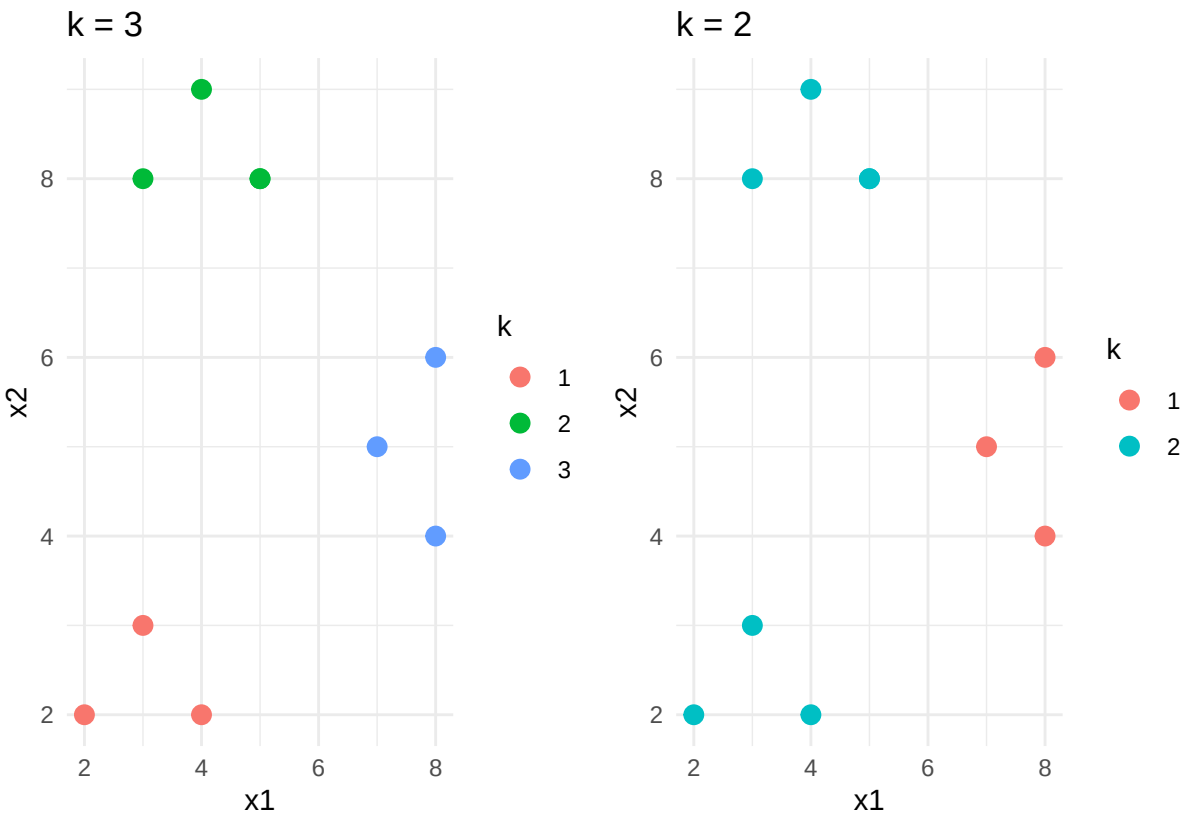



```
krandom2 <- inputs %>% mutate(k = as.factor(sample.int(2, 10, replace = TRUE)), id = 1:10)

newkrandom <- krandom2
for (i in 1:25){
  centroid <- newkrandom %>%
    mutate(centroidk = k) %>%
    group_by(centroidk) %>%
    summarise(x1_c = mean(x1), x2_c = mean(x2))
  newerkrandom <- merge(centroid, krandom2) %>%
    mutate(distance = edistance(x1, x2, x1_c, x2_c)) %>%
    group_by(id) %>%
    slice(which.min(distance)) %>%
    mutate(k = centroidk) %>%
    select(x1, x2, k)
  if (nrow(full_join(newkrandom, newerkrandom)) == 10) {
    break()
  }
  newkrandom <- newerkrandom
}
```

```
## Adding missing grouping variables: `id`
## Joining, by = c("x1", "x2", "k", "id")
## Adding missing grouping variables: `id`
## Joining, by = c("id", "x1", "x2", "k")
```

```
finalplot2 <- newkrandom %>% ggplot(aes(x1, x2, color = k)) + geom_point(size = 3) + ggtitle("k = 2")
finalplot + finalplot2
```



The plots suggest that k = 3 fits the data better than k = 2. This is because when k = 3, the distance between those in the same cluster is smaller as compared to when k = 2.

Dimension reduction

```
library(tidyverse)
library(tidymodels)
library(patchwork)
library(here)
library(ggfortify)
library(Rtsne)
library(rjson)
library(furrr)
```

```
## Loading required package: future
```

```
library(tictoc)
library(ggplot2)

wiki <- read_csv(url("https://raw.githubusercontent.com/ksatinitigan/problem-set-7/master/data/wiki.csv
```

```
## Parsed with column specification:
## cols(
##   .default = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
wikiPCA <- prcomp(wiki, center = TRUE, scale = TRUE)

summary(wikiPCA)
```

```
## Importance of components:
##                            PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.6058 1.90586 1.69219 1.52365 1.46547 1.38228 1.31487
## Proportion of Variance 0.2281 0.06372 0.05024 0.04073 0.03768 0.03352 0.03033
## Cumulative Proportion  0.2281 0.29183 0.34207 0.38280 0.42047 0.45399 0.48433
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     1.20613 1.17385 1.16779 1.13641 1.08654 1.07515 1.04158
## Proportion of Variance 0.02552 0.02417 0.02393 0.02266 0.02071 0.02028 0.01903
## Cumulative Proportion  0.50985 0.53402 0.55795 0.58060 0.60132 0.62160 0.64063
##                           PC15    PC16    PC17    PC18    PC19    PC20    PC21
## Standard deviation     1.01080 0.99808 0.99197 0.96071 0.93339 0.91156 0.90379
## Proportion of Variance 0.01792 0.01748 0.01726 0.01619 0.01528 0.01458 0.01433
## Cumulative Proportion  0.65855 0.67603 0.69329 0.70949 0.72477 0.73935 0.75368
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.8771 0.85951 0.82448 0.80853 0.80231 0.78661 0.75050
## Proportion of Variance 0.0135 0.01296 0.01193 0.01147 0.01129 0.01086 0.00988
## Cumulative Proportion  0.7672 0.78014 0.79206 0.80353 0.81482 0.82568 0.83556
##                           PC29    PC30    PC31   PC32    PC33   PC34    PC35
## Standard deviation     0.73659 0.70268 0.70157 0.6878 0.68203 0.6708 0.64653
## Proportion of Variance 0.00952 0.00866 0.00864 0.0083 0.00816 0.0079 0.00733
## Cumulative Proportion  0.84508 0.85374 0.86238 0.8707 0.87884 0.8867 0.89407
##                           PC36    PC37    PC38    PC39    PC40    PC41    PC42
## Standard deviation     0.64385 0.62790 0.62332 0.61367 0.59686 0.57617 0.57549
## Proportion of Variance 0.00727 0.00692 0.00682 0.00661 0.00625 0.00582 0.00581
## Cumulative Proportion  0.90134 0.90826 0.91507 0.92168 0.92793 0.93375 0.93956
##                           PC43    PC44    PC45    PC46    PC47   PC48    PC49
## Standard deviation     0.5650 0.55613 0.55423 0.54026 0.53701 0.5231 0.51546
## Proportion of Variance 0.0056 0.00543 0.00539 0.00512 0.00506 0.0048 0.00466
## Cumulative Proportion  0.9452 0.95059 0.95598 0.96110 0.96616 0.9710 0.97562
##                           PC50    PC51    PC52    PC53    PC54    PC55    PC56
## Standard deviation     0.50816 0.49831 0.46804 0.46300 0.43911 0.36615 0.33486
## Proportion of Variance 0.00453 0.00436 0.00384 0.00376 0.00338 0.00235 0.00197
## Cumulative Proportion  0.98015 0.98451 0.98835 0.99211 0.99549 0.99784 0.99981
##                           PC57
## Standard deviation     0.10351
## Proportion of Variance 0.00019
## Cumulative Proportion  1.00000
```

```
str(wikiPCA)
```

```
## List of 5
##  $ sdev    : num [1:57] 3.61 1.91 1.69 1.52 1.47 ...
##  $ rotation: num [1:57, 1:57] 0.0218 0.0351 0.0305 0.0342 -0.0814 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:57] "age" "gender" "phd" "yearsexp" ...
```

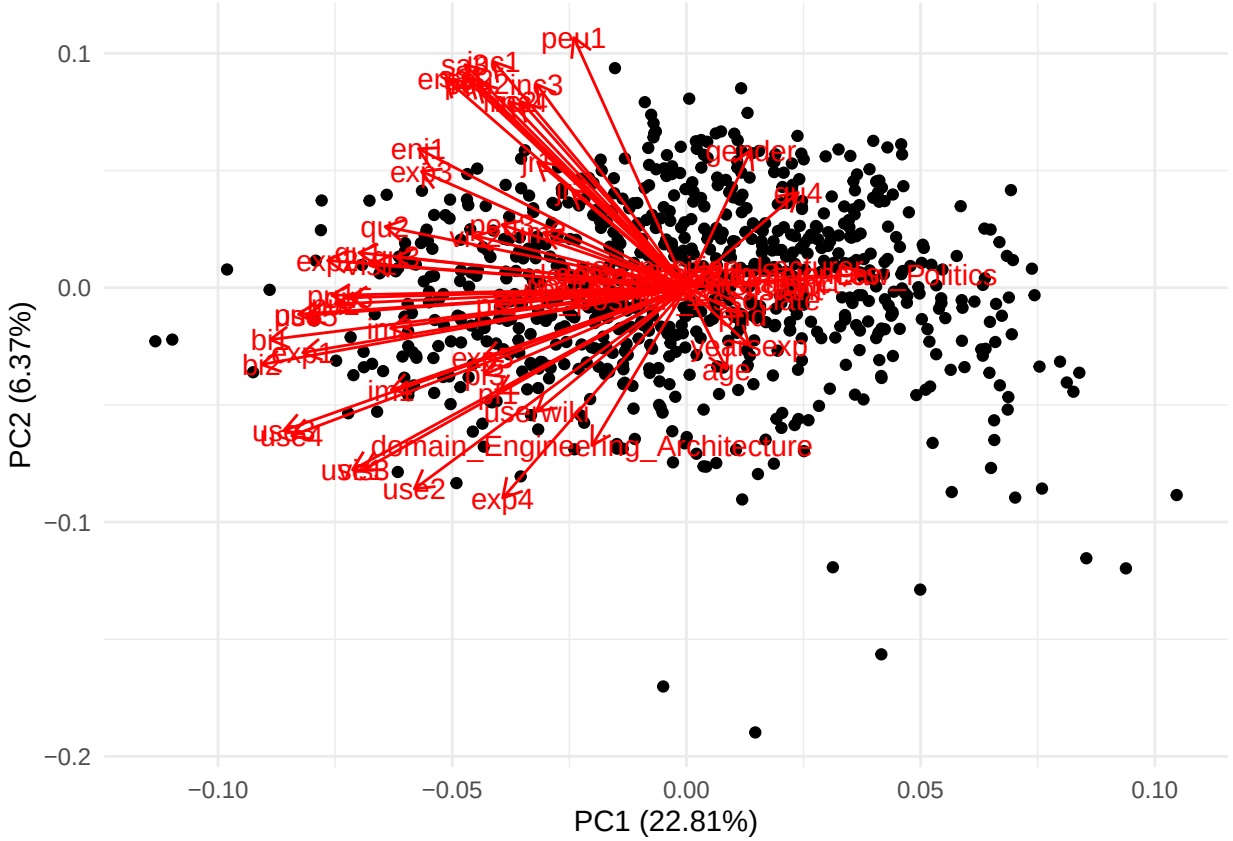
```
##   .. ..$ : chr [1:57] "PC1" "PC2" "PC3" "PC4" ...
##  $ center  : Named num [1:57] 42.166 0.427 0.434 10.409 0.136 ...
##   ..- attr(*, "names")= chr [1:57] "age" "gender" "phd" "yearsexp" ...
##  $ scale   : Named num [1:57] 7.548 0.495 0.496 6.757 0.343 ...
##   ..- attr(*, "names")= chr [1:57] "age" "gender" "phd" "yearsexp" ...
##  $ x       : num [1:800, 1:57] 0.15 3.31 4.68 -1.77 -7.25 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : chr [1:57] "PC1" "PC2" "PC3" "PC4" ...
##  - attr(*, "class")= chr "prcomp"
```

```
wikiPCA$rotation[,1:2]
```

```
##                                        PC1          PC2
## age                             0.021805412 -0.088384981
## gender                          0.035086317  0.149461450
## phd                             0.030501043 -0.030435497
## yearsexp                        0.034190490 -0.062364714
## userwiki                       -0.081363144 -0.134387358
## pu1                            -0.192827065 -0.008273053
## pu2                            -0.190587716 -0.017668791
## pu3                            -0.210862567 -0.028776289
## peu1                           -0.061228008  0.271741292
## peu2                           -0.113718709  0.222367978
## peu3                           -0.100218922  0.068458517
## enj1                           -0.145666175  0.151011732
## enj2                           -0.131109826  0.227602424
## qu1                            -0.178057029  0.038122429
## qu2                            -0.163777789  0.066421876
## qu3                            -0.157956174  0.033472352
## qu4                             0.060796858  0.103458415
## qu5                            -0.183364593 -0.010911790
## vis1                           -0.171153058  0.025207626
## vis2                           -0.114558913  0.056217768
## vis3                           -0.175351292 -0.197634740
## im1                            -0.160432141 -0.111106146
## im2                            -0.077810159  0.059774521
## im3                            -0.160803391 -0.044003603
## sa1                            -0.121658435  0.229926005
## sa2                            -0.117590405  0.226760395
## sa3                            -0.120376196  0.242325421
## use1                           -0.181477170 -0.197827499
## use2                           -0.147851769 -0.218628501
## use3                           -0.218809245 -0.155151571
## use4                           -0.214558397 -0.160864524
## use5                           -0.206538888 -0.029823253
## pf1                            -0.102337996 -0.114370782
## pf2                            -0.103448162 -0.018604706
## pf3                            -0.109632421 -0.094172517
## jr1                            -0.080866885  0.136967544
## jr2                            -0.062216127  0.106296824
## bi1                            -0.226193061 -0.056374273
## bi2                            -0.230923964 -0.083430888
## inc1                           -0.104666756  0.245439824
## inc2                           -0.095802250  0.202021404
```

```
## inc3                                -0.081401727  0.220985795
## inc4                                -0.089707244  0.202022006
## exp1                                -0.208591685 -0.070543836
## exp2                                -0.195043150  0.029560476
## exp3                                -0.144023257  0.126416909
## exp4                                -0.099872875 -0.228494272
## exp5                                -0.110628098 -0.076095685
## domain_Sciences                     -0.021982007  0.014536737
## domain_Health.Sciences               0.017157681  0.015478496
## domain_Engineering_Architecture     -0.051309109 -0.171483803
## domain_Law_Politics                  0.094774659  0.014887154
## uoc_position_Associate              -0.010922081 -0.013134181
## uoc_position_Assistant              -0.007123091 -0.002311281
## uoc_position_Lecturer                0.018040923  0.023591030
## uoc_position_Instructor             -0.004250607  0.003784534
## uoc_position_Adjunct                 0.007848555  0.005301224
```

```
autoplot(wikiPCA, data = wiki, loadings = TRUE, loadings.color = "green", loadings.label = TRUE, loadin
```



The summary shows that bi2 is most strongly negatively correlated on the first principal component. This is followed by bi1, exp1, and exp2. Meanwhile, peu1 is most strongly positively correlated on the second principal component. This is followed by sa3, inc1, and enj2.

```
wikiPCA_PVE <- (wikiPCA$sdev^2)/sum(wikiPCA$sdev^2)
wikiPCA_PVE
```

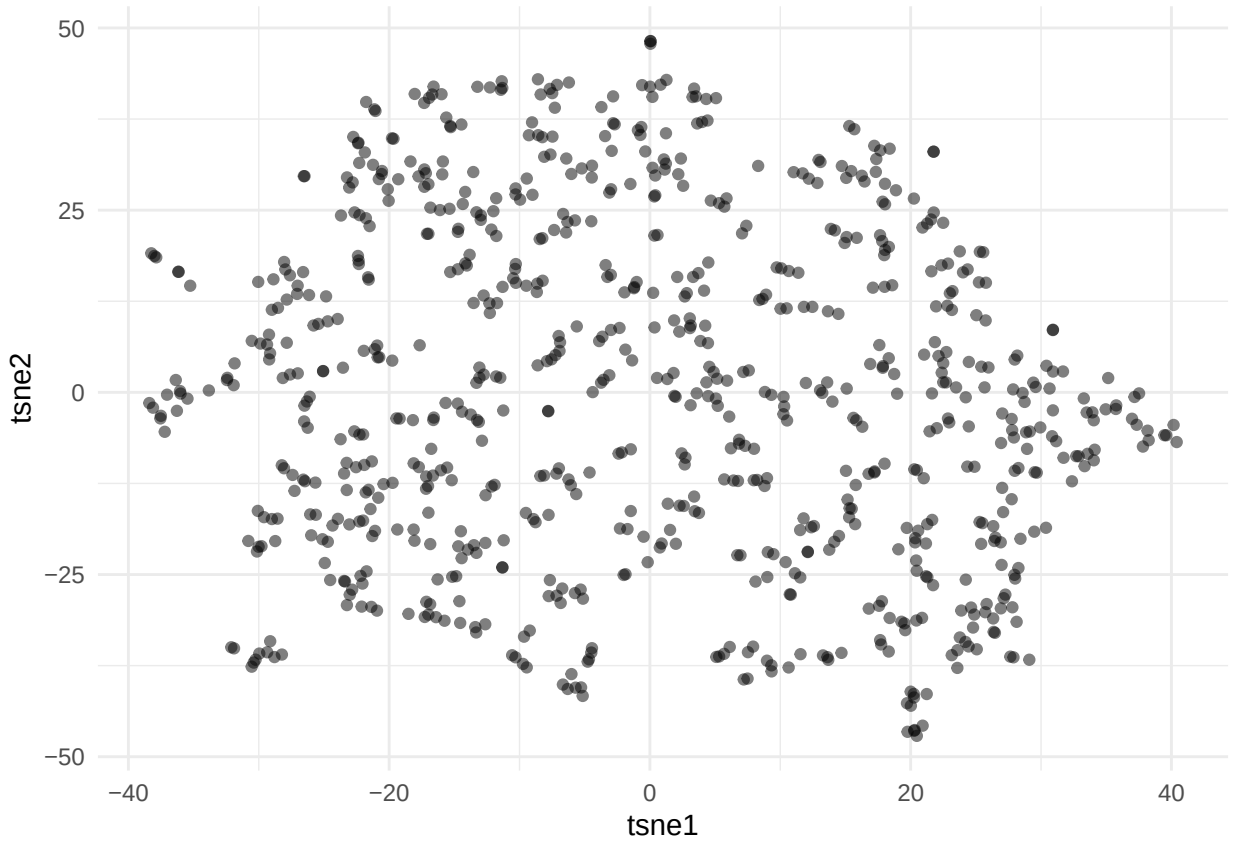
```
##  [1] 0.2281062779 0.0637247454 0.0502370687 0.0407283521 0.0376772356
##  [6] 0.0335209255 0.0303313773 0.0255217752 0.0241742687 0.0239251475
## [11] 0.0226565037 0.0207118345 0.0202799632 0.0190332986 0.0179249263
## [16] 0.0174765005 0.0172633331 0.0161923173 0.0152846094 0.0145779108
## [21] 0.0143303520 0.0134971703 0.0129607608 0.0119257101 0.0114687769
## [26] 0.0112930650 0.0108554417 0.0098814675 0.0095186872 0.0086625377
## [31] 0.0086350227 0.0082987836 0.0081607499 0.0078953167 0.0073334612
## [36] 0.0072727769 0.0069168040 0.0068163401 0.0066067617 0.0062497608
## [41] 0.0058240942 0.0058102814 0.0056003078 0.0054258856 0.0053889842
## [46] 0.0051207775 0.0050593384 0.0048003373 0.0046613631 0.0045302452
## [51] 0.0043563075 0.0038432203 0.0037608469 0.0033827360 0.0023520364
## [56] 0.0019671669 0.0001879532
```

```
sum(wikiPCA_PVE)
```

```
## [1] 1
```

The first principal component explains 22.81% of the variance. The second principal component explains 6.37% of the variance. In total, both explain 29.18% of the variance. The cumulative PVE is 1.

```
wikitSNE <- Rtsne(as.matrix(wiki), perplexity = 5)
wikitSNEplot <- wiki %>% mutate(tsne1 = wikitSNE$Y[, 1], tsne2 = wikitSNE$Y[, 2]) %>% ggplot(aes(tsne1,
wikitSNEplot
```



The plot does not seem to reveal any observable clusters.

Clustering

```
wikiPCAfortifiy <- fortify(wikiPCA)
wikiScaled <- scale(wiki)

wikik2 <- kmeans(wikiScaled, 2, nstart = 25, iter.max = 50)
wikik2PCA <- cbind(wikiPCAfortifiy, group = wikik2$cluster)

wikik2PCAplot <- ggplot(wikik2PCA) + geom_point(aes(x = PC1, y = PC2, col = factor(group), text = rowna
```

```
## Warning: Ignoring unknown aesthetics: text
```

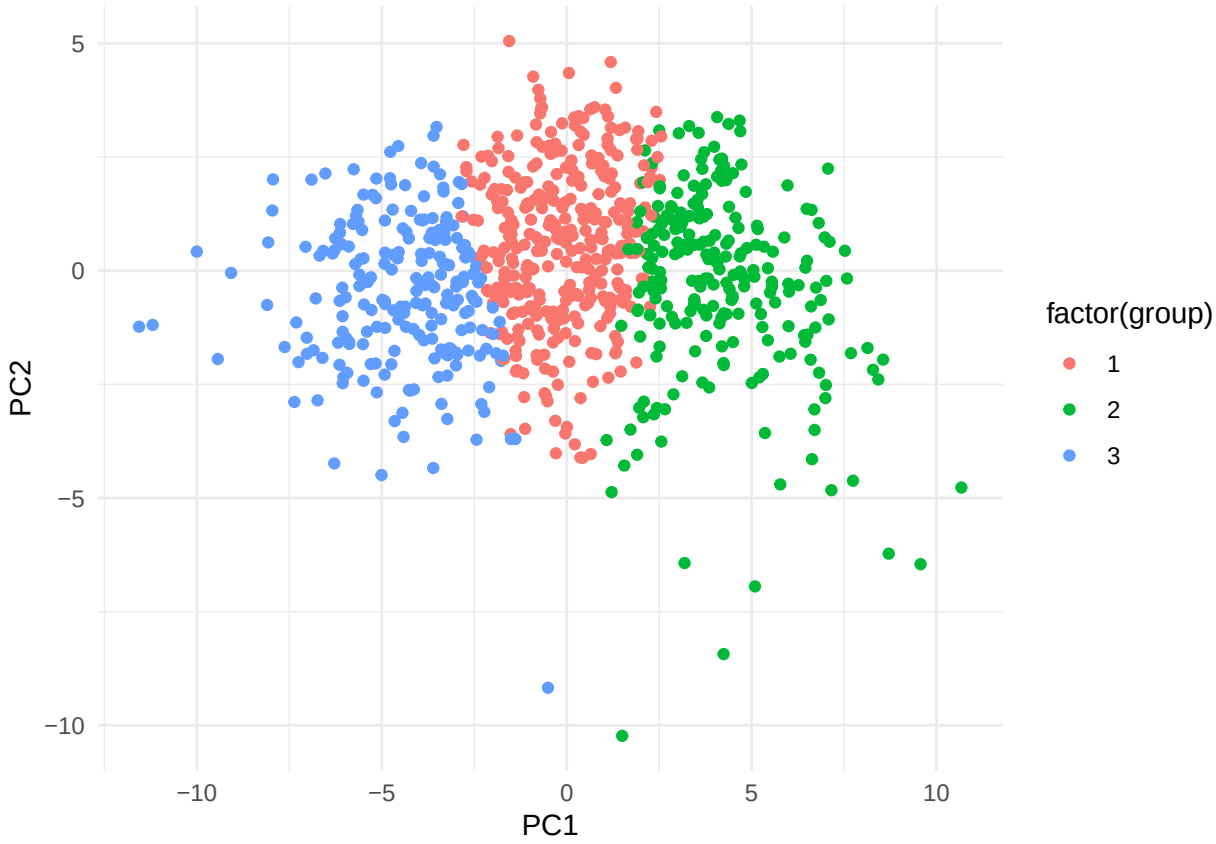
```
wikik2PCAplot
```


```
wikik3 <- kmeans(wikiScaled, 3, nstart = 25, iter.max = 50)
wikik3PCA <- cbind(wikiPCAfortifiy, group = wikik3$cluster)

wikik3PCAplot <- ggplot(wikik3PCA) + geom_point(aes(x = PC1, y = PC2, col = factor(group), text = rowna
```

```
## Warning: Ignoring unknown aesthetics: text
```

```
wikik3PCAplot
```

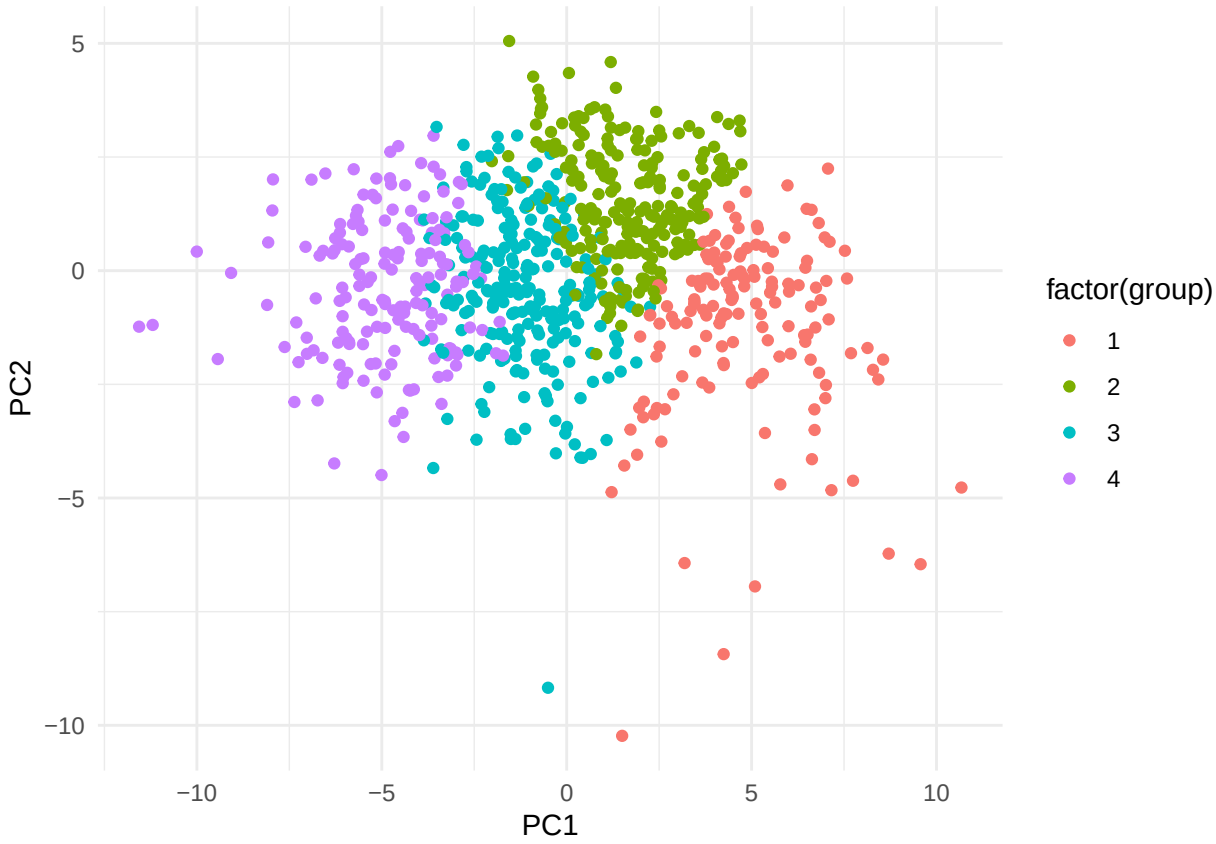



```
wikik4 <- kmeans(wikiScaled, 4, nstart = 25, iter.max = 50)
wikik4PCA <- cbind(wikiPCAfortifiy, group = wikik4$cluster)

wikik4PCAplot <- ggplot(wikik4PCA) + geom_point(aes(x = PC1, y = PC2, col = factor(group), text = rowna
```

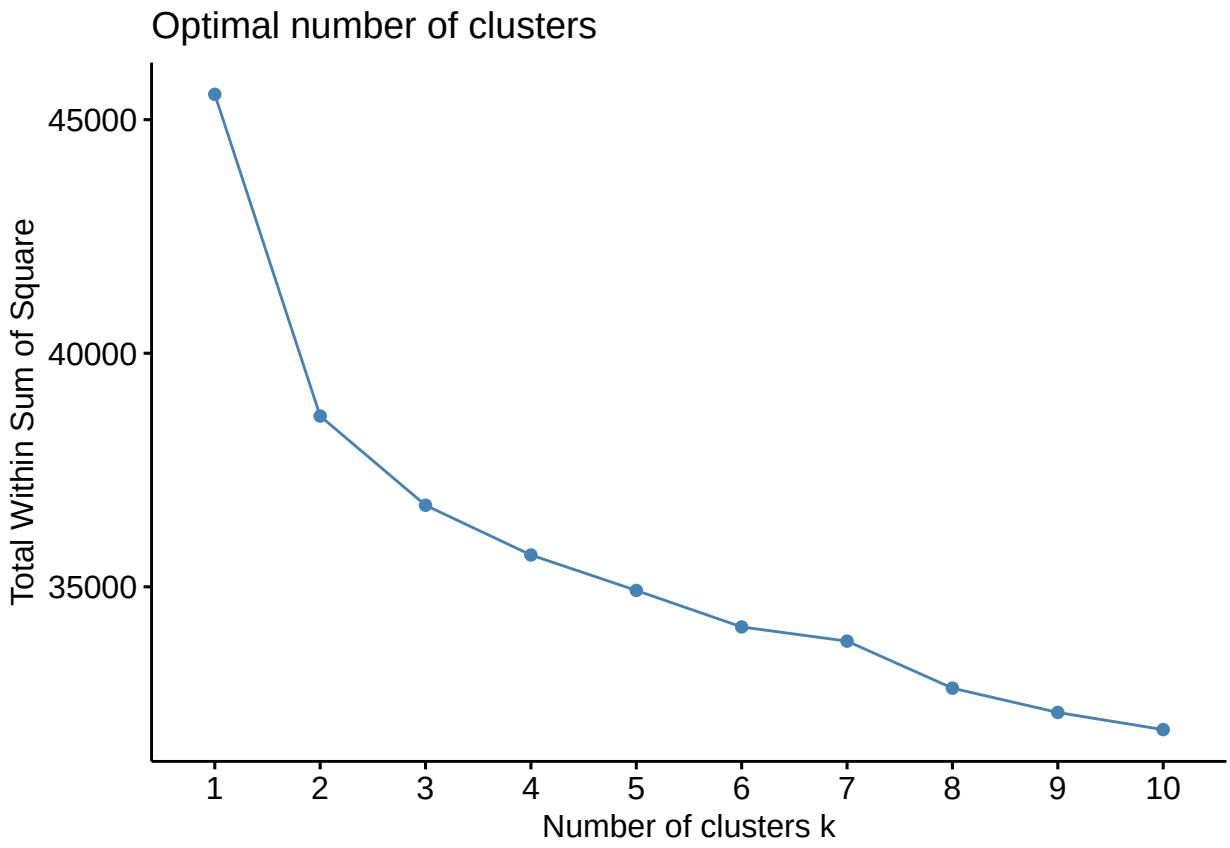
```
## Warning: Ignoring unknown aesthetics: text
```

```
wikik4PCAplot
```

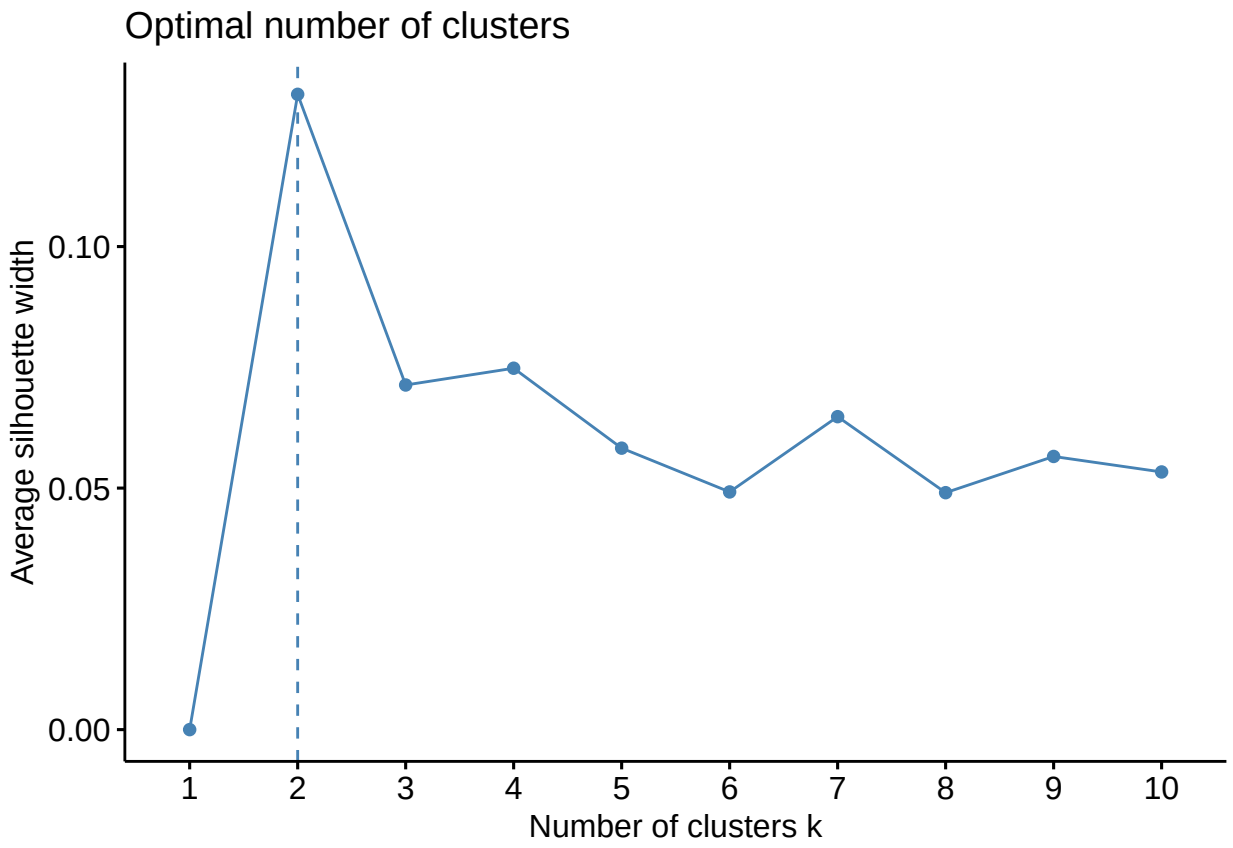


The plots show that the first principal component divides the different clusters as k goes from 2 to 4. There is least overlap when k = 2 and this suggests that this may be the best fit.

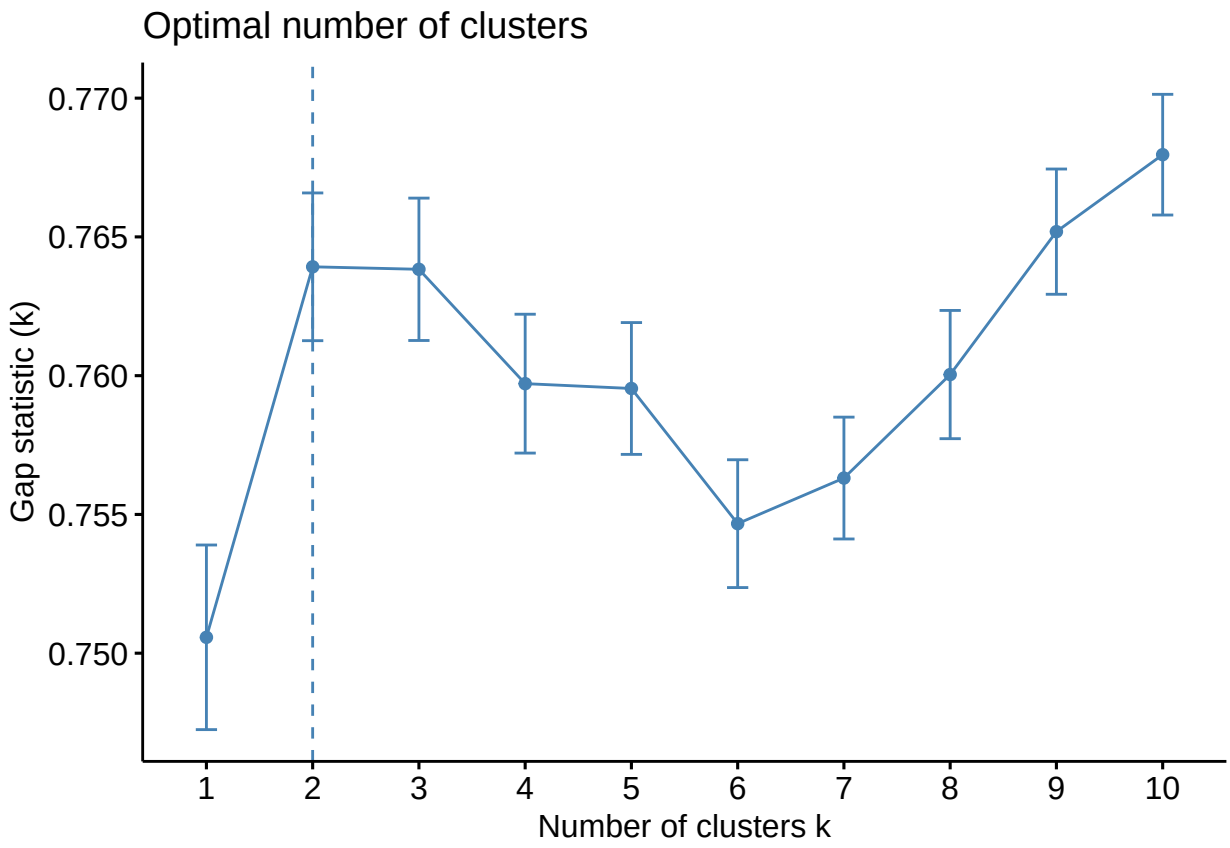
```
fviz_nbclust(wikiScaled, kmeans, method = "wss")
```




```
fviz_nbclust(wikiScaled, kmeans, method = "silhouette")
```




```
wikigapstat <- clusGap(wikiScaled, FUN = kmeans, nstart = 25, iter.max = 50, K.max = 10, B = 50)
fviz_gap_stat(wikigapstat)
```

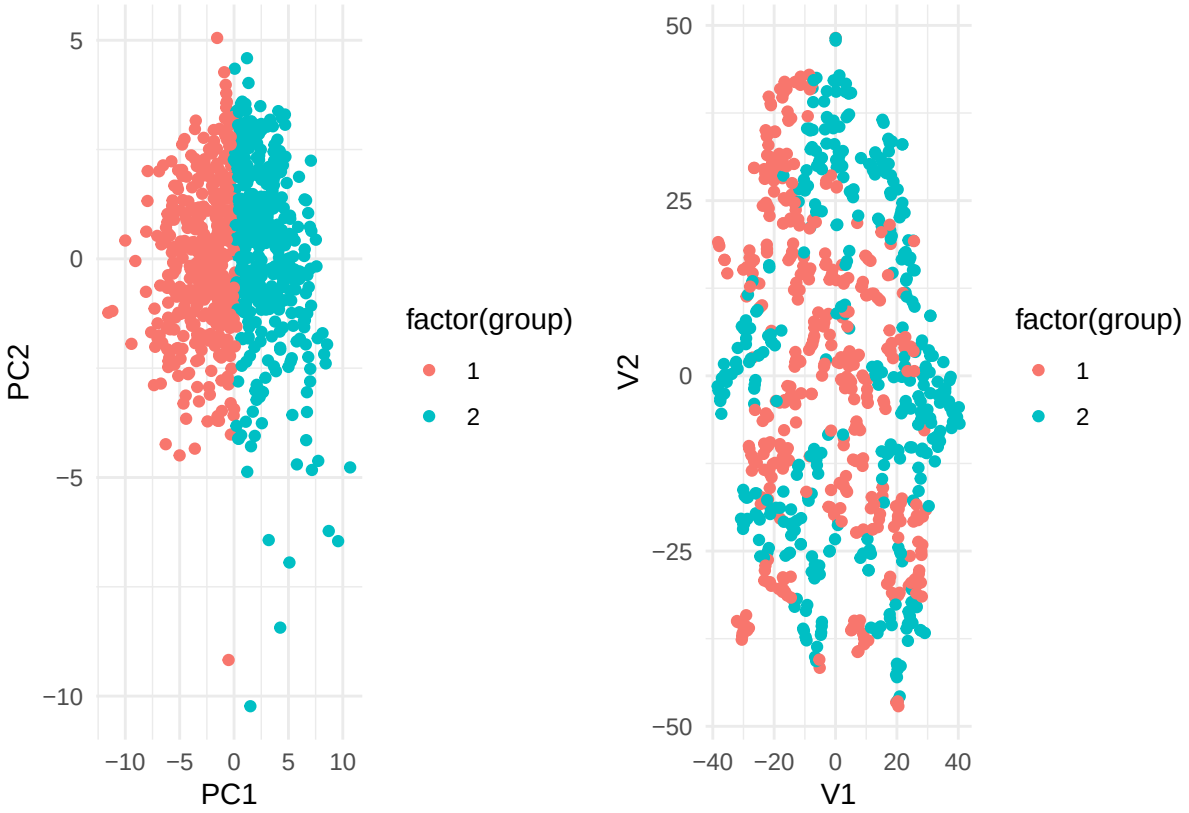


All methods suggest k = 2.

```
wikik2tSNE <- cbind(wikitSNE$Y, group = wikik2$cluster)
wikik2tSNEplot <- ggplot(wikik2tSNE) + geom_point(aes(x = V1, y = V2, col = factor(group), text = rowna
```

```
## Warning: Ignoring unknown aesthetics: text
```

```
wikik2PCAplot + wikik2tSNEplot
```



The PCA plot shows less overlap than the tSNE plot. The plots also show that t-SNE is a non-linear technique for dimensionality reduction.