Identificación de perfiles estudiantiles con técnicas de clustering

Pablo Parrilla Cañadas

2025-05-31

```
rm(list=ls())
library(dplyr)
```

```
##
## Adjuntando el paquete: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(fastDummies)
```

```
## Warning: package 'fastDummies' was built under R version 4.4.2
```

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

```
## Cargando paquete requerido: ggplot2

## Warning: package 'ggplot2' was built under R version 4.4.3

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

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.4.2

## Cargando paquete requerido: lattice
```

```
library(gridExtra)
```

```
##
## Adjuntando el paquete: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

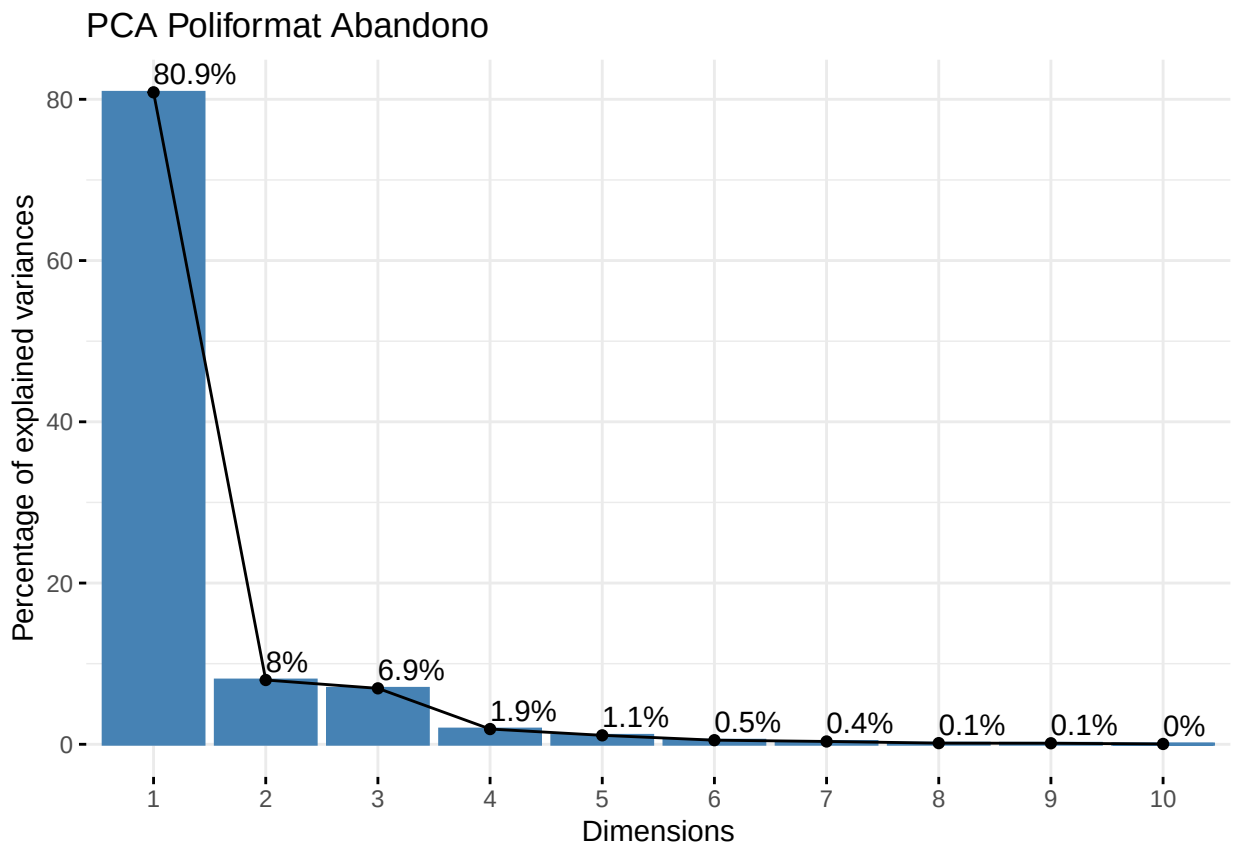
```
load("../../Datos/Capítulos/Caracterización.Rdata")
```

Preparar y concatenar

```
num_vars <- poliformat[poliformat$abandono==1,] %>% select(where(is.numeric))
```

```
## Adding missing grouping variables: `dni_hash`
```

```
num_vars= num_vars[,2:length(num_vars)]



datos_pca <- num_vars

preprocess <- preProcess(datos_pca, method = "medianImpute")
datos_imputados <- predict(preprocess, newdata = datos_pca)


datos_filtrados <- datos_imputados[, sapply(datos_imputados, function(col) var(col, na.rm = TRUE) != 0)]



pca_poliformat <- prcomp(datos_filtrados, scale. = TRUE)

fviz_eig(pca_poliformat, addlabels = TRUE, main="PCA Poliformat Abandono")
```




```
ambas=academicas[academicas$abandono==1,] %>% left_join(sociodemografia[academicas$abandono==1,], by="d
```

```
numeric_vars <- ambas %>% select(where(is.numeric))
categorical_vars <- ambas %>% select(where(~is.factor(.)))
```

```
categorical_dummy <- fastDummies::dummy_cols(categorical_vars, remove_first_dummy = TRUE, remove_select
```

```
combined <- bind_cols(numeric_vars, categorical_dummy)
```

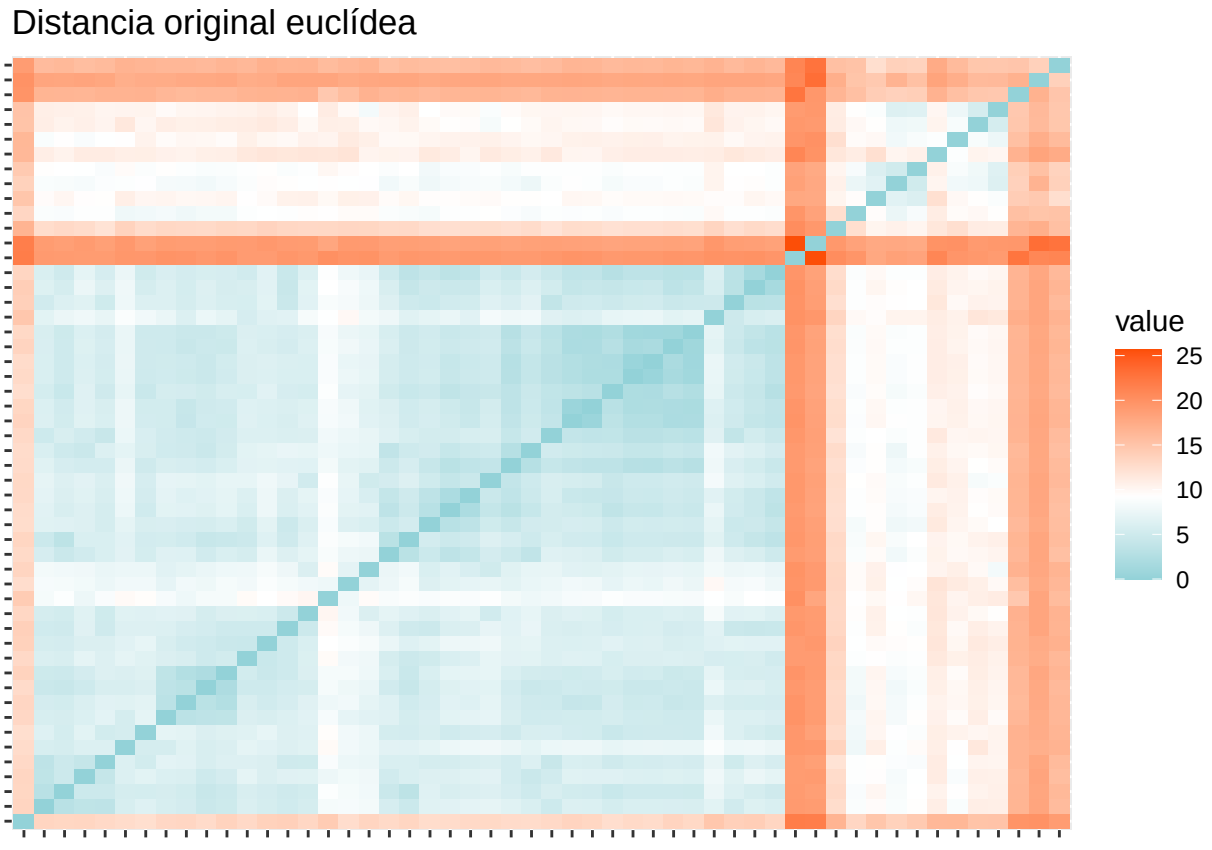
```
combined$actividades <- NULL
```

```
combined$nota14[is.na(combined$nota14)] <- median(combined$nota14, na.rm = TRUE)
```

```
combined <- combined %>%
  select(-rend_total_ultimo, -rend_total_penultimo, -rend_total_antepenultimo)
```

```r
combined_scaled <- scale(combined)

pca_df=pca_poliformat$x[,1:3]



final_dataset <- as.data.frame(cbind(combined_scaled, pca_df))
final_dataset <- final_dataset[, colSums(is.na(final_dataset)) < nrow(final_dataset)]

rownames(final_dataset)=1:52
```

```r
midist <- get_dist(final_dataset, stand = TRUE, method = "euclidean")
fviz_dist(midist, show_labels = TRUE, lab_size = 0.3,
          gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"))+labs(title = "Distancia or
```




```r
set.seed(10)
myN = c(7,25,37)  # m
myhopkins = NULL
myseed = sample(1:52, 2)
for (i in myN) {
  for (j in myseed) {
    tmp = get_clust_tendency(data = final_dataset, n = i, graph = FALSE, seed = j)
    myhopkins = c(myhopkins, tmp$hopkins_stat)
  }
```

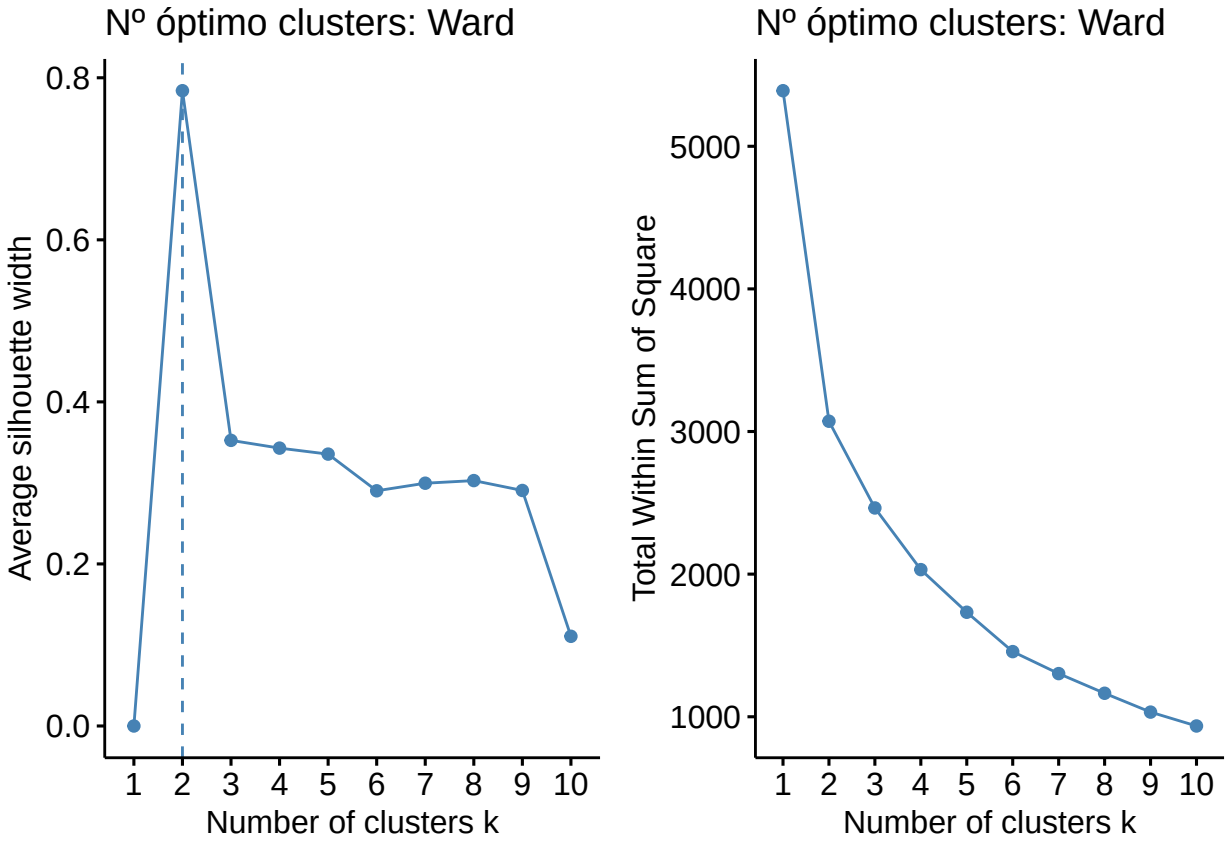
```
}
summary(myhopkins)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7353  0.7629  0.7902  0.7801  0.7961  0.8128
```

```
datos_elegidos=final_dataset
```

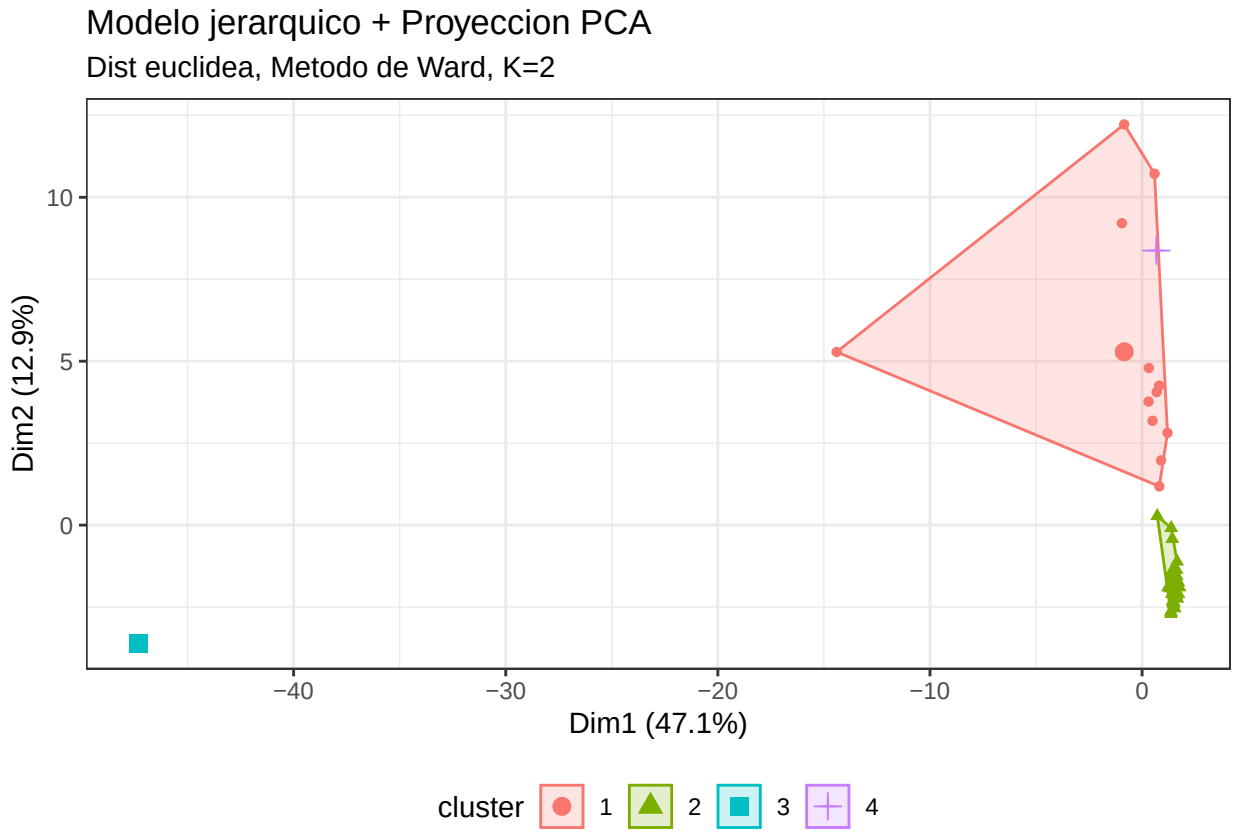
Método de Ward

```
p1 = fviz_nbclust(x =datos_elegidos, FUNcluster = hcut, method = "silhouette",
                  hc_method = "ward.D2", k.max = 10, verbose = FALSE,
                  hc_metric = "euclidean") + labs(title = "Nº óptimo clusters: Ward")
p2 = fviz_nbclust(x = datos_elegidos, FUNcluster = hcut, method = "wss",
                  hc_method = "ward.D2", k.max = 10, verbose = FALSE,
                  hc_metric = "euclidean") + labs(title = "Nº óptimo clusters: Ward")
grid.arrange(p1, p2, nrow = 1)
```



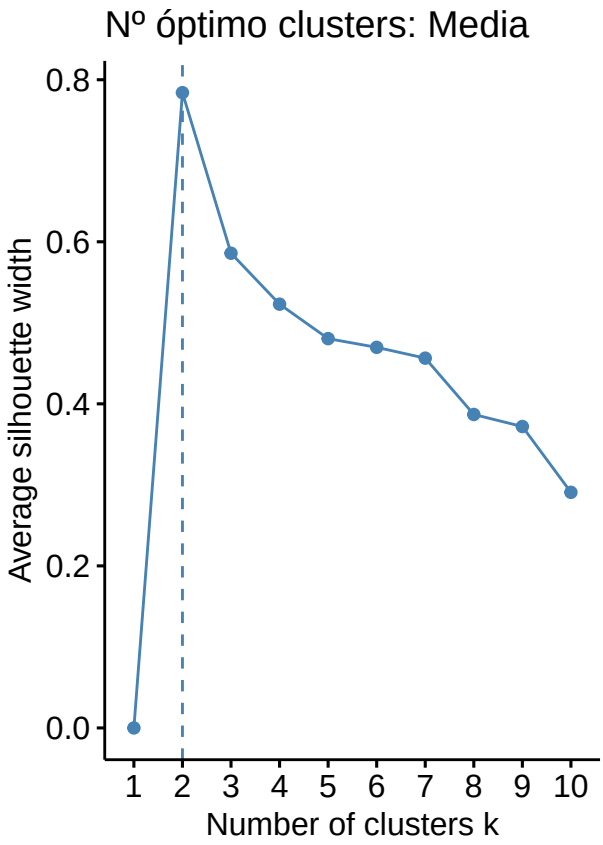

```
K=4

clust1 <- hclust(midist, method="ward.D2")
grupos1 <- cutree(clust1, k=K)

fviz_cluster(object = list(data=datos_elegidos, cluster=grupos1), stand = FALSE,
```

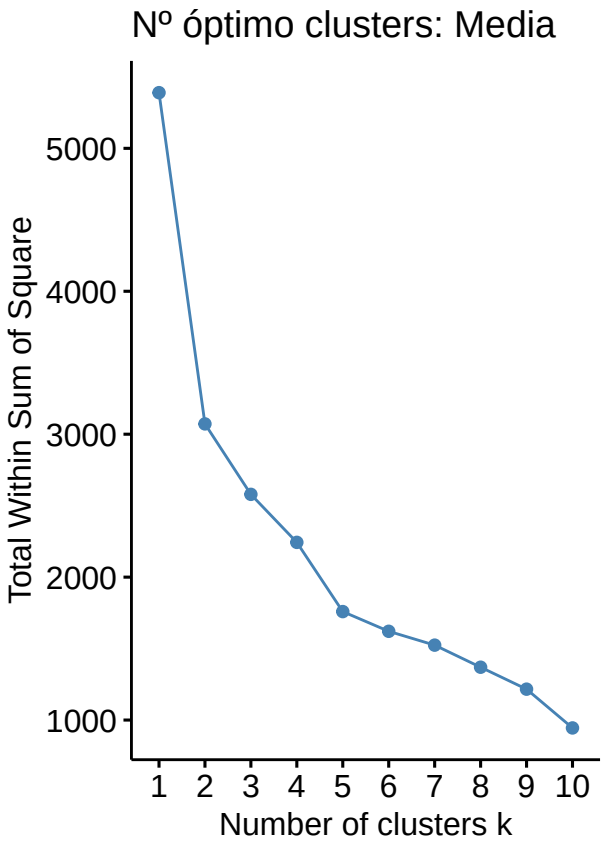
```
            ellipse.type = "convex", geom = "point", show.clust.cent = TRUE,
            labelsize = 8)  +
  labs(title = "Modelo jerarquico + Proyeccion PCA",
       subtitle = "Dist euclidea, Metodo de Ward, K=2") +
  theme_bw() +
  theme(legend.position = "bottom")
```



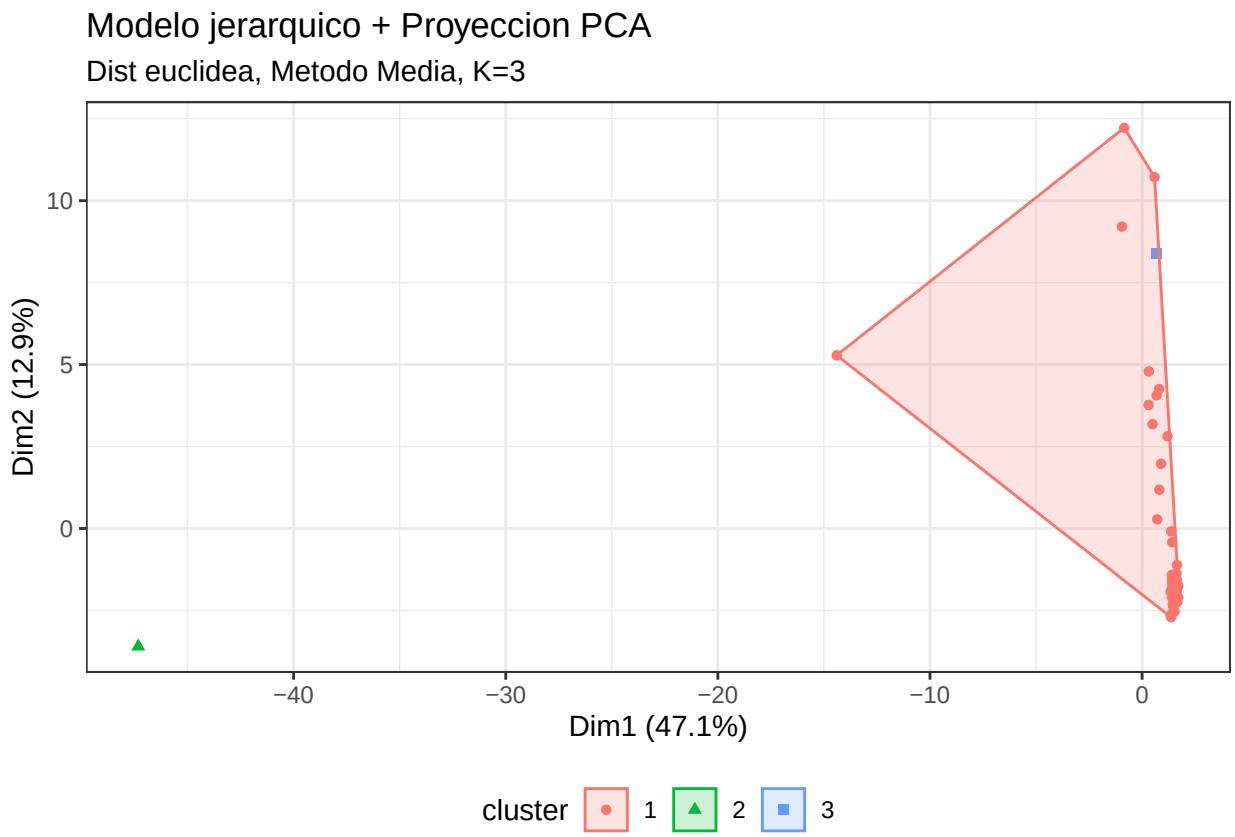
Método de la media

```
p1 = fviz_nbclust(x = datos_elegidos, FUNcluster = hcut, method = "silhouette",
                  hc_method = "average", k.max = 10, verbose = FALSE,
                  hc_metric = "euclidean") + labs(title = "Nº óptimo clusters: Media")
p2 = fviz_nbclust(x = datos_elegidos, FUNcluster = hcut, method = "wss",
                  hc_method = "average", k.max = 10, verbose = FALSE,
                  hc_metric = "euclidean") + labs(title = "Nº óptimo clusters: Media")
grid.arrange(p1, p2, nrow = 1)
```





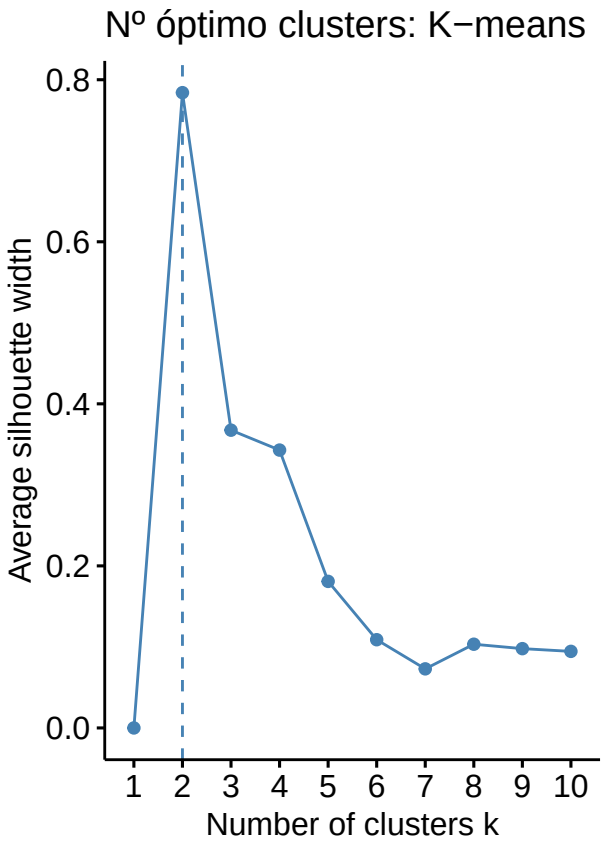

```r
K=3

clust1 <- hclust(midist, method="average")
grupos1 <- cutree(clust1, k=K)

fviz_cluster(object = list(data=datos_elegidos, cluster=grupos1), stand = FALSE,
             ellipse.type = "convex", geom = "point", show.clust.cent = FALSE,
             labelsize = 8)  +
  labs(title = "Modelo jerarquico + Proyeccion PCA",
       subtitle = "Dist euclidea, Metodo Media, K=3") +
  theme_bw() +
  theme(legend.position = "bottom")
```

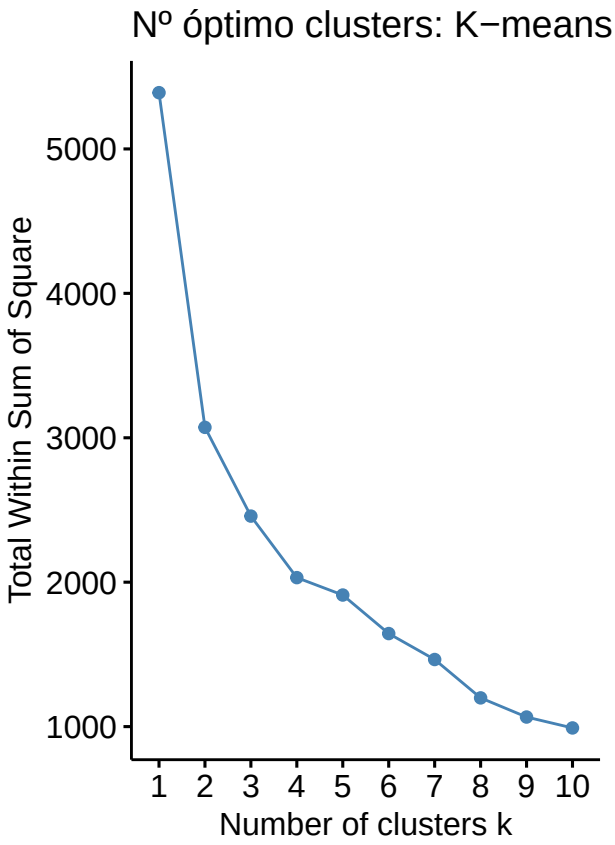


Partición

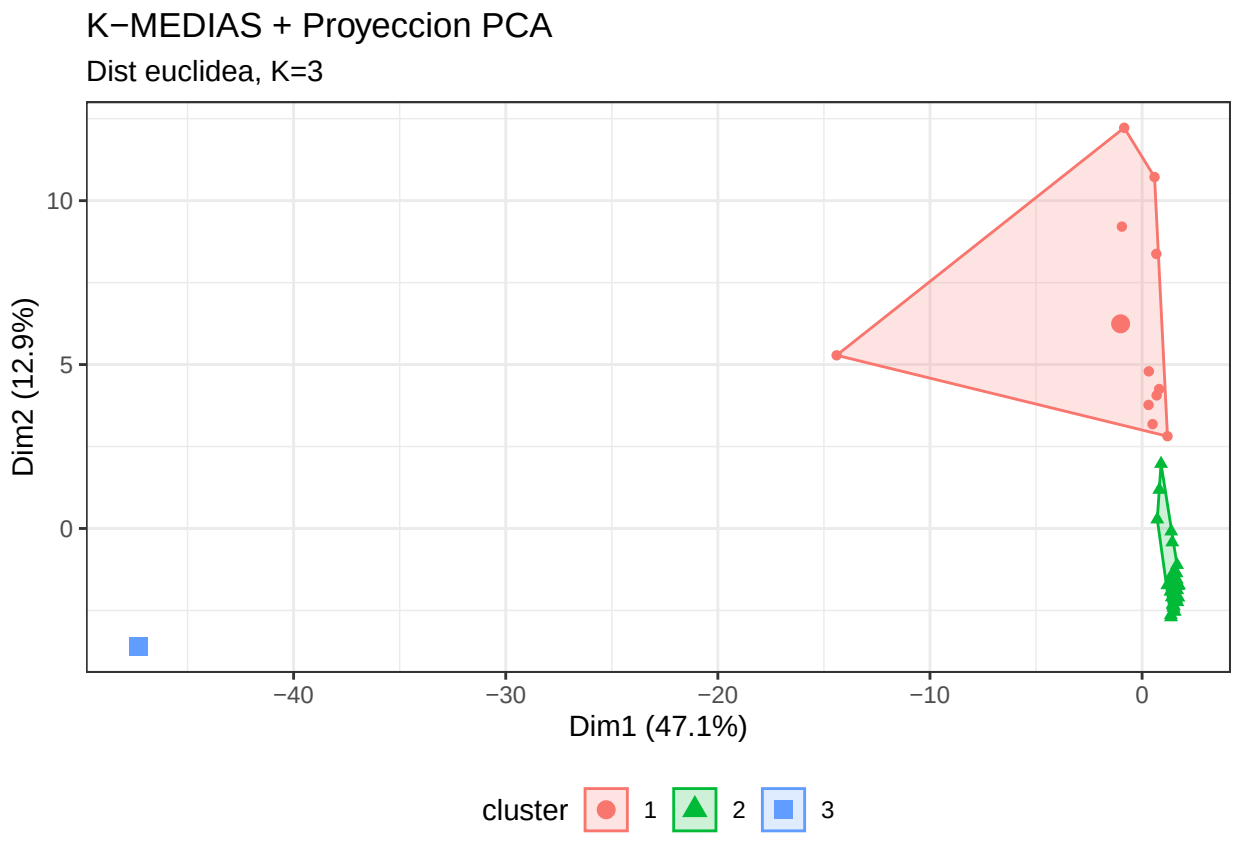
k-means

```
p1 = fviz_nbclust(x = datos_elegidos, FUNcluster = kmeans, method = "silhouette",
             k.max = 10, verbose = FALSE) +
  labs(title = "Nº óptimo clusters: K-means")
p2 = fviz_nbclust(x = datos_elegidos, FUNcluster = kmeans, method = "wss",
             k.max = 10, verbose = FALSE) +
  labs(title = "Nº óptimo clusters: K-means")
grid.arrange(p1, p2, nrow = 1)
```



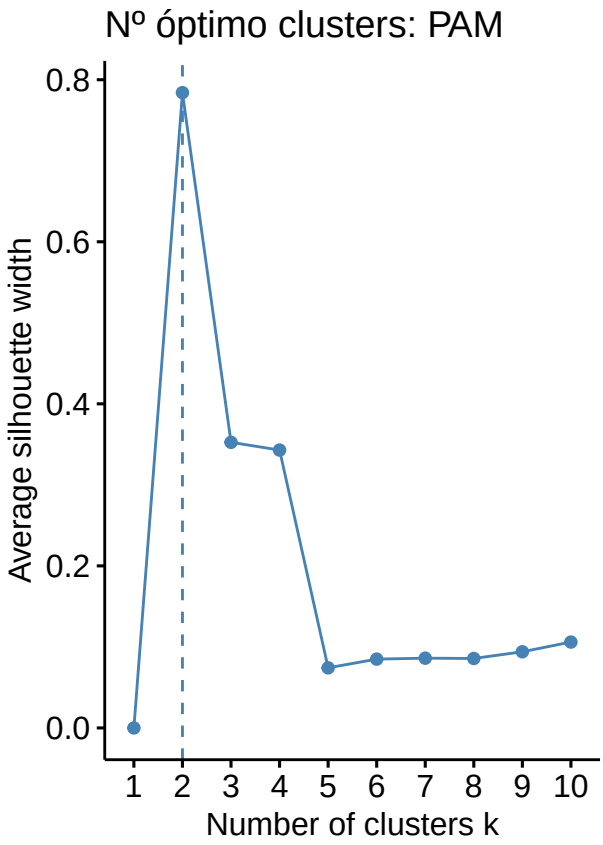


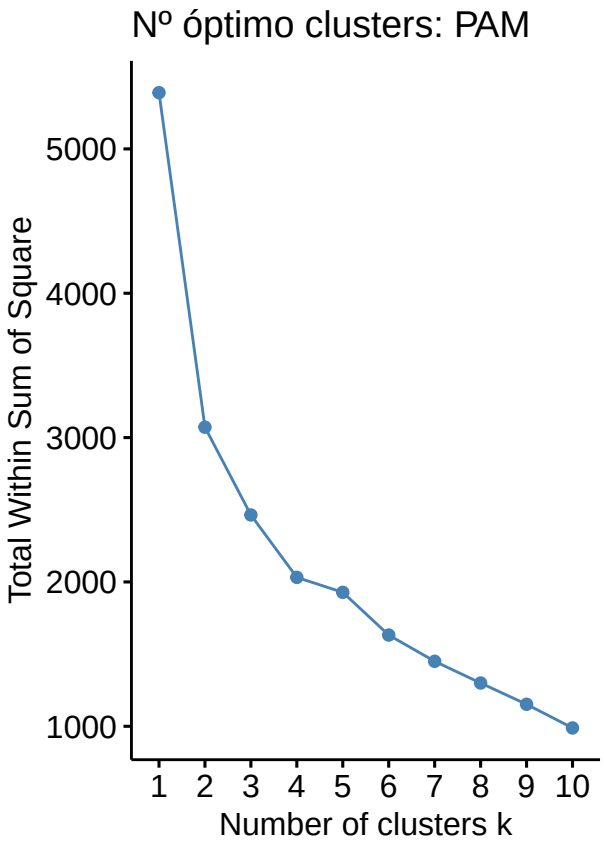

```r
k=3
clust3 <- kmeans(datos_elegidos, centers = k, nstart = 20)

p1 = fviz_cluster(object = list(data=datos_elegidos, cluster=clust3$cluster), stand = FALSE,
             ellipse.type = "convex", geom = "point", show.clust.cent = TRUE,
             labelsize = 8)  +
  labs(title = "K-MEDIAS + Proyeccion PCA",
       subtitle = "Dist euclidea, K=3") +
  theme_bw() +
  theme(legend.position = "bottom")

grid.arrange(p1,nrow = 1)
```



pam

```
p1 = fviz_nbclust(x = datos_elegidos, FUNcluster = pam, method = "silhouette",
             k.max = 10, verbose = FALSE) +
  labs(title = "Nº óptimo clusters: PAM")
p2 = fviz_nbclust(x = datos_elegidos, FUNcluster = pam, method = "wss",
             k.max = 10, verbose = FALSE) +
  labs(title = "Nº óptimo clusters: PAM")
grid.arrange(p1, p2, nrow = 1)
```

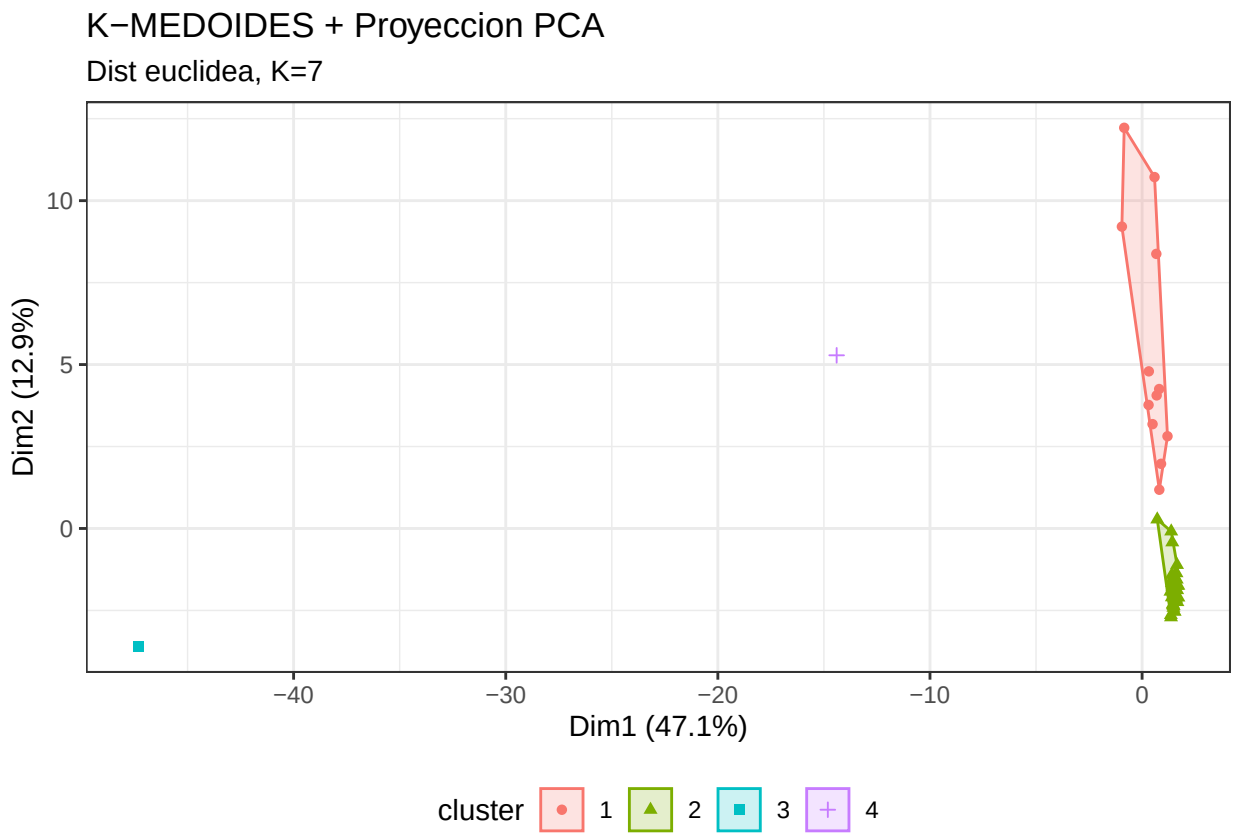





```
k=4
clust4 <- pam(datos_elegidos, k = k)
table(clust4$clustering)
```

```
##
##  1  2  3  4
## 12 38  1  1
```

```
p1 = fviz_cluster(object = list(data=datos_elegidos, cluster=clust4$clustering), stand = FALSE,
             ellipse.type = "convex", geom = "point", show.clust.cent = FALSE,
             labelsize = 8)  +
  labs(title = "K-MEDOIDES + Proyeccion PCA",
       subtitle = "Dist euclidea, K=7") +
  theme_bw() +
  theme(legend.position = "bottom")
grid.arrange(p1,nrow = 1)
```



Cuáles son esos dos estudiantes porsaqueros?

```
cluster_sizes <- table(clust4$clustering)
outlier_clusters <- names(cluster_sizes[cluster_sizes == 1])


outlier_ids <- which(clust4$clustering %in% outlier_clusters)


datos_elegidos[outlier_ids, ]
```

```
##    cred_mat1  cred_mat2 cred_mat3  cred_mat4 cred_sup_espec  cred_sup
## 29 -1.785002 -0.1247317  4.772668 -0.2652387      -0.138675 -0.138675
## 32 -1.785002  1.0260189  2.044395 -0.2652387      -0.138675 -0.138675
##    cred_mat_normal cred_ptes_acta cred_mat_sem_a cred_mat_sem_b cred_mat_anu
## 29       6.3207794      6.3207794      1.9260983      3.5526505   -0.4017367
## 32      -0.2136248     -0.2136248     -0.4831793     -0.6291645    1.2052102
##    cred_mat_total anyo_inicio_estudios cred_sup_1o cred_sup_2o cred_sup_3o
## 29      4.0531458           -0.5336537    1.785002    2.491429  -0.3485026
## 32      0.1177262           -1.5247250    1.785002    1.612824   1.5882141
##    cred_sup_4o  practicas cred_sup_tit   data_nac alta_universitat anyo_ingreso
## 29  -0.2426124 -0.2450451     1.529163 -0.3810125       -0.1408913   -0.5336537
## 32  -0.2426124 -0.2450451     1.590851 -0.7058101       -0.6642017   -1.5247250
##        nota10     nota14 curso_mas_bajo_2 curso_mas_alto_2 curso_mas_alto_3
## 29 -0.7470751 -0.5044405        4.9516897        -0.138675        4.0024031
## 32 -0.8894307 -0.8119763       -0.1980676        -0.138675       -0.2450451
```

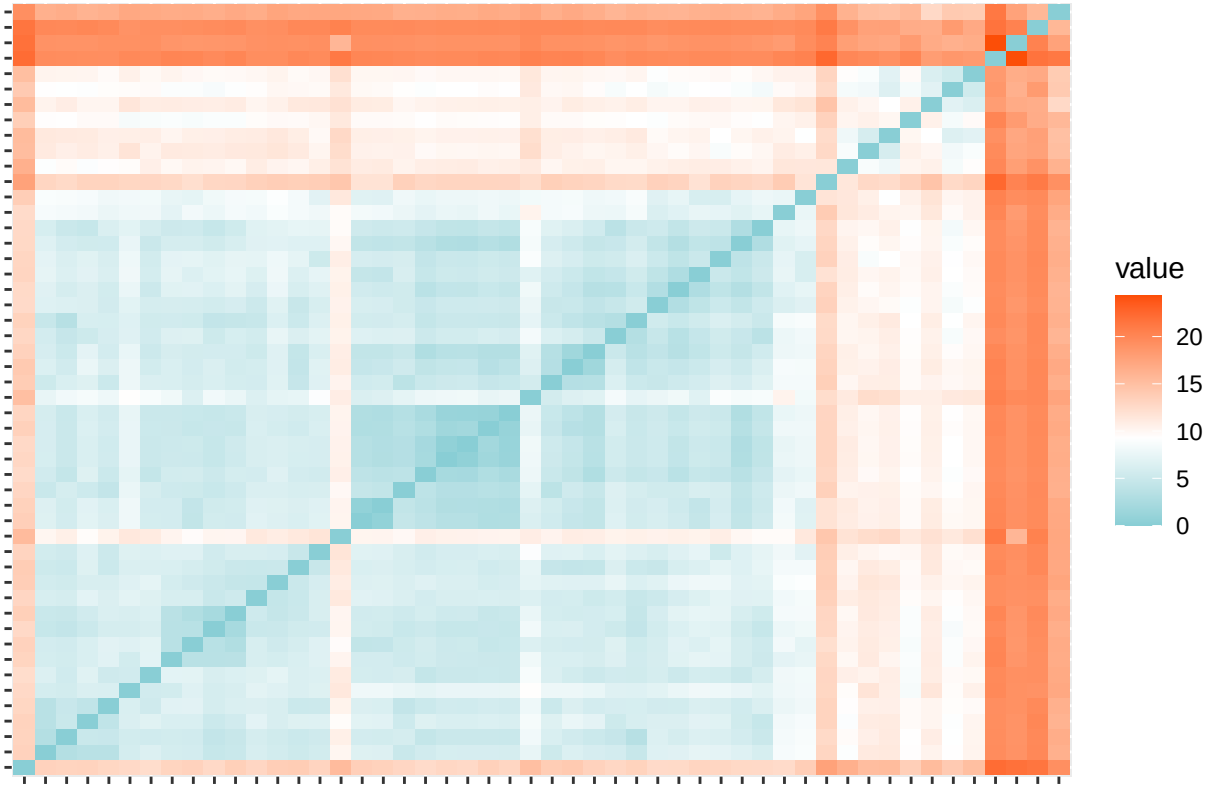
```
##    es_adaptado_1 nacionalitat_XXX    sexe_V prov_origen_ESPANYA
## 29     -0.138675        -0.138675 0.4222815          -0.4832354
## 32     -0.138675        -0.138675 0.4222815          -0.4832354
##    tipo_ingreso_BMA tipo_ingreso_NAP estudios_p_3 estudios_p_4 estudios_p_5
## 29        -1.631591         1.631591   -0.5129651   -0.7514555    1.2032465
## 32        -1.631591         1.631591    1.9119609   -0.7514555   -0.8151025
##    estudios_p_6 estudios_m_2 estudios_m_3 estudios_m_4 estudios_m_5
## 29    -0.138675    -0.138675    -0.357668   -0.7205767    0.9529531
## 32    -0.138675    -0.138675    -0.357668   -0.7205767    0.9529531
##    dedicacion_TP desplazado_1 discapacidad_1  becado_2 preferencia_seleccion_2
## 29    -0.2450451   -0.7205767      -0.138675  7.072428              -0.3230126
## 32    -0.2450451   -0.7205767      -0.138675 -0.138675              -0.3230126
##    preferencia_seleccion_3 preferencia_seleccion_Baja
## 29              -0.5424304                 -0.6602253
## 32              -0.5424304                 -0.6602253
##    preferencia_seleccion_Desconocido  mes_julio mes_septiembre mes_octubre
## 29                        -0.1980676 -0.8819589     -0.7829309  -0.3230126
## 32                        -0.1980676 -0.8819589     -0.7829309  -0.3230126
##    mes_diciembre       PC1       PC2       PC3
## 29    -0.1980676 -44.67668 -3.951647 -2.393982
## 32     4.9516897 -14.19229 11.819561  7.025431
```

```
datos_filtrados <- final_dataset[-outlier_ids, ]

ambas=ambas[-outlier_ids, ]
```

Filtrado

```
midist <- get_dist(datos_filtrados, stand = TRUE, method = "euclidean")
fviz_dist(midist, show_labels = TRUE, lab_size = 0.3,
          gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"))+labs(title = "Distancia eu
```

Distancia euclídea sin outliers



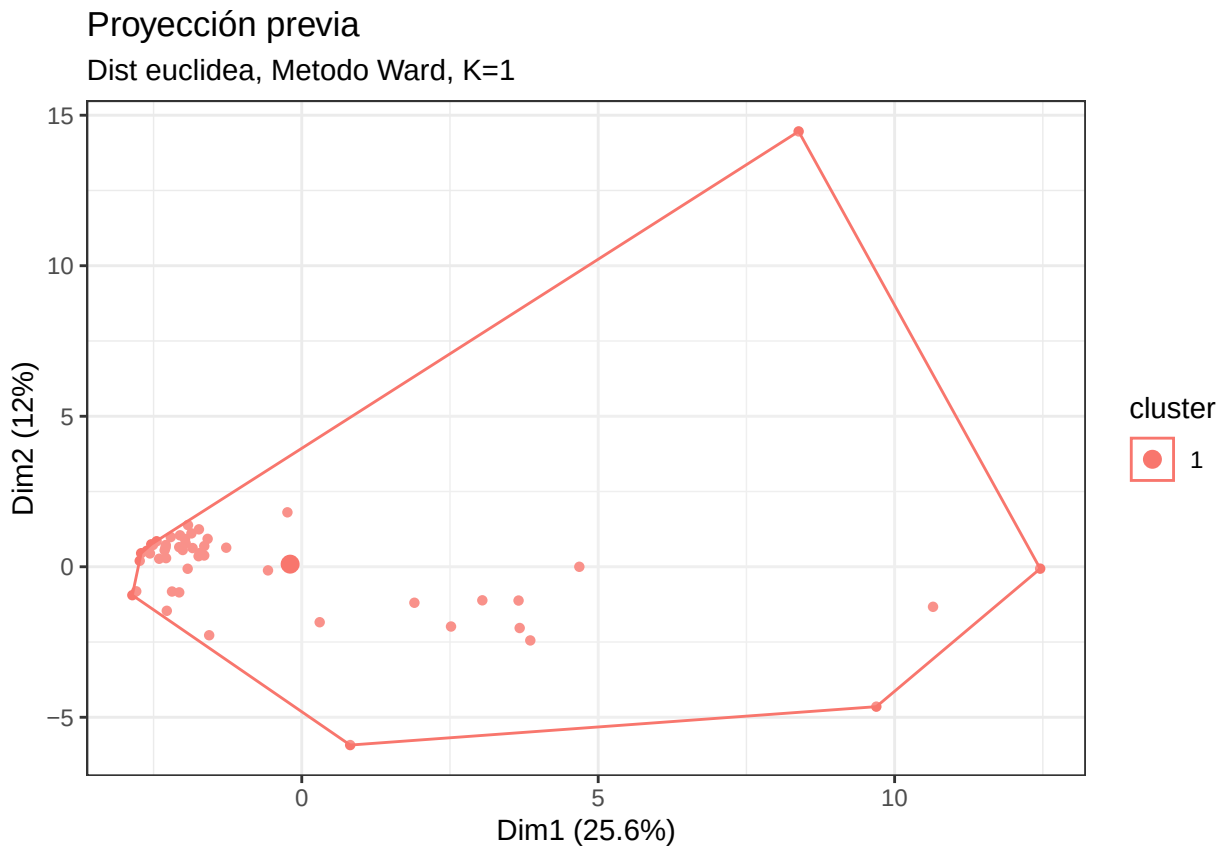

```
set.seed(10)
myN = c(7,25,37)  # m
myhopkins = NULL
myseed = sample(1:52, 2)
for (i in myN) {
  for (j in myseed) {
    tmp = get_clust_tendency(data = datos_filtrados, n = i, graph = FALSE, seed = j)
    myhopkins = c(myhopkins, tmp$hopkins_stat)
  }
}
summary(myhopkins)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7682  0.7767  0.7985  0.8042  0.8223  0.8603
```

```
datos_elegidos=datos_filtrados
```

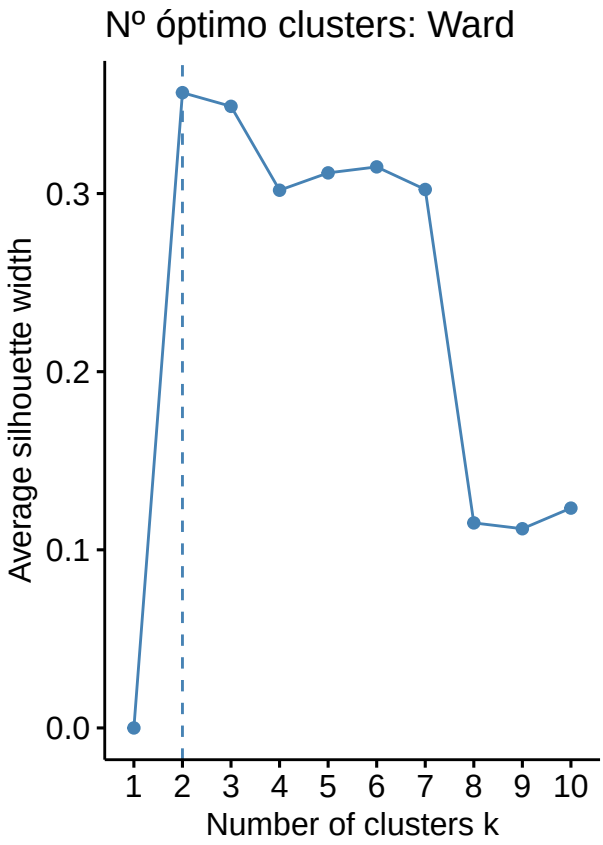
```
K=1

clust1 <- hclust(midist, method="ward.D2")
grupos1 <- cutree(clust1, k=K)

fviz_cluster(object = list(data=datos_elegidos, cluster=grupos1), stand = FALSE,
             ellipse.type = "convex", geom = "point", show.clust.cent = TRUE,
             labelsize = 8)  +
```

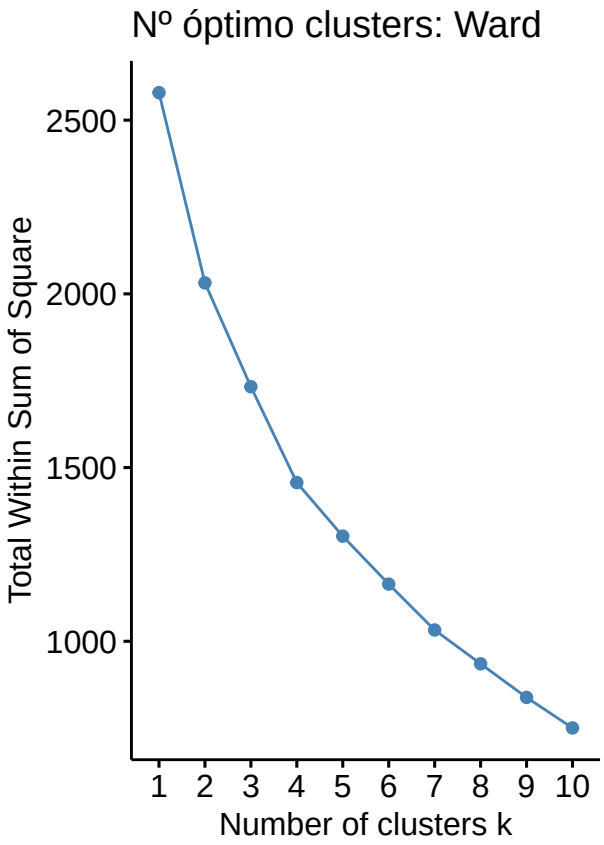
```
  labs(title = "Proyección previa",
       subtitle = "Dist euclidea, Metodo Ward, K=1") +
  theme_bw()
```



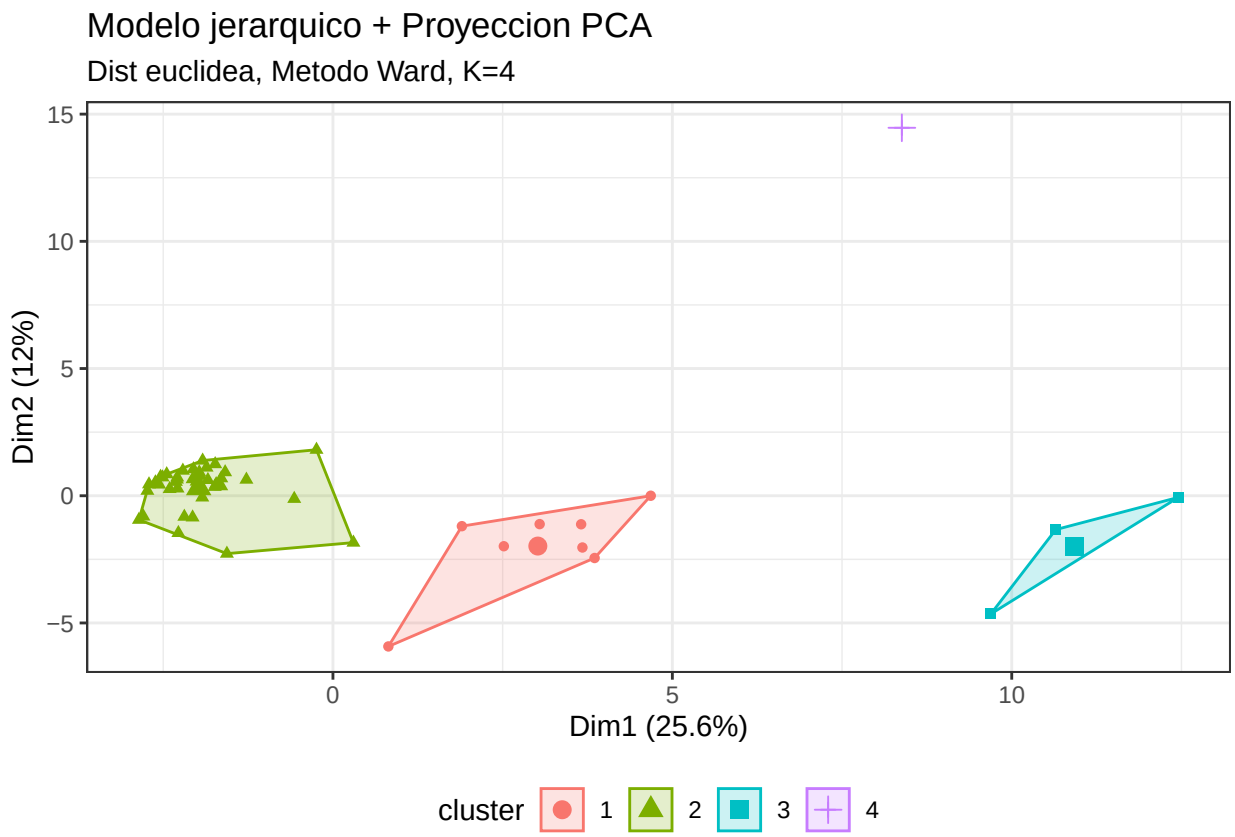
Método de Ward

```
p1 = fviz_nbclust(x =datos_elegidos, FUNcluster = hcut, method = "silhouette",
                  hc_method = "ward.D2", k.max = 10, verbose = FALSE,
                  hc_metric = "euclidean") + labs(title = "Nº óptimo clusters: Ward")
p2 = fviz_nbclust(x = datos_elegidos, FUNcluster = hcut, method = "wss",
                  hc_method = "ward.D2", k.max = 10, verbose = FALSE,
                  hc_metric = "euclidean") + labs(title = "Nº óptimo clusters: Ward")
grid.arrange(p1, p2, nrow = 1)
```



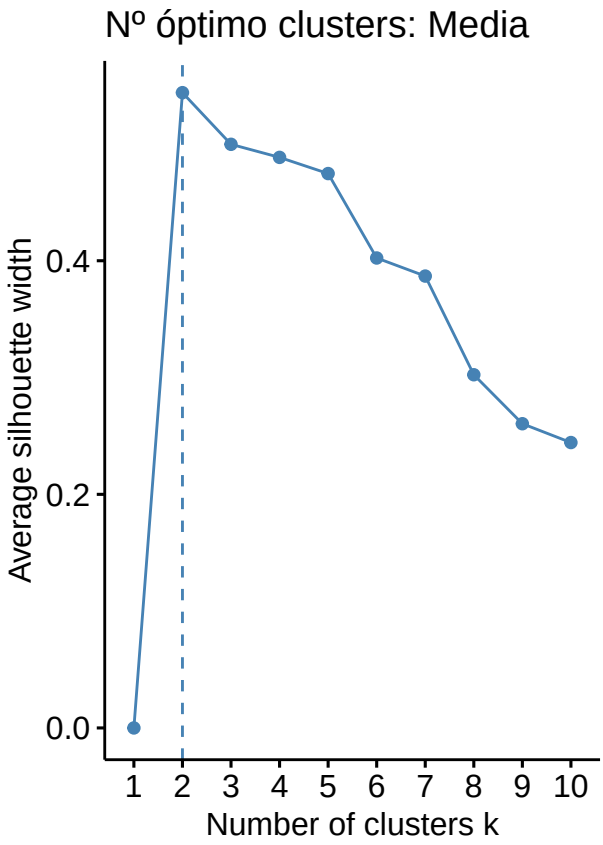


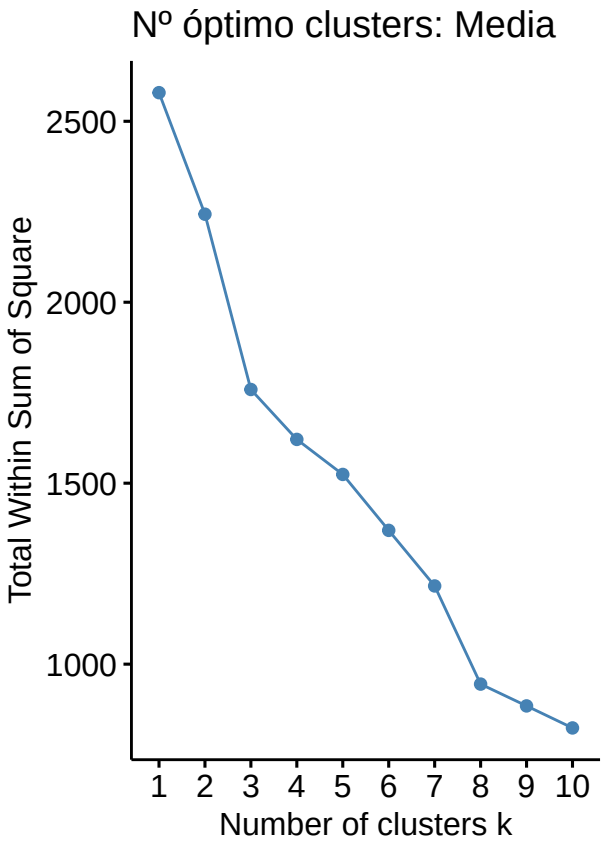

```
K=4

clust1 <- hclust(midist, method="ward.D2")
grupos1 <- cutree(clust1, k=K)

fviz_cluster(object = list(data=datos_elegidos, cluster=grupos1), stand = FALSE,
             ellipse.type = "convex", geom = "point", show.clust.cent = TRUE,
             labelsize = 8)  +
  labs(title = "Modelo jerarquico + Proyeccion PCA",
       subtitle = "Dist euclidea, Metodo Ward, K=4") +
  theme_bw() +
  theme(legend.position = "bottom")
```



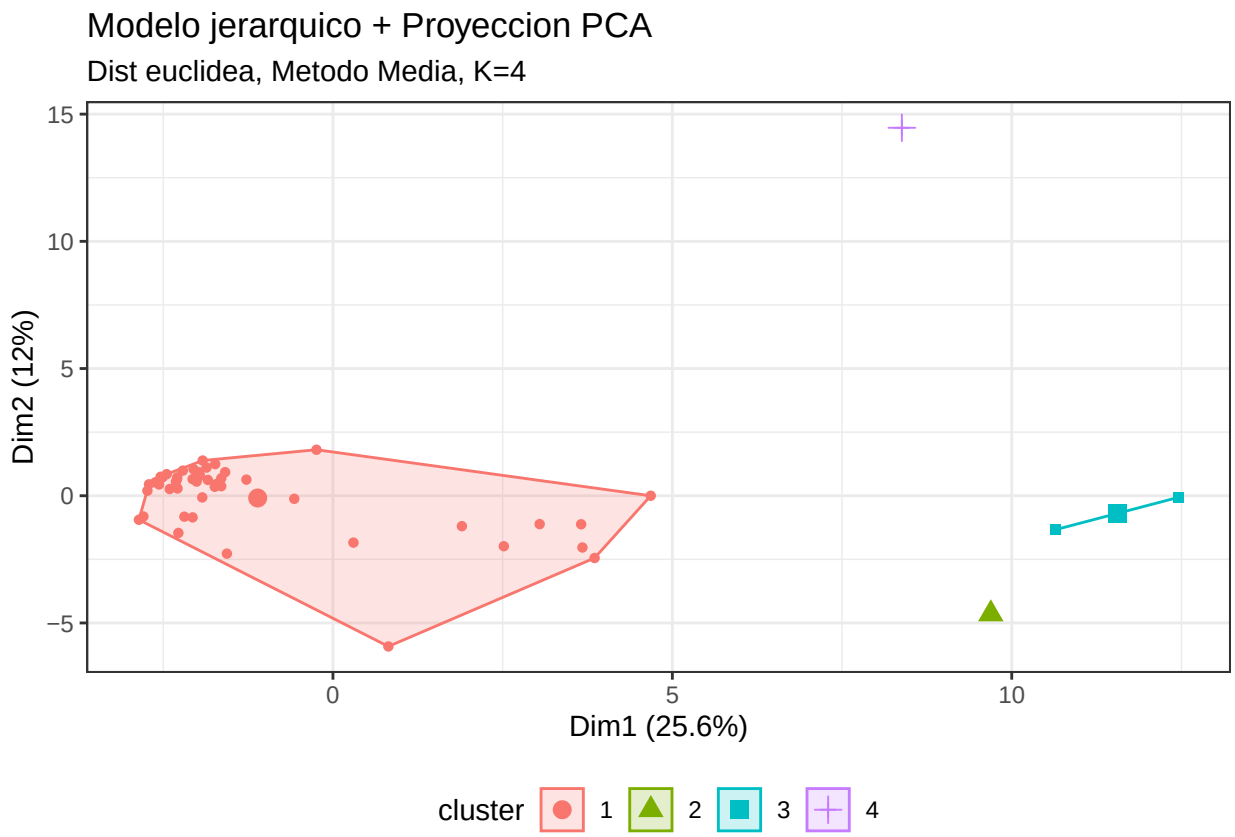
Método de la media

```
p1 = fviz_nbclust(x = datos_elegidos, FUNcluster = hcut, method = "silhouette",
                  hc_method = "average", k.max = 10, verbose = FALSE,
                  hc_metric = "euclidean") + labs(title = "Nº óptimo clusters: Media")
p2 = fviz_nbclust(x = datos_elegidos, FUNcluster = hcut, method = "wss",
                  hc_method = "average", k.max = 10, verbose = FALSE,
                  hc_metric = "euclidean") + labs(title = "Nº óptimo clusters: Media")
grid.arrange(p1, p2, nrow = 1)
```





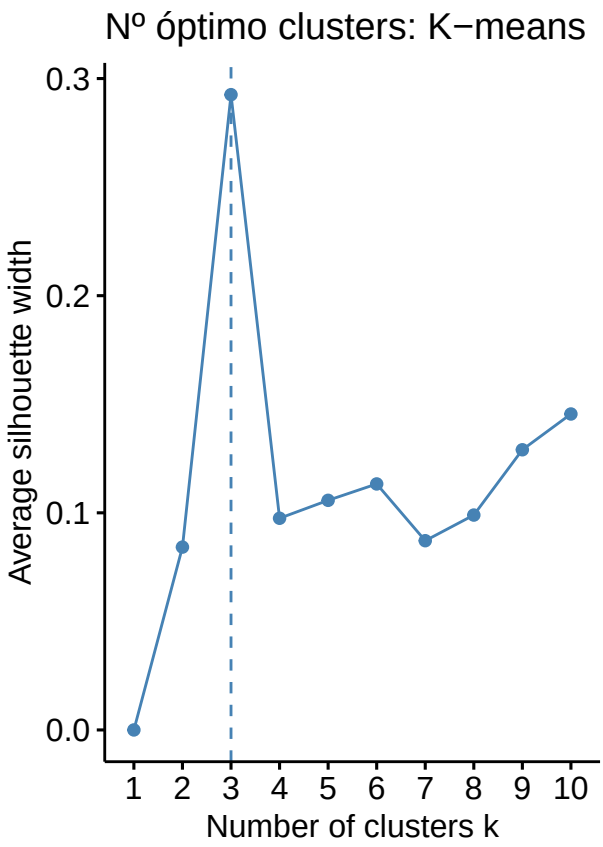

```r
K=4

clust1 <- hclust(midist, method="average")
grupos1 <- cutree(clust1, k=K)

fviz_cluster(object = list(data=datos_elegidos, cluster=grupos1), stand = FALSE,
             ellipse.type = "convex", geom = "point", show.clust.cent = TRUE,
             labelsize = 8)  +
  labs(title = "Modelo jerarquico + Proyeccion PCA",
       subtitle = "Dist euclidea, Metodo Media, K=4") +
  theme_bw() +
  theme(legend.position = "bottom")
```

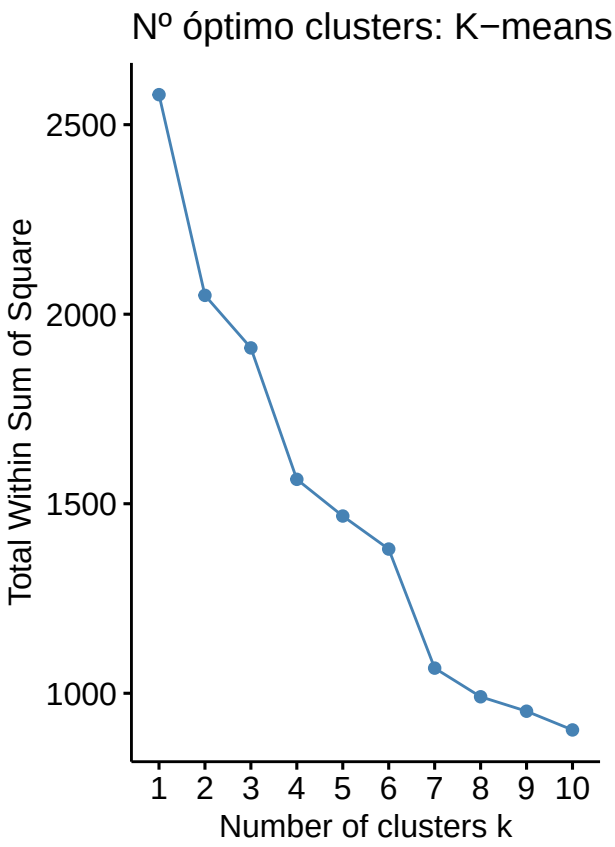


Partición

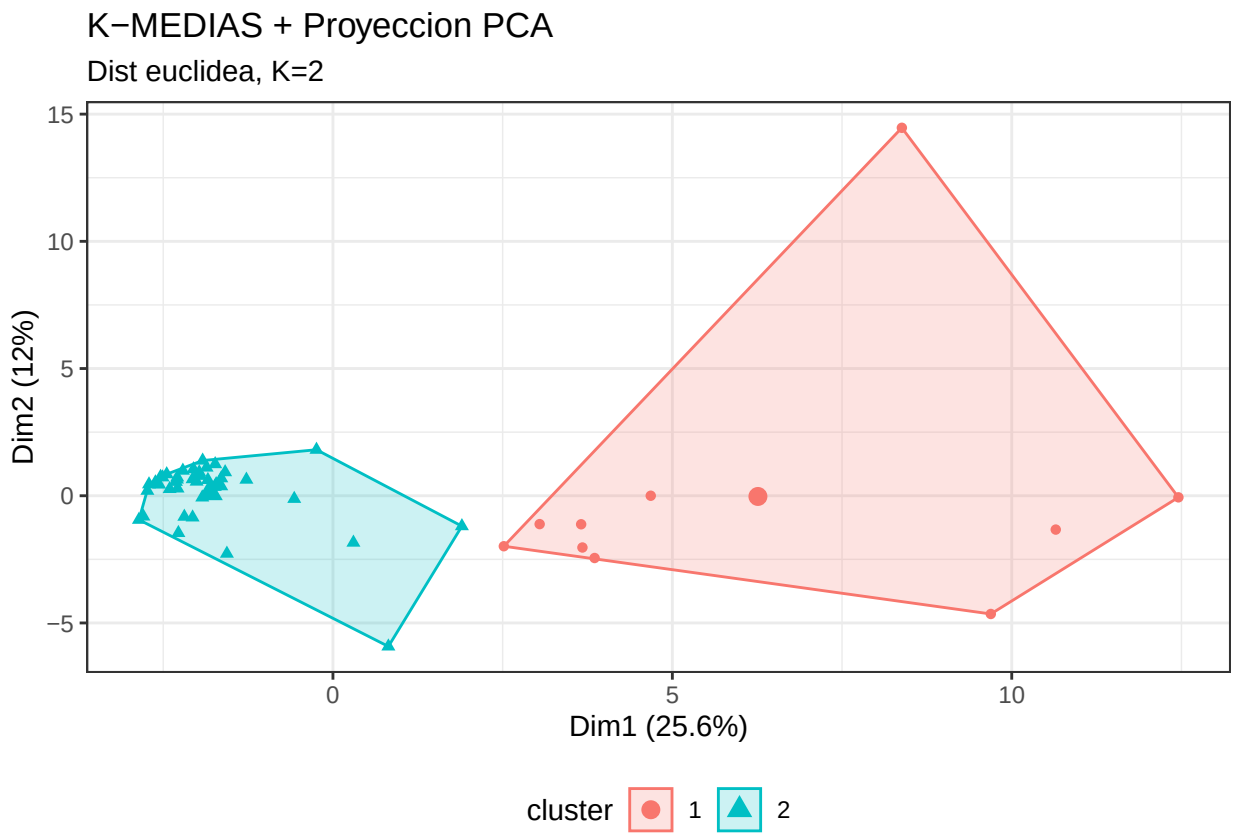
k-means

```
p1 = fviz_nbclust(x = datos_elegidos, FUNcluster = kmeans, method = "silhouette",
             k.max = 10, verbose = FALSE) +
  labs(title = "Nº óptimo clusters: K-means")
p2 = fviz_nbclust(x = datos_elegidos, FUNcluster = kmeans, method = "wss",
             k.max = 10, verbose = FALSE) +
  labs(title = "Nº óptimo clusters: K-means")
grid.arrange(p1, p2, nrow = 1)
```



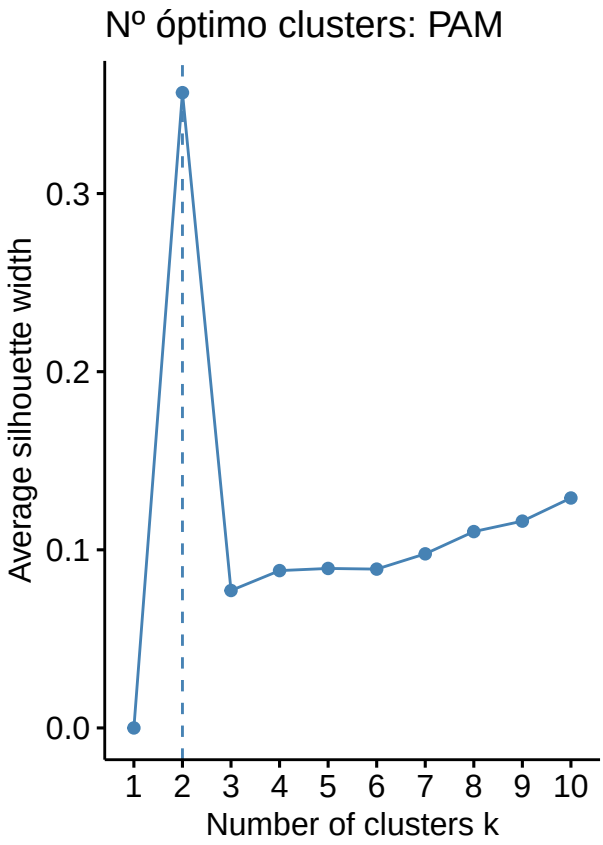


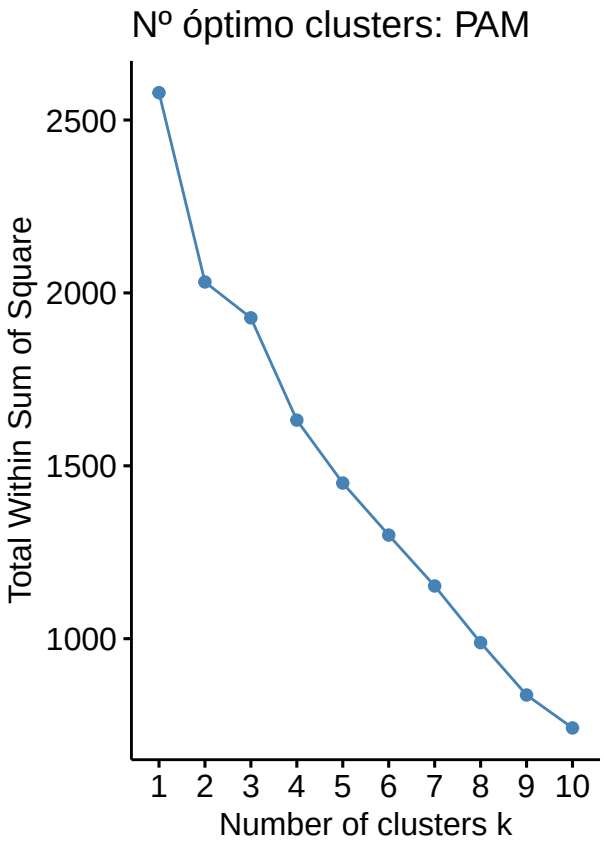

```
k=2
clust3 <- kmeans(datos_elegidos, centers = k, nstart = 20)

p1 = fviz_cluster(object = list(data=datos_elegidos, cluster=clust3$cluster), stand = FALSE,
             ellipse.type = "convex", geom = "point", show.clust.cent = TRUE,
             labelsize = 8)  +
  labs(title = "K-MEDIAS + Proyeccion PCA",
       subtitle = "Dist euclidea, K=2") +
  theme_bw() +
  theme(legend.position = "bottom")

grid.arrange(p1,nrow = 1)
```



pam

```
p1 = fviz_nbclust(x = datos_elegidos, FUNcluster = pam, method = "silhouette",
             k.max = 10, verbose = FALSE) +
  labs(title = "Nº óptimo clusters: PAM")
p2 = fviz_nbclust(x = datos_elegidos, FUNcluster = pam, method = "wss",
             k.max = 10, verbose = FALSE) +
  labs(title = "Nº óptimo clusters: PAM")
grid.arrange(p1, p2, nrow = 1)
```

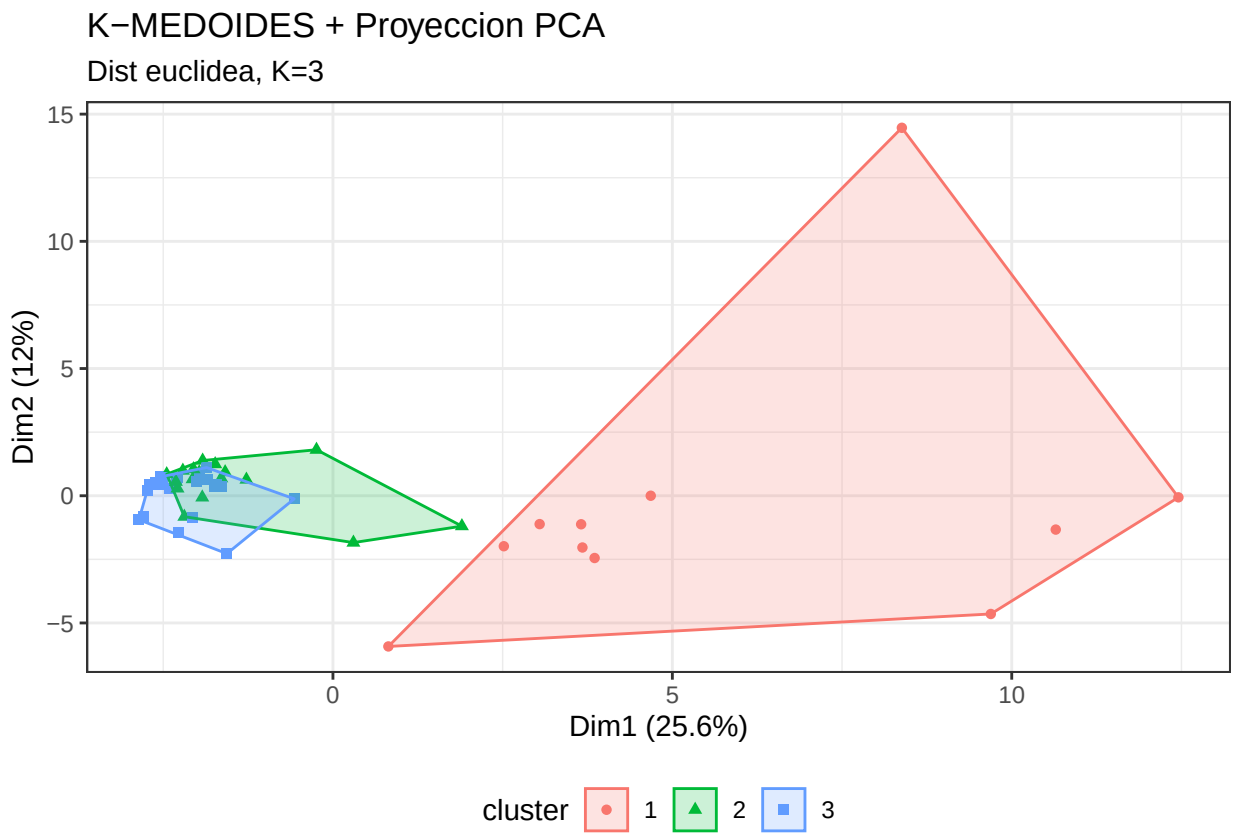





```
k=3
clust4 <- pam(datos_elegidos, k = k)
table(clust4$clustering)
```

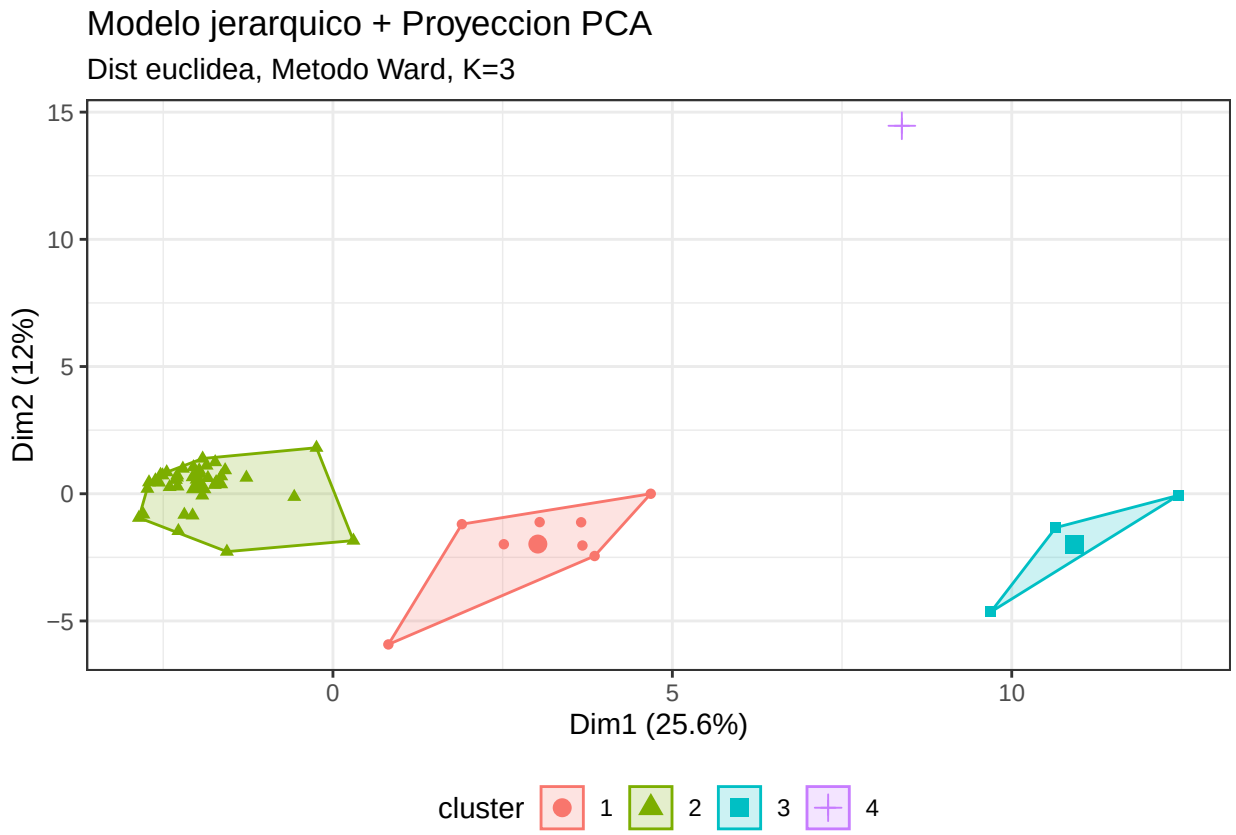
```
##
##  1  2  3
## 11 17 22
```

```
p1 = fviz_cluster(object = list(data=datos_elegidos, cluster=clust4$clustering), stand = FALSE,
             ellipse.type = "convex", geom = "point", show.clust.cent = FALSE,
             labelsize = 8)  +
  labs(title = "K-MEDOIDES + Proyeccion PCA",
       subtitle = "Dist euclidea, K=3") +
  theme_bw() +
  theme(legend.position = "bottom")
grid.arrange(p1,nrow = 1)
```



Final: Ward

```
clust1 <- hclust(midist, method="ward.D2")
grupos1 <- cutree(clust1, k=4)

fviz_cluster(object = list(data=datos_elegidos, cluster=grupos1), stand = FALSE,
             ellipse.type = "convex", geom = "point", show.clust.cent = TRUE,
             labelsize = 8)  +
  labs(title = "Modelo jerarquico + Proyeccion PCA",
       subtitle = "Dist euclidea, Metodo Ward, K=3") +
  theme_bw() +
  theme(legend.position = "bottom")
```




```r
datos_filtrados$cluster <- as.factor(cutree(clust1, k = 4))
resumen_centros <- aggregate(. ~ cluster, data = datos_filtrados, FUN = mean, na.rm = TRUE)
#write.csv(resumen_centros, "resumen_centros_clusters.csv", row.names = FALSE)
print(resumen_centros)
```

```
##   cluster  cred_mat1  cred_mat2  cred_mat3  cred_mat4 cred_sup_espec  cred_sup
## 1       1 -1.4585890  1.8498518  0.6567394 -0.1215677      -0.138675 -0.138675
## 2       2  0.5889139 -0.4385728 -0.4016423 -0.2652387      -0.138675 -0.138675
## 3       3 -1.7850025  0.1193669  0.6645792  3.9491095      -0.138675 -0.138675
## 4       4 -1.7850025  0.6075642  1.1976900 -0.2652387       7.072428  7.072428
##   cred_mat_normal cred_ptes_acta cred_mat_sem_a cred_mat_sem_b cred_mat_anu
## 1      -0.2136248     -0.2136248     -0.3971337      -1.274886    1.6069469
## 2      -0.1620374     -0.1620374      0.2051857       0.354792   -0.4017367
## 3       0.6576291      0.6576291     -1.8599094      -1.285135    0.6695612
## 4      -0.2136248     -0.2136248     -0.4831793      -2.351088   -0.4017367
##   cred_mat_total anyo_inicio_estudios cred_sup_1o cred_sup_2o cred_sup_3o
## 1    -0.07357889           -0.8433635   1.4585890   0.5844566 -0.05453666
## 2     0.11772623            0.4574175  -0.5889139  -0.4638788 -0.34850258
## 3    -1.77710541           -3.0113319   1.7850025   2.3050578  3.61715530
## 4    -2.72452123            0.4574175   1.7850025   1.9323163  1.58821406
##   cred_sup_4o  practicas cred_sup_tit   data_nac alta_universitat anyo_ingreso
## 1  -0.2426124 -0.2450451    0.8076310 -0.4216122       -0.3044258   -0.8433635
## 2  -0.2426124 -0.2450451   -0.5241736  0.3540556        0.3686478    0.4574175
## 3   3.9626691  4.0024031    2.8803112 -2.1132662       -1.6236041   -3.0113319
## 4  -0.2426124 -0.2450451    1.6966023 -2.6545955       -5.8973057    0.4574175
```

```
##        nota10      nota14 curso_mas_bajo_2 curso_mas_alto_2 curso_mas_alto_3
## 1  0.24066261  0.07952094       -0.1980676        -0.138675        0.2858859
## 2  0.09876763  0.08822931       -0.1980676        -0.138675       -0.2450451
## 3 -0.34140057 -0.89837171        1.5185182         2.265026        1.1707710
## 4 -3.01776326  0.02265059       -0.1980676        -0.138675       -0.2450451
##   es_adaptado_1 nacionalitat_XXX     sexe_V prov_origen_ESPANYA
## 1     -0.138675        -0.138675 -0.6070297          -0.1691324
## 2     -0.138675        -0.138675  0.1333521           0.0457802
## 3     -0.138675         2.265026 -0.4926618          -0.4832354
## 4      7.072428        -0.138675  0.4222815           2.0295888
##   tipo_ingreso_BMA tipo_ingreso_NAP estudios_p_3 estudios_p_4 estudios_p_5
## 1       -1.6315906        1.6315906   0.09326638   0.27685201   -0.3105152
## 2        0.5423570       -0.5423570  -0.06626822  -0.04787666    0.1409576
## 3       -1.6315906        1.6315906   0.29534355   0.61962117   -0.8151025
## 4        0.6011123       -0.6011123  -0.51296511  -0.75145546   -0.8151025
##   estudios_p_6 estudios_m_2 estudios_m_3 estudios_m_4 estudios_m_5
## 1    -0.138675  -0.13867505   0.02980567   0.32025631   -0.2858859
## 2    -0.138675   0.05109081  -0.11294780  -0.06320848    0.1183668
## 3    -0.138675  -0.13867505   0.67559517   0.66720064   -1.0291894
## 4     7.072428  -0.13867505   2.74212158  -0.72057669   -1.0291894
##   dedicacion_TP desplazado_1 discapacidad_1  becado_2 preferencia_seleccion_2
## 1   -0.24504509   0.06004806    -0.13867505 -0.138675              0.51682019
## 2   -0.02149518   0.04635289     0.05109081 -0.138675             -0.05780226
## 3    1.17077098  -0.72057669    -0.13867505 -0.138675             -0.32301262
## 4   -0.24504509   1.36108931    -0.13867505 -0.138675             -0.32301262
##   preferencia_seleccion_3 preferencia_seleccion_Baja
## 1              -0.5424304                -0.6602253
## 2               0.1379867                 0.2432409
## 3               0.2410802                -0.6602253
## 4              -0.5424304                -0.6602253
##   preferencia_seleccion_Desconocido  mes_julio mes_septiembre mes_octubre
## 1                        -0.1980676 -0.8819589      0.7437843  0.09690379
## 2                        -0.1980676  0.3249322     -0.1936724 -0.05780226
## 3                         1.5185182 -0.8819589      0.5741493  0.79676446
## 4                         4.9516897 -0.8819589      1.2526894 -0.32301262
##   mes_diciembre       PC1        PC2        PC3
## 1     0.4456521 1.1632041 -0.3823566  0.7416181
## 2    -0.1980676 1.2204499 -0.1673215 -0.2033983
## 3    -0.1980676 0.7917431  0.3838783 -0.6864949
## 4    -0.1980676 0.8110127  0.3975199 -0.7757741
```

```
dim(resumen_centros)
```

```
## [1]  4 57
```