PruebaExcel

Juan Francisco Pallardó Latorre

2024-05-16

```
dataEx = read_excel("rank_prim.xlsx")
```

He eliminado variables como Nombre, Identificador, Latitud, Longitud, barrio, Direccion, NotaReviews, NumReviews y Mixto.

```
eliminar_v = c(1,2,4,5,55,68:70,73)
prueba = dataEx[, -eliminar_v]
```

Elimino las variables constantes: Comedor, Webcam, Internado mixto, Internado especial, Internado femenino, Internado masculino, Chino, Italiano, Portugués, Árabe, Japonés, Griego, Inglés a distancia, Ruso, Euskera, ADU, EP, PEV, PIP, ZC, PPEV, PPEC y PEPLI.

```
eliminar_v = c(6,19:23, 29:38, 40, 44,45,47:50)
prueba = prueba[, -eliminar_v]
```

Arreglo de las variables Inst+Gimnasio, Gimnas+Piscina y Inst+Gimnas+Piscina.

```
for (i in 1:length(prueba$`Inst + Gimnasio`)) {
  if (prueba$`Inst + Gimnasio`[i] == 1) {
    prueba$`Instalaciones deportivas`[i] <- 1
  }
}

for (i in 1:length(prueba$`Inst + Gimnasio`)) {
  if (prueba$`Gimnas + Piscina`[i] == 1) {
    prueba$Piscina[i] <- 1
    prueba$`Inst + Gimnasio`[i] <- 1
  }
}
prueba = subset(prueba, select = - `Gimnas + Piscina`)

for (i in 1:length(prueba$Piscina)) {
  if (prueba$`Inst + Gimnas + Piscina`[i] == 1) {
    prueba$Piscina[i] <- 1
    prueba$`Instalaciones deportivas`[i] <- 1
    prueba$`Inst + Gimnasio`[i] = 1
  }
}
prueba = subset(prueba, select = - `Inst + Gimnas + Piscina`)
names(prueba)[9] = "Gimnasio"
```

Ahora creamos las variables dummy de las variables Tipo, index_gl_1, Religion, Precio y Horario. (One-Hot encoding)

```
var_dummy = as.data.frame(model.matrix(~ Tipo -1, data=prueba))
prueba = cbind(prueba[-which(names(prueba) == "Tipo")], var_dummy)

var_dummy = as.data.frame(model.matrix(~ index_gl_1 - 1, data=prueba))
prueba = cbind(prueba[-which(names(prueba) == "index_gl_1")], var_dummy)


var_dummy = as.data.frame(model.matrix(~ Religion - 1, data=prueba))
prueba = cbind(prueba[-which(names(prueba) == "Religion")], var_dummy)

var_dummy = as.data.frame(model.matrix(~ Precio - 1, data=prueba))
prueba = cbind(prueba[-which(names(prueba) == "Precio")], var_dummy)

var_dummy = as.data.frame(model.matrix(~ Horario - 1, data=prueba))
prueba = cbind(prueba[-which(names(prueba) == "Horario")], var_dummy)
```

Elimino algunas variables dummy que son redundantes como: TipoPúblico, ReligiosoLaico y PrecioEntre 100 y 300€. Además de la variable Huerto Escolar que se repite su información en Huerto.

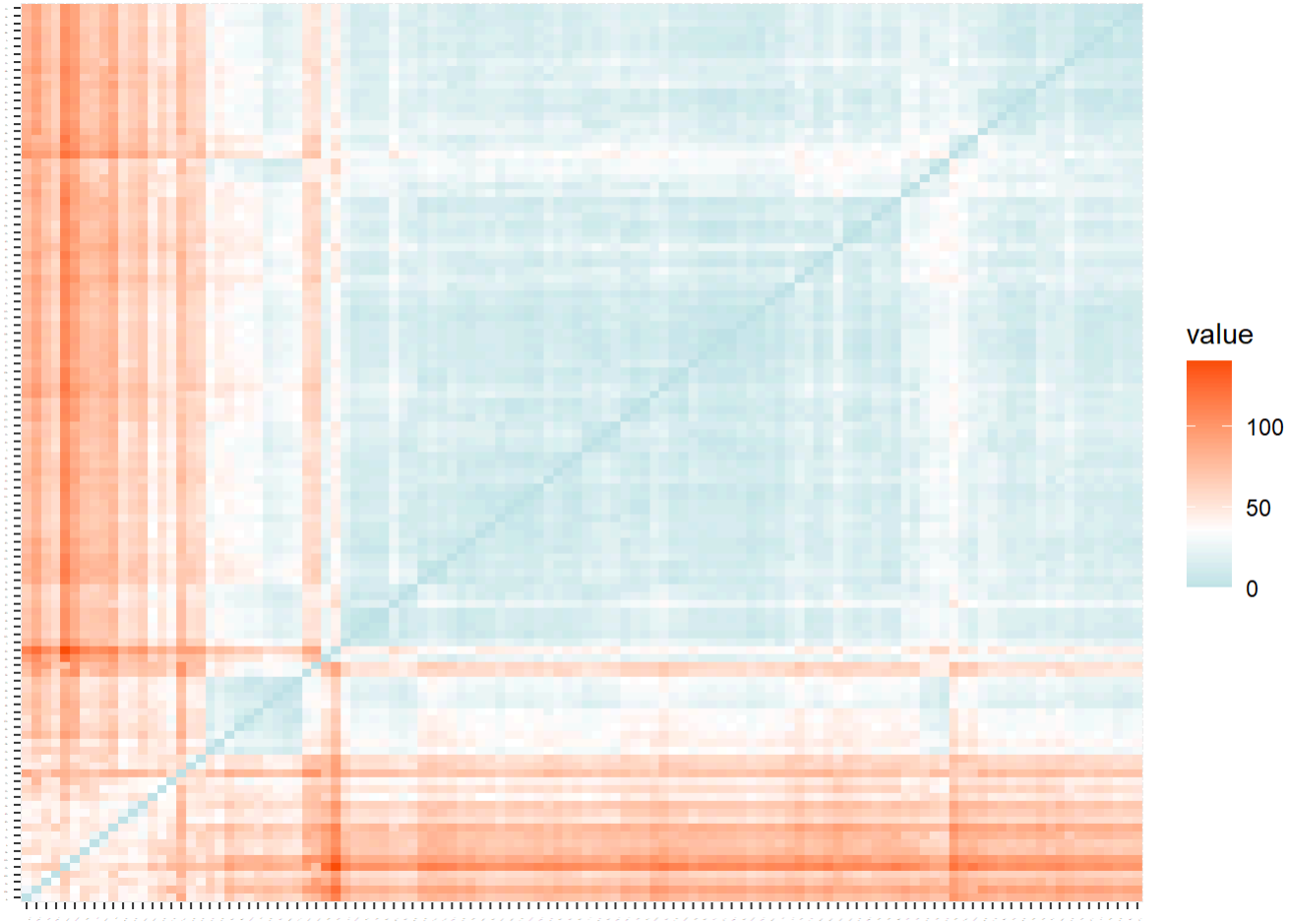
```
prueba_bin = prueba[,-c(12, 59, 63, 65)]
```

Quito las variables socioeconomicas: paro, renta_media, riesgo_pobreza, index_equi, index_soci, index_glob, index_gl1NoVulnerable, index_gl_1Pot.Vulnerable, PrecioGratuito o < 100€.

```
eliminar_v = c(27:32, 58:60,62)
prueba_bin = prueba_bin[,-eliminar_v]
```

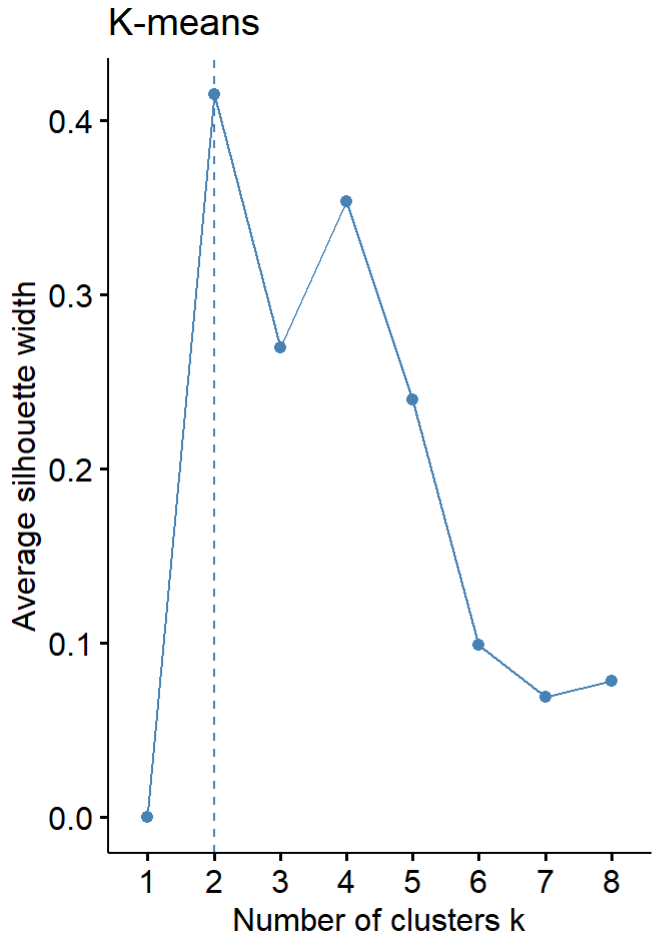
Hago el escalado y centrado de los datos, usare la distancia de Manhattan (ver diferencias con Euclídea). Printeo una matriz de todas las instancias y su distancia.

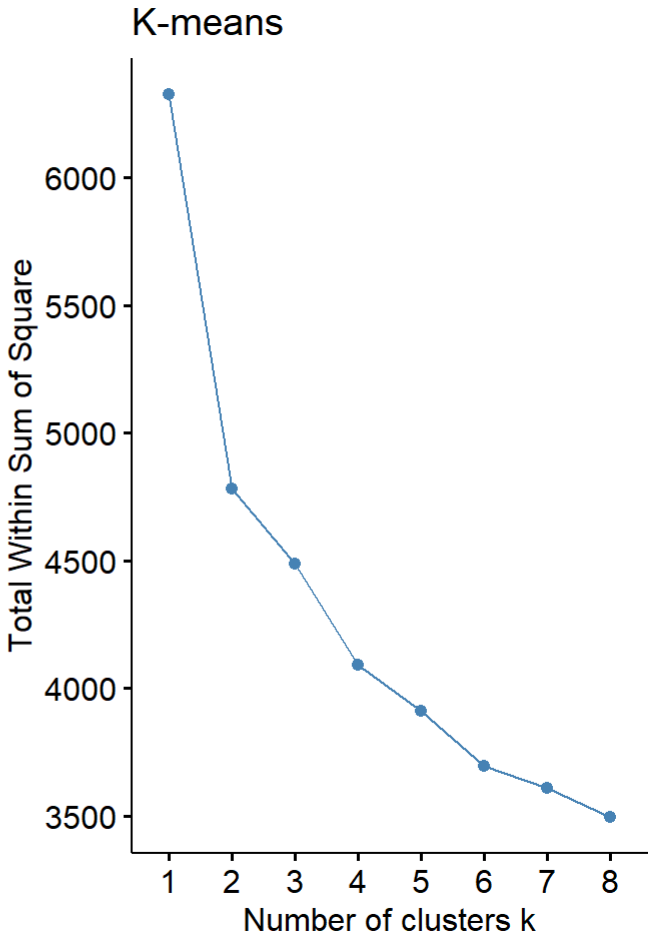
```
schools_cluster = scale(prueba_bin,center=TRUE, scale=TRUE)
midist <- get_dist(schools_cluster, stand = FALSE, method = "manhattan")
fviz_dist(midist, show_labels = TRUE, lab_size = 0.3,
          gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"))
```



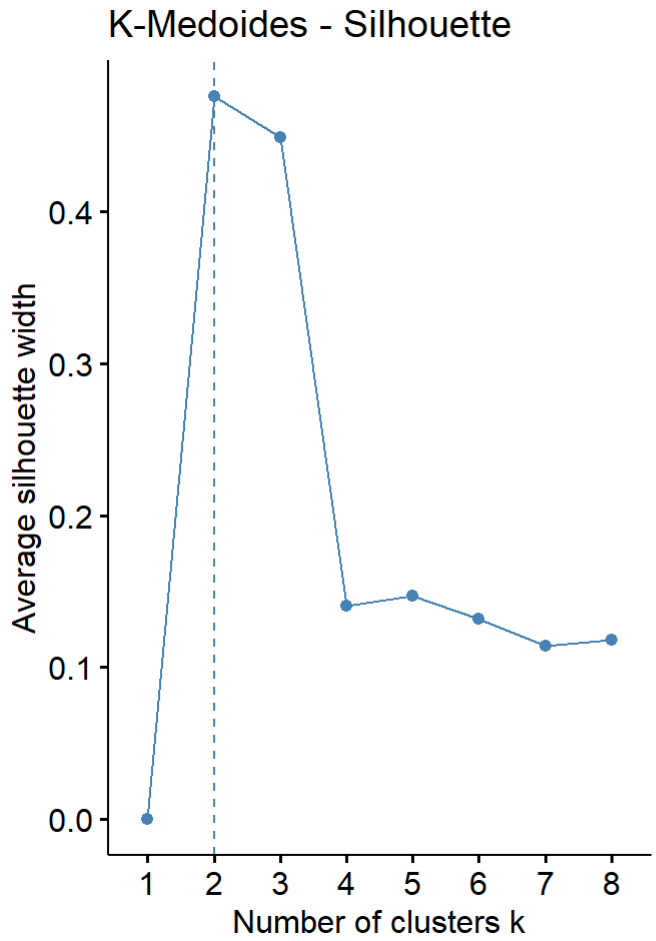
En el gráfico obtenido a partir de las distancias podemos observar posibles agrupaciones entre centros que serían los cuadrados azules que se forman a través de la diagonal principal de la matriz. El siguiente paso será realizar un método de partición, en concreto realizaremos el algoritmo **k-medias**. Sin embargo, previo a realizarlo hay que determinar el número óptimo de clusters. Para ello nos basaremos en el coeficiente de Silhouette (A mayor más relación entre clusters) y en la Suma de Cuadrados Residual (A menor mejor).

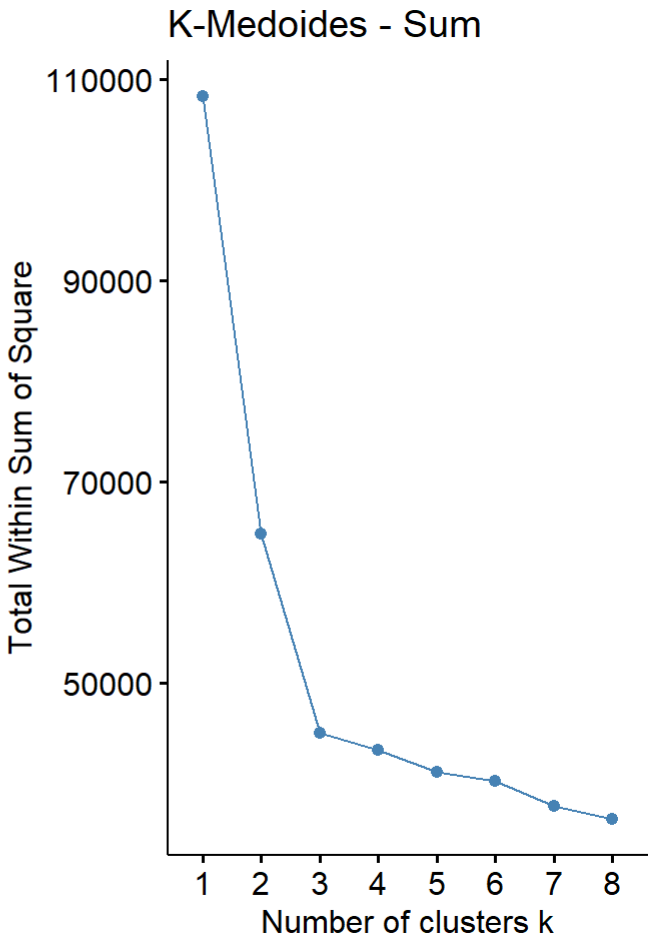
```
p1 = fviz_nbclust(x = schools_cluster, FUNcluster = kmeans, method = "silhouette",
             k.max = 8, verbose = FALSE) +
  labs(title = "K-means")
p2 = fviz_nbclust(x = schools_cluster, FUNcluster = kmeans, method = "wss",
             k.max = 8, verbose = FALSE) +
  labs(title = "K-means")
p3 = fviz_nbclust(x = schools_cluster, FUNcluster = pam, method = "silhouette",
             k.max = 8, verbose = FALSE, diss = midist) +
  labs(title = "K-Medoides - Silhouette")

p4 = fviz_nbclust(x = schools_cluster, FUNcluster = pam, method = "wss",
             k.max = 8, verbose = FALSE, diss = midist) +
  labs(title = "K-Medoides - Sum")

grid.arrange(p1,p2,nrow = 1)
```






```
grid.arrange(p3,p4,nrow = 1)
```

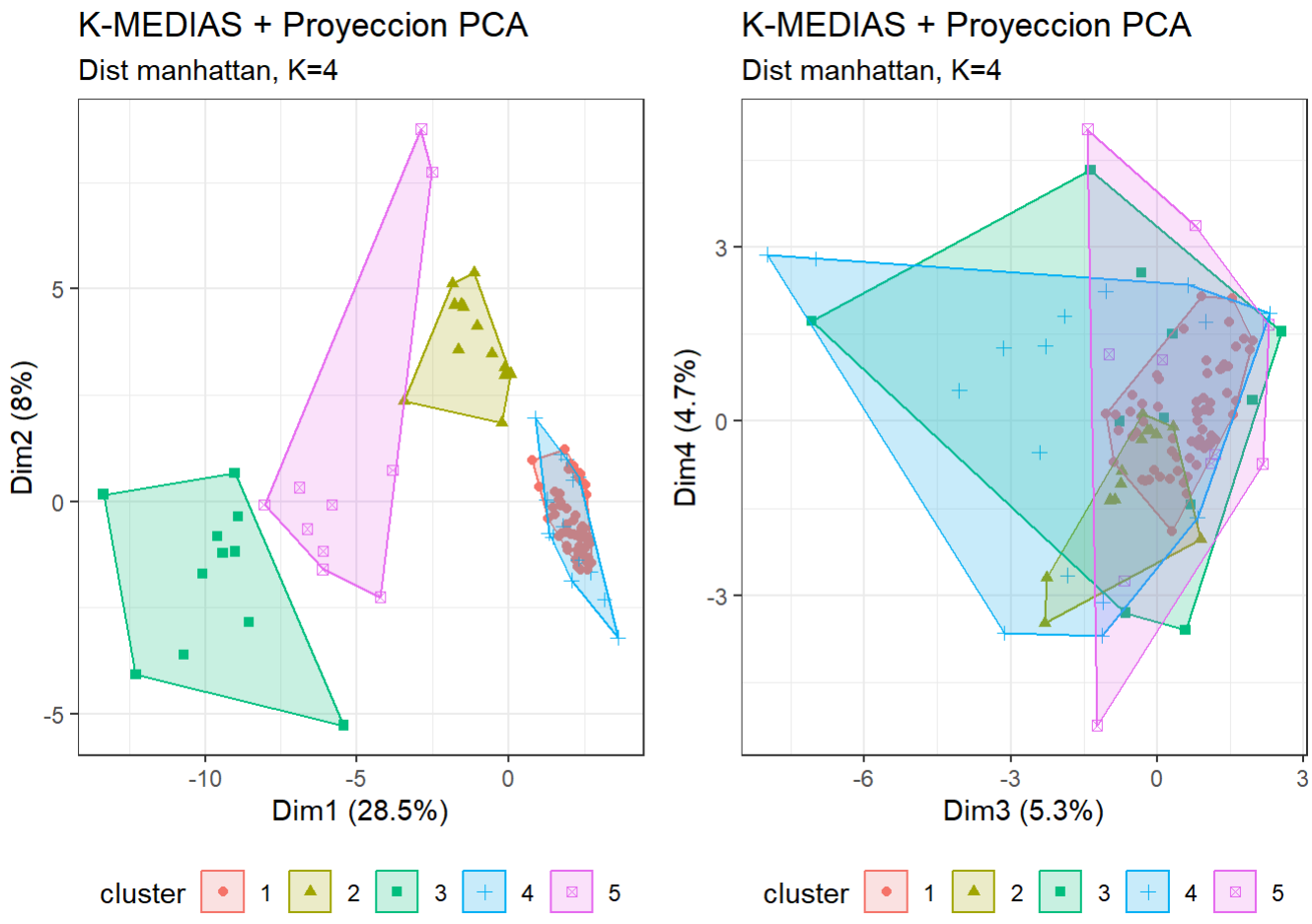




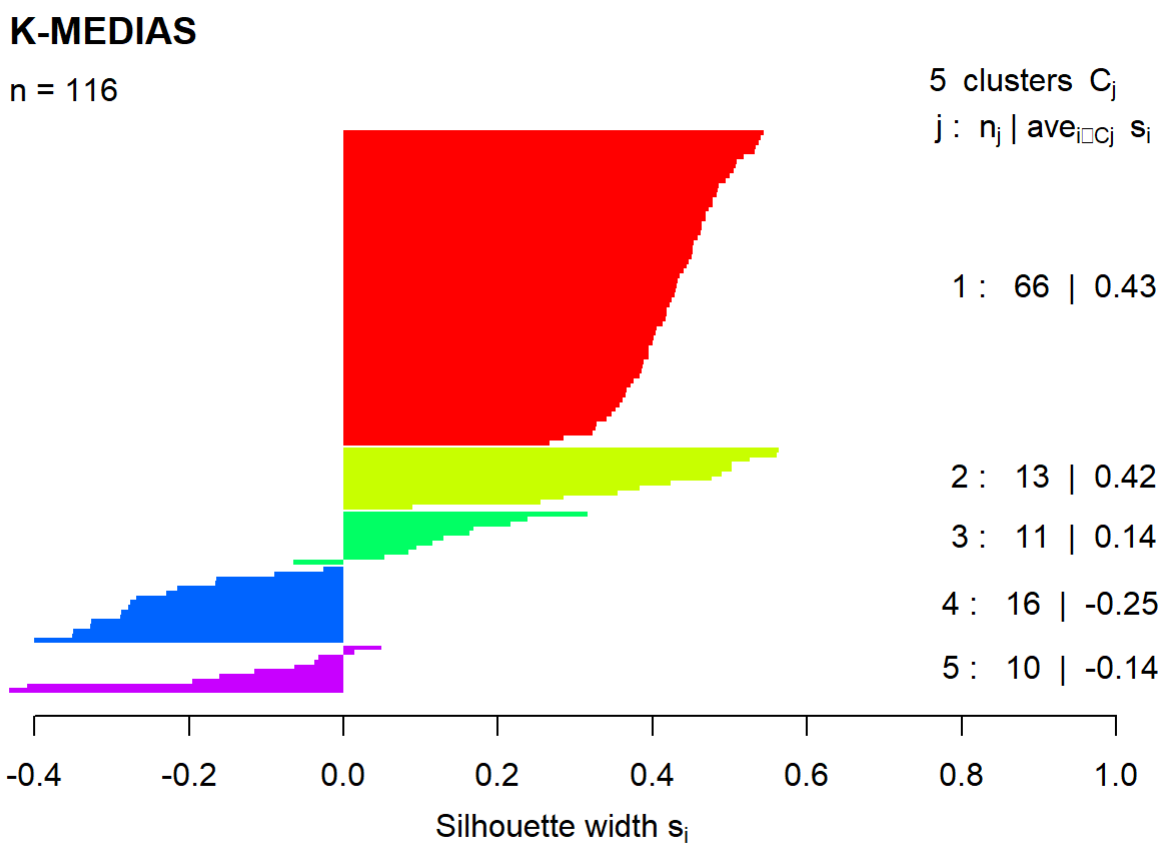
```
set.seed(100)
clust3 <- kmeans(midist, centers = 5, nstart = 20)
table(clust3$cluster)
```

```
##
##  1  2  3  4  5
## 66 13 11 16 10
```

```
p1 = fviz_cluster(object = list(data=schools_cluster, cluster=clust3$cluster), stand = FALSE,
             ellipse.type = "convex", geom = "point", show.clust.cent = FALSE,
             labelsize = 8)  +
  labs(title = "K-MEDIAS + Proyeccion PCA",
       subtitle = "Dist manhattan, K=4") +
  theme_bw() +
  theme(legend.position = "bottom")
p2 = fviz_cluster(object = list(data=schools_cluster, cluster=clust3$cluster), stand = FALSE,
             ellipse.type = "convex", geom = "point", show.clust.cent = FALSE,
             labelsize = 8, axes = 3:4)  +
  labs(title = "K-MEDIAS + Proyeccion PCA",
       subtitle = "Dist manhattan, K=4") +
  theme_bw() +
  theme(legend.position = "bottom")
grid.arrange(p1, p2, nrow = 1)
```

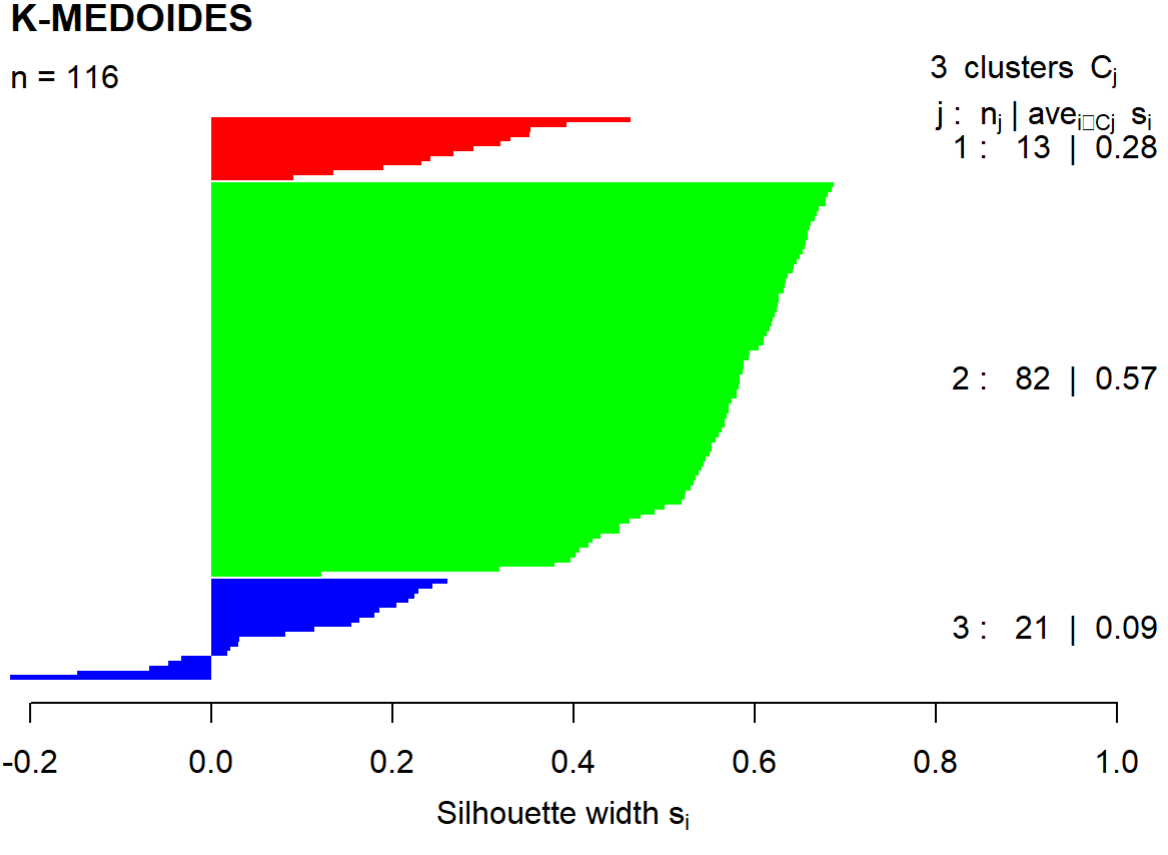



```
plot(silhouette(clust3$cluster, midist), col=rainbow(5), border=NA, main = "K-MEDIAS")
```



Average silhouette width : 0.26

```
clust4 <- pam(schools_cluster, k = 3)
plot(silhouette(clust4$clustering, midist), col=rainbow(3), border=NA, main = "K-MEDOIDES")
```



Average silhouette width : 0.45

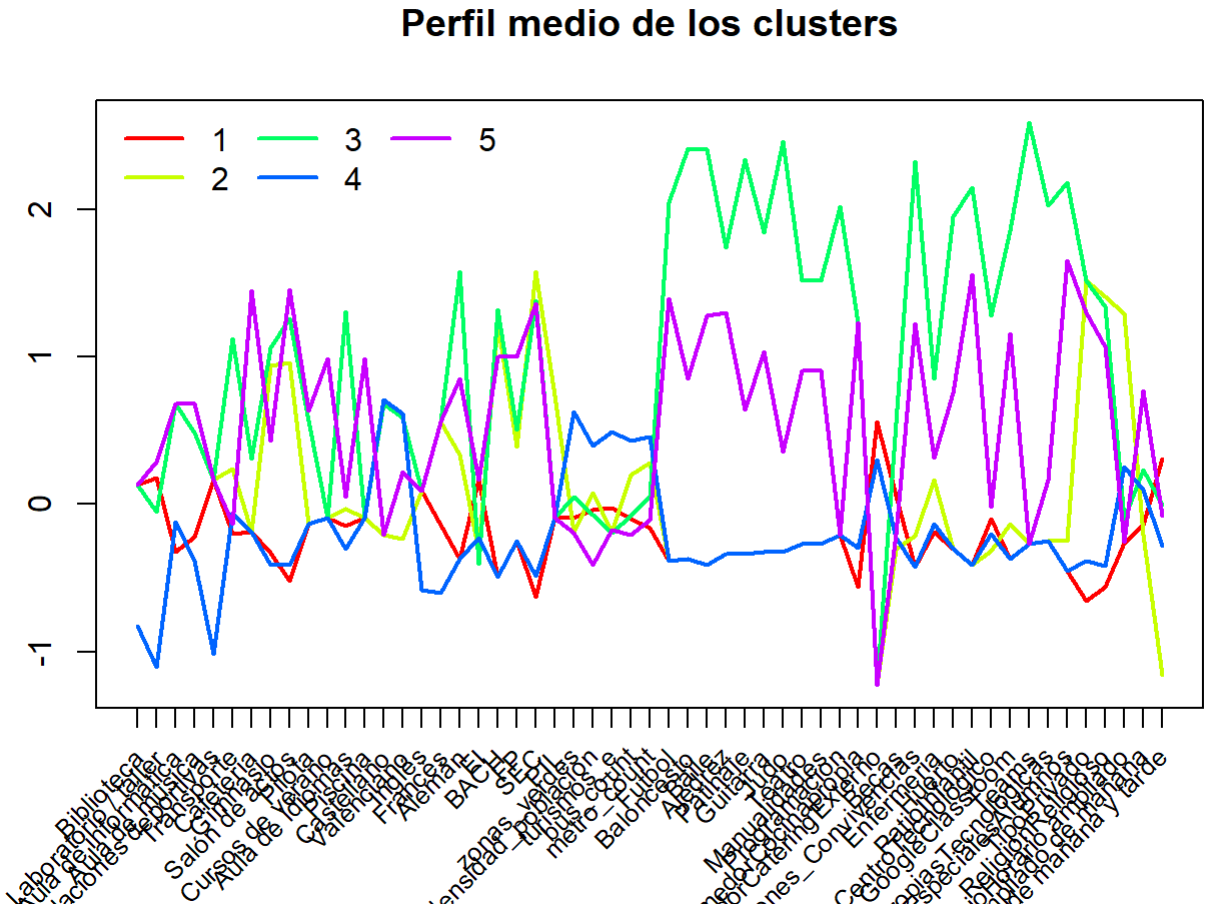
```
sil = data.frame(silhouette(clust3$cluster, midist))

mal = sil$sil_width < -0.045
malClasifiS = sil[mal,]
```

```
misclust = factor(clust3$cluster)

mediasCluster = aggregate(schools_cluster, by = list("cluster" = misclust), mean)[,-1]
rownames(mediasCluster) = paste0("c",1:5)
kable(t(round(mediasCluster,2)))
```

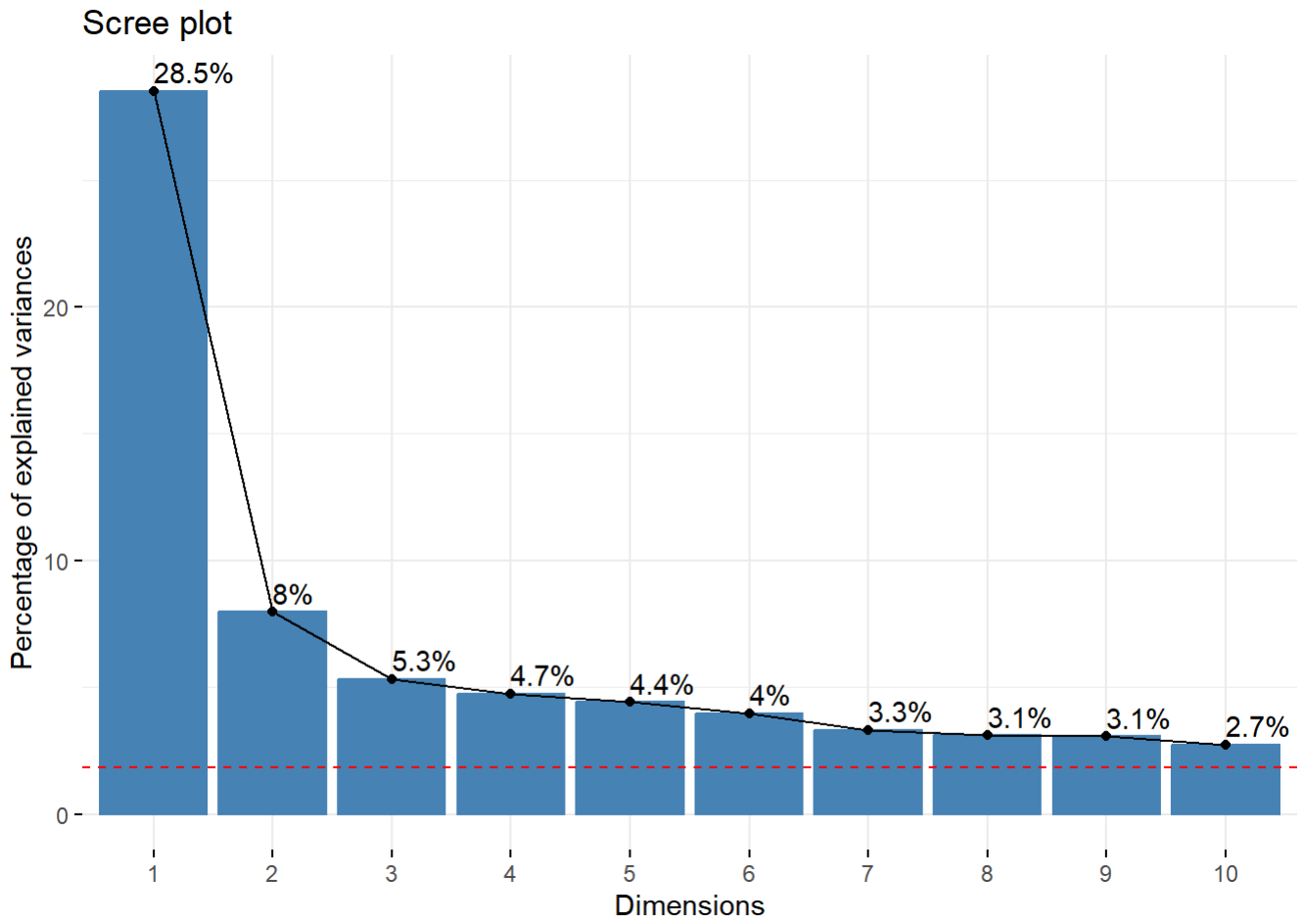
	c1	c2	c3	c4	c5
Biblioteca	0.13	0.13	0.13	-0.82	0.13
Laboratorio / taller	0.18	0.29	-0.05	-1.11	0.29
Aula de informática	-0.32	0.68	0.68	-0.12	0.68
Aula de música	-0.22	0.68	0.49	-0.39	0.68
Instalaciones deportivas	0.16	0.16	0.16	-1.01	0.16
Transporte	-0.20	0.24	1.12	-0.06	-0.13
Cafetería	-0.19	-0.19	0.31	-0.19	1.45
Gimnasio	-0.33	0.94	1.06	-0.41	0.43
Salón de actos	-0.52	0.96	1.26	-0.41	1.45
Pilota	-0.13	-0.13	0.56	-0.13	0.63
Cursos de verano	-0.09	-0.09	-0.09	-0.09	0.98
Aula de idiomas	-0.14	-0.03	1.31	-0.31	0.05
Piscina	-0.09	-0.09	-0.09	-0.09	0.98
Castellano	-0.21	-0.21	0.68	0.71	-0.21
Valenciano	-0.23	-0.23	0.58	0.61	0.22
Inglés	0.09	0.09	0.09	-0.58	0.09
Francés	-0.14	0.56	0.56	-0.60	0.56
Alemán	-0.37	0.34	1.58	-0.37	0.85
EI	0.16	-0.32	-0.41	-0.23	0.16
BACH	-0.50	1.23	1.32	-0.50	1.00
FP	-0.25	0.39	0.51	-0.25	1.00
SEC	-0.63	1.58	1.38	-0.49	1.36
PIL	-0.09	0.74	-0.09	-0.09	-0.09
zonas_verdes	-0.09	-0.19	0.05	0.63	-0.19
densidad_poblacion	-0.04	0.08	-0.07	0.40	-0.42

	c1	c2	c3	c4	c5
turismos_e	-0.02	-0.19	-0.19	0.49	-0.17
bus_count	-0.10	0.20	-0.08	0.43	-0.20
metro_count	-0.16	0.28	0.05	0.46	-0.11
Futbol	-0.38	-0.38	2.04	-0.38	1.40
Baloncesto	-0.37	-0.37	2.41	-0.37	0.85
Baile	-0.41	-0.41	2.40	-0.41	1.28
Ajedrez	-0.34	-0.34	1.74	-0.34	1.30
Patinaje	-0.34	-0.34	2.34	-0.34	0.64
Guitarra	-0.32	-0.32	1.84	-0.32	1.04
Judo	-0.32	-0.32	2.46	-0.32	0.36
Teatro	-0.27	-0.27	1.52	-0.27	0.91
Manualidades	-0.27	-0.27	1.52	-0.27	0.91
Programacion	-0.21	-0.21	2.02	-0.21	-0.21
ComedorCocinapropia	-0.56	1.23	1.23	-0.30	1.23
ComedorCateringExterno	0.56	-1.23	-1.23	0.30	-1.23
Becas	0.05	-0.30	0.55	-0.23	-0.20
Excursiones_Convivencias	-0.43	-0.22	2.32	-0.43	1.22
Enfermeria	-0.19	0.16	0.85	-0.13	0.32
Huerto	-0.31	-0.31	1.95	-0.31	0.76
PatioInfantil	-0.41	-0.41	2.15	-0.41	1.56
CentroTecnologico	-0.10	-0.32	1.28	-0.20	-0.02
GoogleClassroom	-0.37	-0.13	1.85	-0.37	1.16
Teams	-0.27	-0.27	2.59	-0.27	-0.27
HerramientasPropiasTecnologicas	-0.25	-0.25	2.03	-0.25	0.17
NecesidadesEspecialesAlumnos	-0.45	-0.25	2.18	-0.45	1.65
TipoPrivado	-0.65	1.51	1.51	-0.38	1.30
ReligionReligioso	-0.56	1.41	1.34	-0.42	1.07
HorarioHorario ampliado	-0.26	1.29	-0.10	0.25	-0.25
HorarioHorario ampliado de mañana	-0.14	-0.21	0.23	0.10	0.77
HorarioHorario ampliado de mañana y tarde	0.31	-1.16	0.00	-0.29	-0.08

```
par(mar = c(5, 4, 4, 2) + 0.1)
matplot(t(mediasCluster), type = "l", col = rainbow(5), ylab = "", xlab = "", lwd = 2,
        lty = 1, main = "Perfil medio de los clusters", xaxt = "n")
axis(side = 1, at = 1:ncol(schools_cluster), labels = FALSE)
text(x = 1:ncol(schools_cluster), y = par("usr")[3] - 0.3,
     labels = colnames(schools_cluster), srt = 45, adj = 1, xpd = TRUE, cex = 0.8)
legend("topleft", as.character(1:5), col = rainbow(5), lwd = 2, ncol = 3, bty = "n")
```




```
res.pca = PCA(prueba_bin, scale.unit = TRUE, graph = FALSE, ncp = 10)
eig.val <- get_eigenvalue(res.pca)
VPmedio = 100 * (1/nrow(eig.val))
fviz_eig(res.pca, addlabels = TRUE) +
  geom_hline(yintercept=VPmedio, linetype=2, color="red")
```

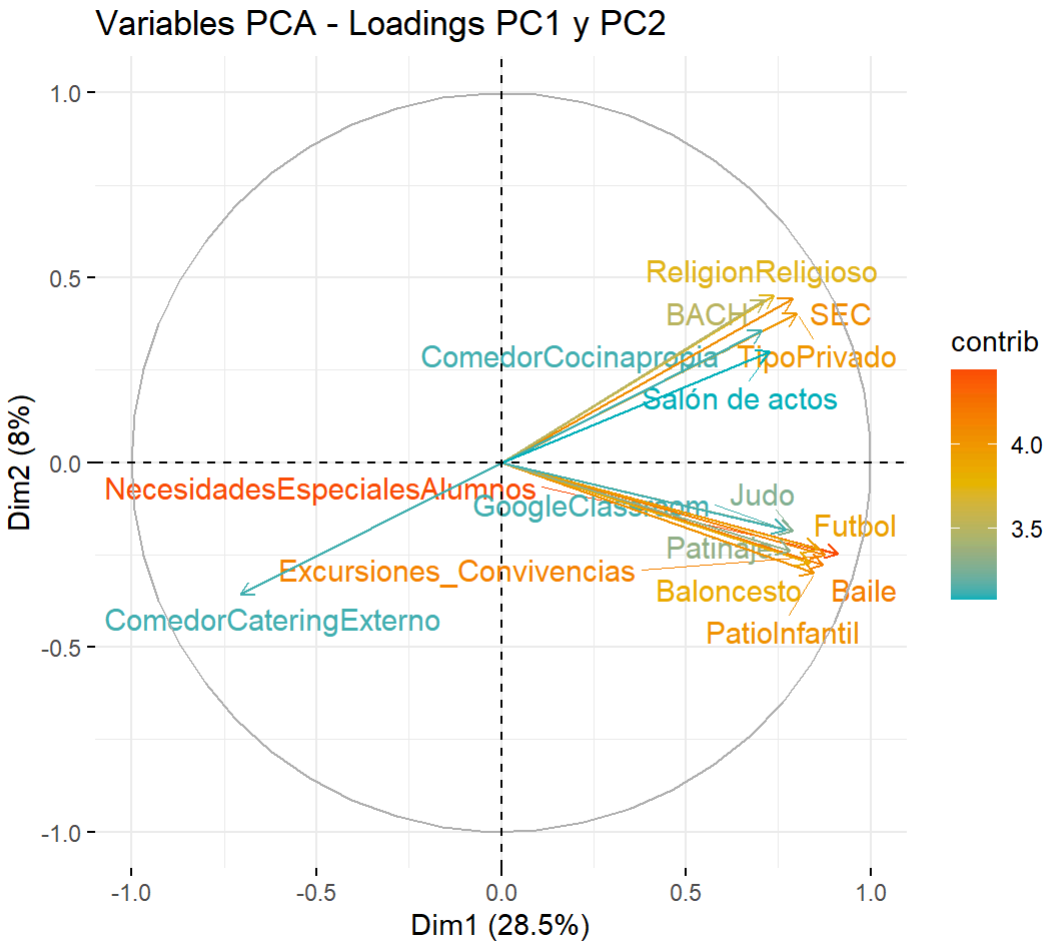



```
kable(eig.val[1:6,])
```

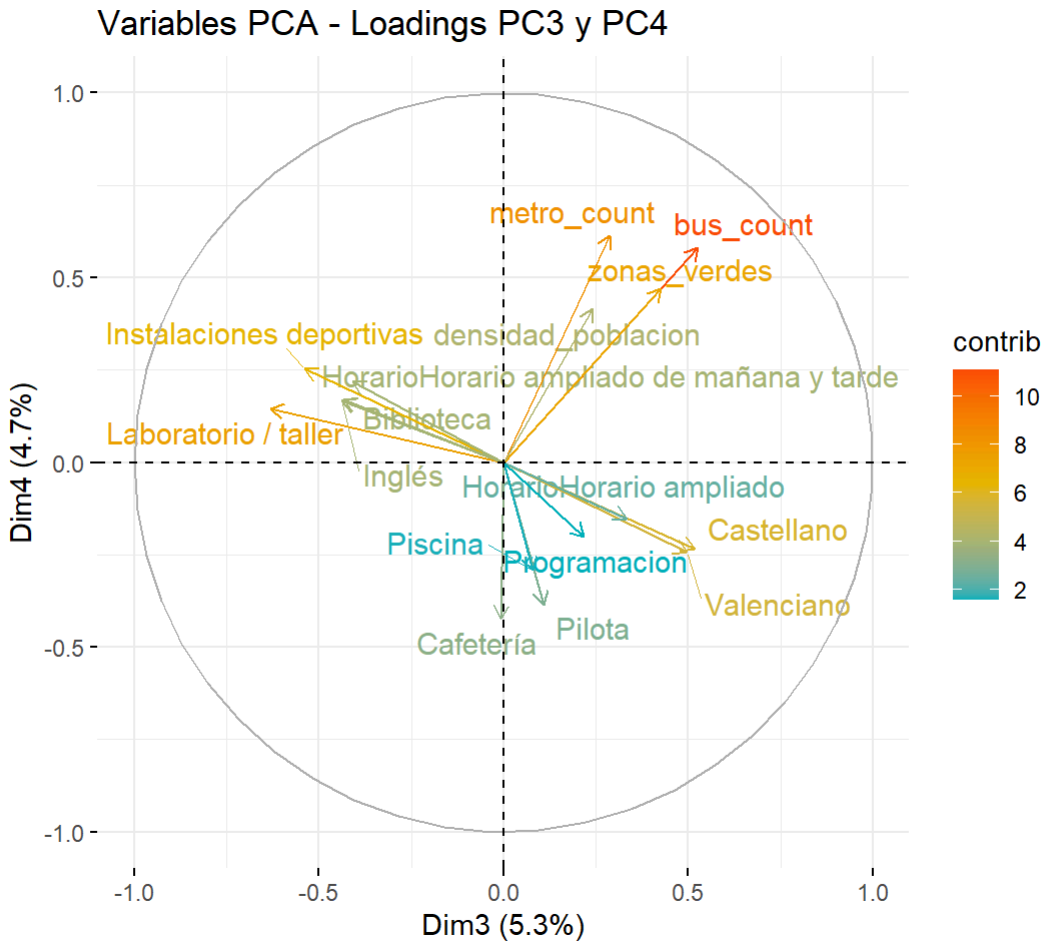
	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	15.686406	28.520738	28.52074
Dim.2	4.380264	7.964116	36.48485
Dim.3	2.921370	5.311581	41.79644
Dim.4	2.611000	4.747272	46.54371
Dim.5	2.427787	4.414159	50.95787
Dim.6	2.176983	3.958152	54.91602

```
K = 4
res.pca = PCA(prueba_bin, scale.unit = TRUE, graph = FALSE, ncp = K)
```

```
fviz_pca_var(res.pca, axes = c(1,2), repel = TRUE, col.var = "contrib",

             select.var = list(contrib = 16) ,
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),

             labelsize = 4,

             title = 'Variables PCA - Loadings PC1 y PC2')
```



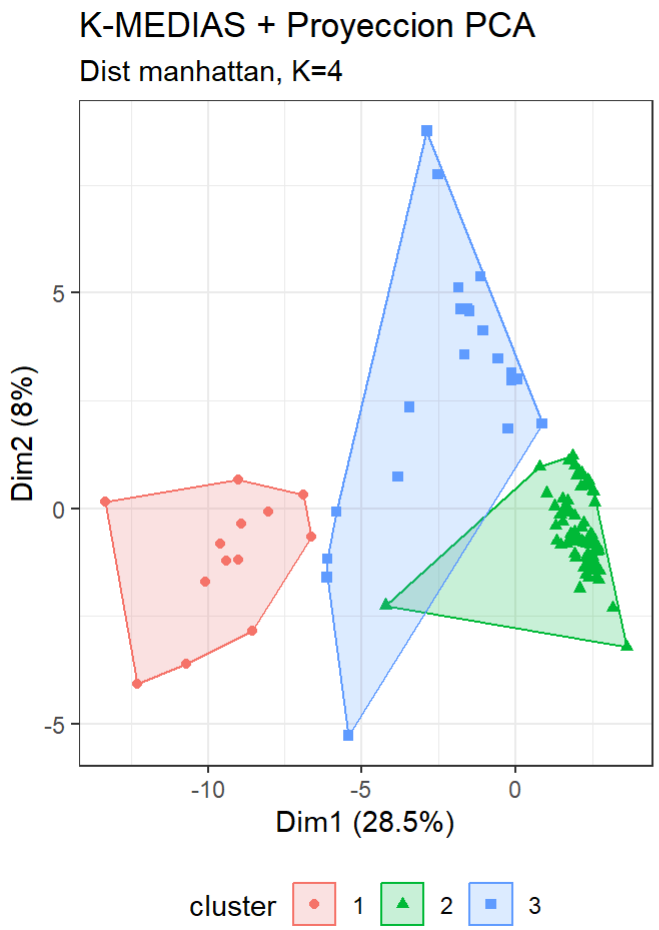

```
fviz_pca_var(res.pca, axes = c(3,4), repel = TRUE, col.var = "contrib",
             select.var = list(contrib=16),
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),

             labelsize = 4,

             title = 'Variables PCA - Loadings PC3 y PC4')
```

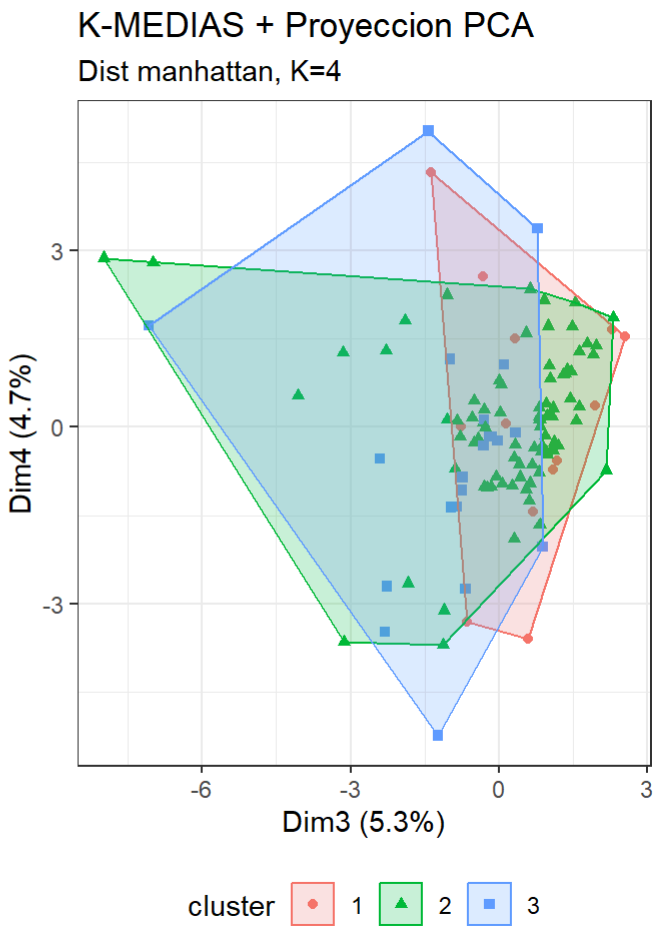



```
table(clust4$cluster)
```

```
##
##  1  2  3
## 13 82 21
```

```
p1 = fviz_cluster(object = list(data=schools_cluster, cluster=clust4$cluster), stand = FALSE,
             ellipse.type = "convex", geom = "point", show.clust.cent = FALSE,
             labelsize = 8)  +
  labs(title = "K-MEDIAS + Proyeccion PCA",
       subtitle = "Dist manhattan, K=4") +
  theme_bw() +
  theme(legend.position = "bottom")
p2 = fviz_cluster(object = list(data=schools_cluster, cluster=clust4$cluster), stand = FALSE,
             ellipse.type = "convex", geom = "point", show.clust.cent = FALSE,
             labelsize = 8, axes = 3:4)  +
  labs(title = "K-MEDIAS + Proyeccion PCA",
       subtitle = "Dist manhattan, K=4") +
  theme_bw() +
  theme(legend.position = "bottom")
grid.arrange(p1, p2, nrow = 1)
```






```
sil = data.frame(silhouette(clust4$cluster, midist))

mal = sil$sil_width < -0.045
malClasifiS = sil[mal,]
```

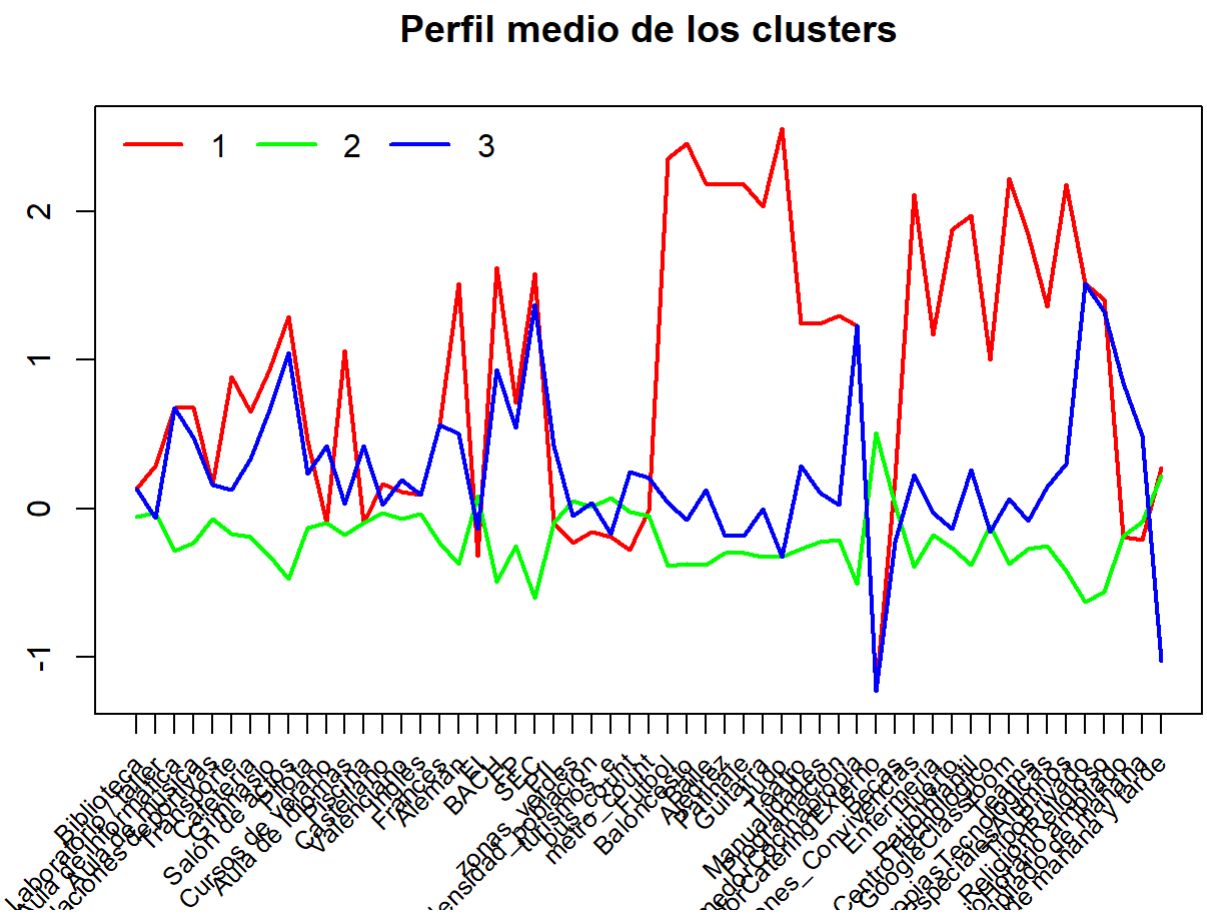
```
misclust = factor(clust4$cluster)

mediasCluster = aggregate(schools_cluster, by = list("cluster" = misclust), mean)[,-1]
rownames(mediasCluster) = paste0("c",1:3)
kable(t(round(mediasCluster,2)))
```

	c1	c2	c3
Biblioteca	0.13	-0.05	0.13
Laboratorio / taller	0.29	-0.03	-0.07
Aula de informática	0.68	-0.28	0.68
Aula de música	0.68	-0.23	0.48
Instalaciones deportivas	0.16	-0.07	0.16
Transporte	0.89	-0.17	0.12
Cafetería	0.65	-0.19	0.33
Gimnasio	0.94	-0.32	0.67
Salón de actos	1.29	-0.47	1.05

	c1	c2	c3
Pilota	0.46	-0.13	0.23
Cursos de verano	-0.09	-0.09	0.42
Aula de idiomas	1.06	-0.18	0.03
Piscina	-0.09	-0.09	0.42
Castellano	0.17	-0.03	0.02
Valenciano	0.11	-0.07	0.20
Inglés	0.09	-0.04	0.09
Francés	0.56	-0.23	0.56
Alemán	1.51	-0.37	0.50
EI	-0.32	0.09	-0.14
BACH	1.62	-0.50	0.93
FP	0.71	-0.25	0.54
SEC	1.58	-0.60	1.37
PIL	-0.09	-0.09	0.42
zonas_verdes	-0.23	0.05	-0.05
densidad_poblacion	-0.15	0.01	0.04
turismos_e	-0.19	0.07	-0.17
bus_count	-0.28	-0.02	0.25
metro_count	-0.01	-0.05	0.21
Futbol	2.36	-0.38	0.04
Baloncesto	2.45	-0.37	-0.08
Baile	2.19	-0.38	0.12
Ajedrez	2.18	-0.30	-0.18
Patinaje	2.18	-0.30	-0.18
Guitarra	2.03	-0.32	0.00
Judo	2.55	-0.32	-0.32
Teatro	1.24	-0.27	0.29
Manualidades	1.24	-0.22	0.10
Programacion	1.30	-0.21	0.02
ComedorCocinapropia	1.23	-0.51	1.23
ComedorCateringExterno	-1.23	0.51	-1.23
Becas	0.20	0.03	-0.22

	c1	c2	c3
Excursiones_Convivencias	2.11	-0.39	0.23
Enfermeria	1.17	-0.18	-0.03
Huerto	1.88	-0.26	-0.14
PatioInfantil	1.97	-0.38	0.26
CentroTecnologico	1.00	-0.12	-0.16
GoogleClassroom	2.22	-0.37	0.07
Teams	1.84	-0.27	-0.08
HerramientasPropiasTecnologicas	1.36	-0.25	0.15
NecesidadesEspecialesAlumnos	2.18	-0.42	0.30
TipoPrivado	1.51	-0.63	1.51
ReligionReligioso	1.41	-0.56	1.32
HorarioHorario ampliado	-0.19	-0.19	0.84
HorarioHorario ampliado de mañana	-0.21	-0.09	0.49
HorarioHorario ampliado de mañana y tarde	0.27	0.22	-1.03

```
par(mar = c(5, 4, 4, 2) + 0.1)
matplot(t(mediasCluster), type = "l", col = rainbow(3), ylab = "", xlab = "", lwd = 2,
        lty = 1, main = "Perfil medio de los clusters", xaxt = "n")
axis(side = 1, at = 1:ncol(schools_cluster), labels = FALSE)
text(x = 1:ncol(schools_cluster), y = par("usr")[3] - 0.3,
     labels = colnames(schools_cluster), srt = 45, adj = 1, xpd = TRUE, cex = 0.8)
legend("topleft", as.character(1:3), col = rainbow(3), lwd = 2, ncol = 3, bty = "n")
```



-Cluster 1 (rojo): Presentan laboratorio/taller, aula de música, transporte privado, cafetería. Otras instalaciones como gimnasio y salon de actos (similar al C3). Presentan aula de idiomas. Se encuentran en este cluster los centros que dan Aleman. Se localizan los centros desde primaria hasta bachiller (no infantil). Presentan numerosas extraescolares (futbol, ajedrez, teatro), además de numerosos servicios (teams, convivencias…). Capacidad para alumnos con necesidades especiales. Centros privados y religiosos con horarios de mañana y tarde. Hacen cursos de verano

-Cluster 2 (verde): No presentan muchas instalaciones, se sitúan la mayoría de centros de EI unicamente de ahi las pocas instalaciones. Presenta servicio de catering a diferencia de los otros 2. Generalmente público y laicos con horarios ampliados de mañana y tarde.

-Cluster 3 (azul): Bastante similar al cluster 1. Esta mejor comunicado con paradas de bus. No tienen laboratorio/taller, presentan piscina pero no tienen cursos de verano. Hay menos centros que ofrecen francés. Se seitúan en barrios más poblados y no ofrecen tantas extraescolares (a excepción de teatro parece). Son privados y religiosos pero con horarios ampliados de mañana.