PruebaExcel

Juan Francisco Pallardó Latorre

2024-05-16

```
dataEx = read_excel("rank_SEC_BACH.xlsx")
load("datosCOMPLETOS.rdata")

a = datosCOMPLETOS[datosCOMPLETOS$SEC == 1, ]

a = a[,c(1, 59, 61, 76:105)]

p = dataEx[ , -c(32,34)]

b = merge ( p , a , by = 'Identificador')

b = na.omit(b)
```

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

```
eliminar_v = c(1,2,4,5,14, 31:35, 50, 51, 54, 77, 78 )
prueba = b[, -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)[6] = "Gimnasio"
```

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

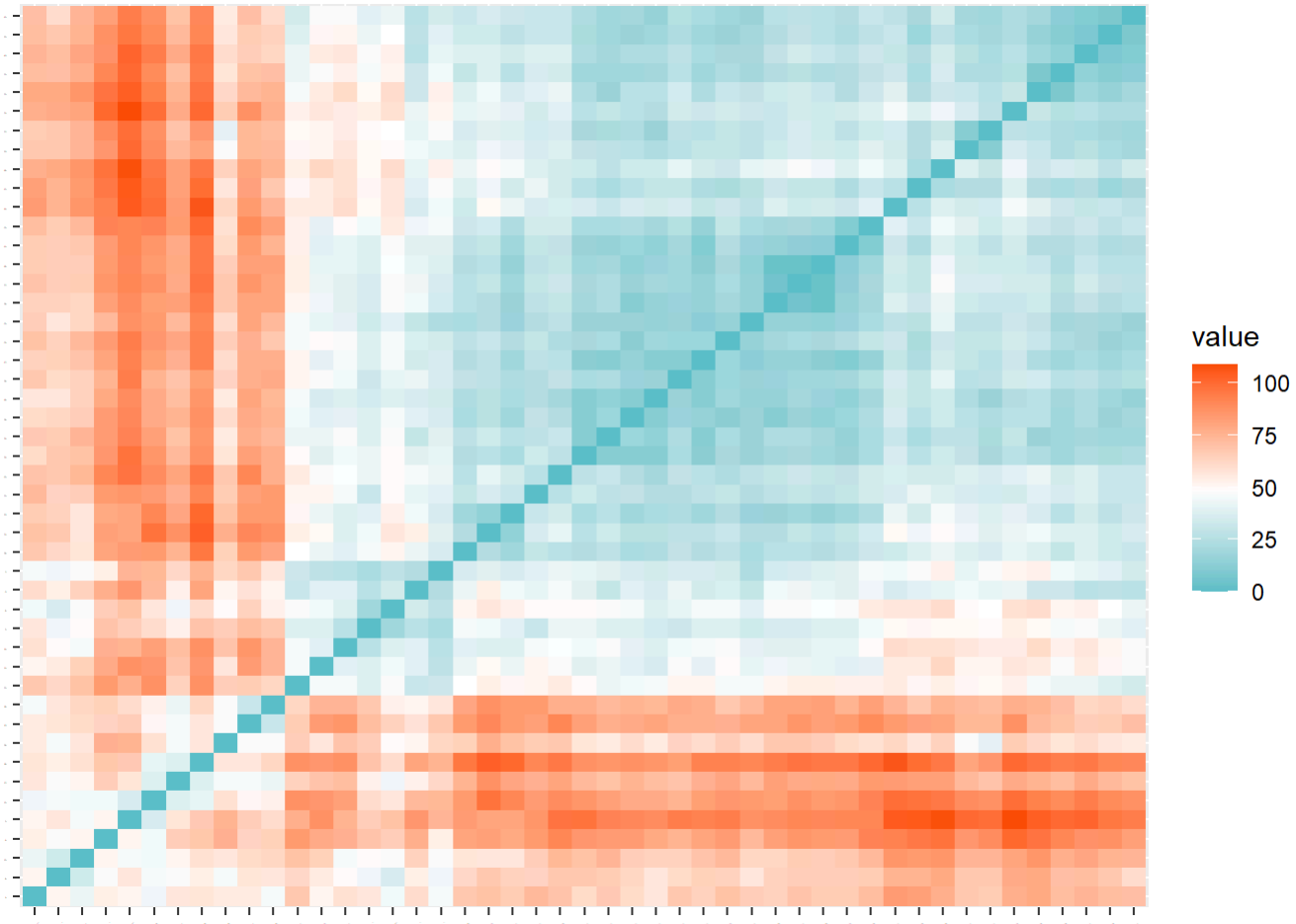
```
var_dummy = as.data.frame(model.matrix(~ Tipo -1, data=prueba))
prueba = cbind(prueba[-which(names(prueba) == "Tipo")], var_dummy)

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, PrecioEntre 100 y 300€ y Horario Ampliado Mañana y tarde. Además de instalaciones deportivas.

```
prueba_bin = prueba[,-c(2, 59, 63, 65, 68)]
```

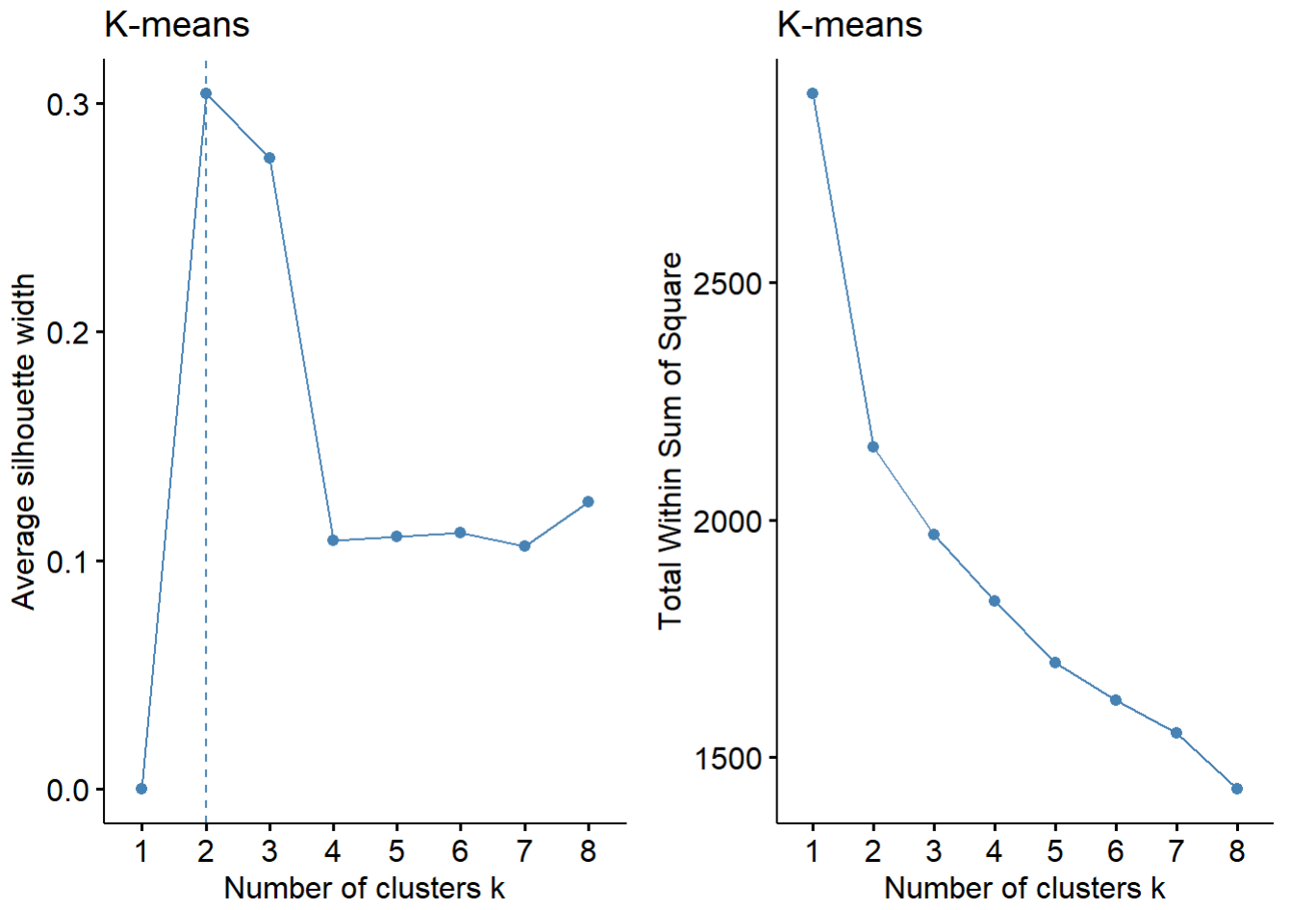
Hago el escalado y centrado de los datos, usare la distancia de Manhattan (ver diferencias con Euclídea). Printeo una matriz de todas las instancias y su distancia.

```
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"))
```

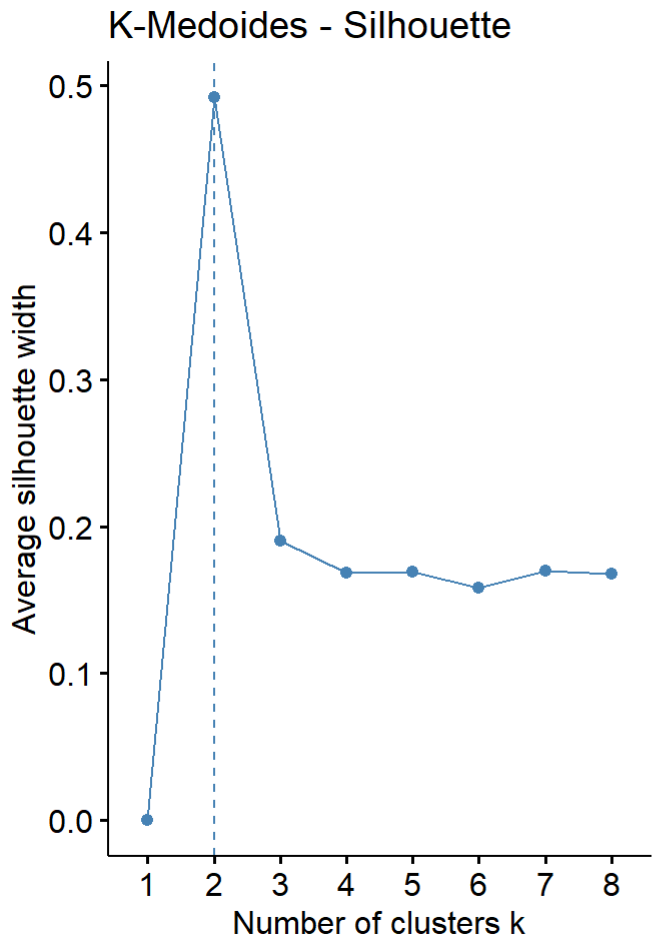


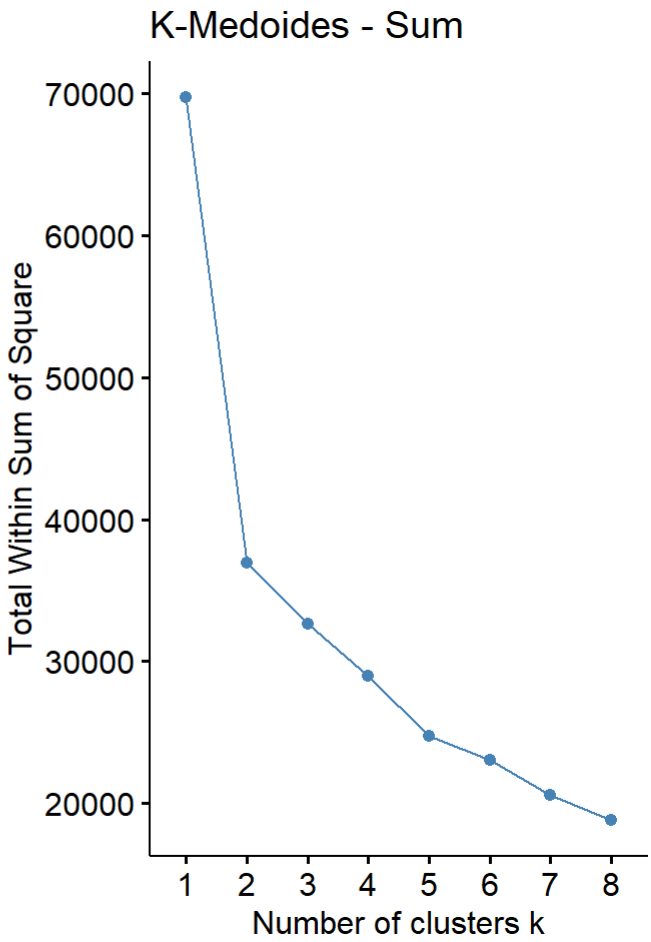
En el gráfico obtenido a partir de las distancias podemos observar posibles agrupaciones entre centros que serían los cuadrados azules que se forman a través de la diagonal principal de la matriz. El siguiente paso será realizar un método de partición, en concreto realizaremos el algoritmo **k-medias**. 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)
```

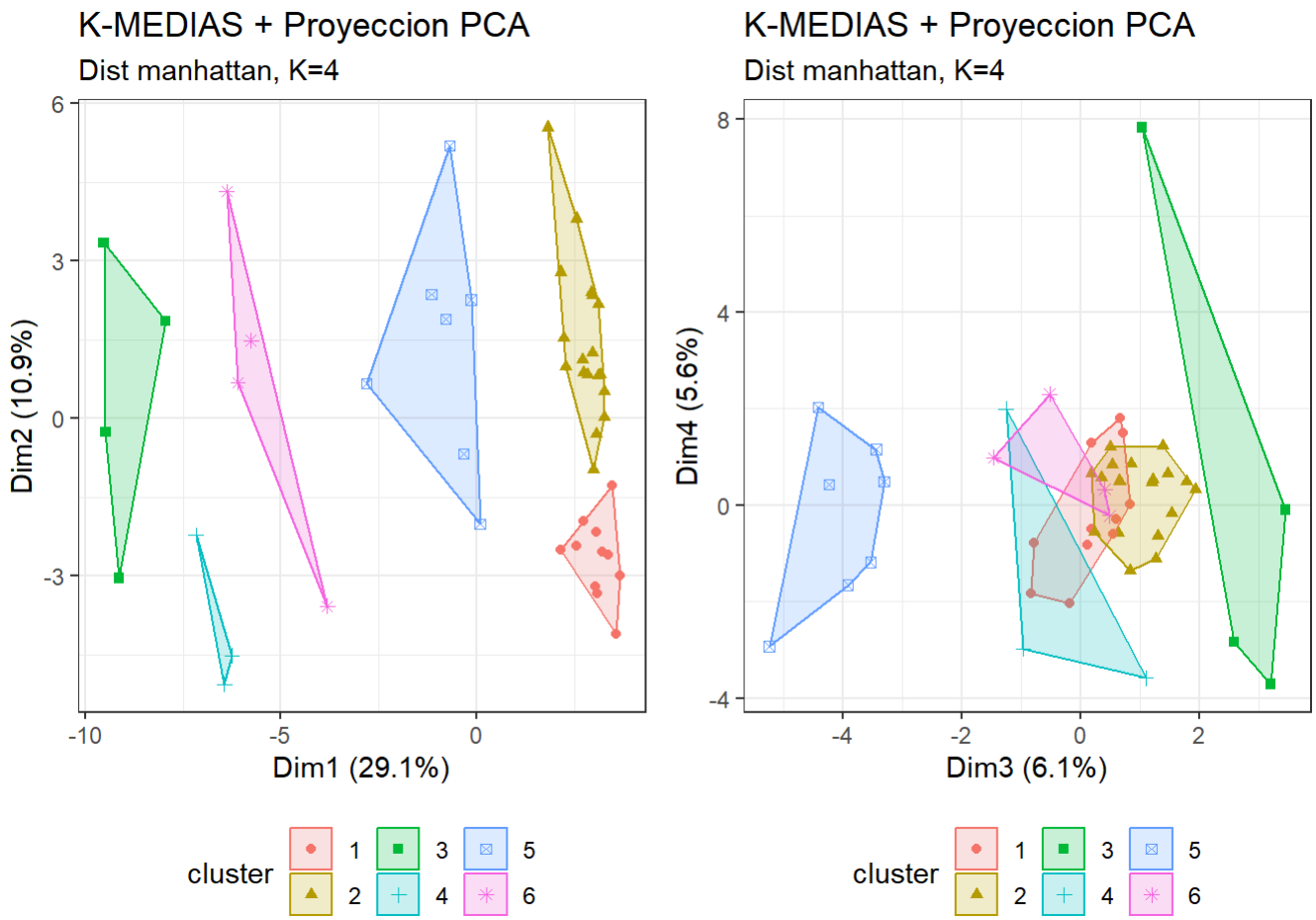




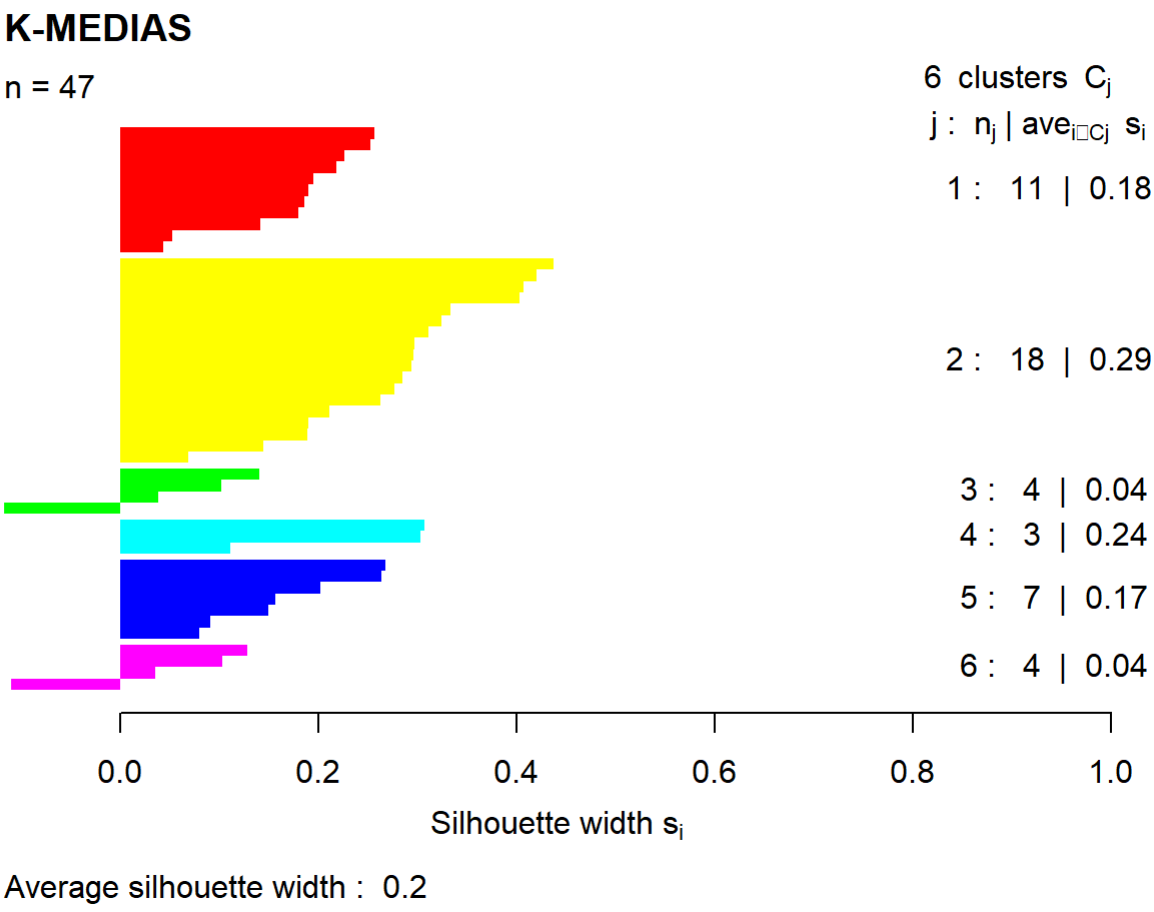

```
set.seed(100)
clust3 <- kmeans(midist, centers = 6, nstart = 20)
table(clust3$cluster)
```

```
##
##  1  2  3  4  5  6
## 11 18  4  3  7  4
```

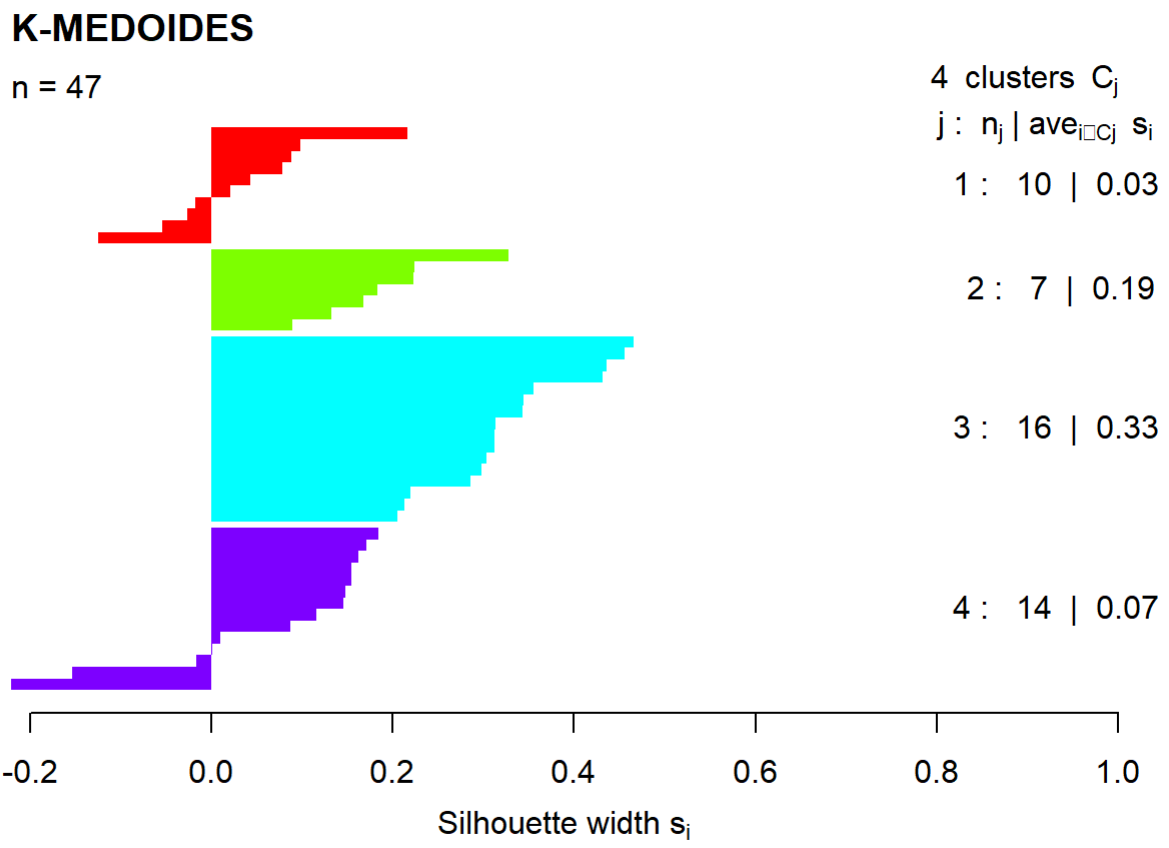
```
p1 = fviz_cluster(object = list(data=schools_cluster, cluster=clust3$cluster), stand = FALSE,
             ellipse.type = "convex", geom = "point", show.clust.cent = FALSE,
             labelsize = 8)  +
  labs(title = "K-MEDIAS + Proyeccion PCA",
       subtitle = "Dist manhattan, K=4") +
  theme_bw() +
  theme(legend.position = "bottom")
p2 = fviz_cluster(object = list(data=schools_cluster, cluster=clust3$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(6), border=NA, main = "K-MEDIAS")
```



```
clust4 <- pam(schools_cluster, k = 4)
plot(silhouette(clust4$clustering, midist), col=rainbow(4), border=NA, main = "K-MEDOIDES")
```




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

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

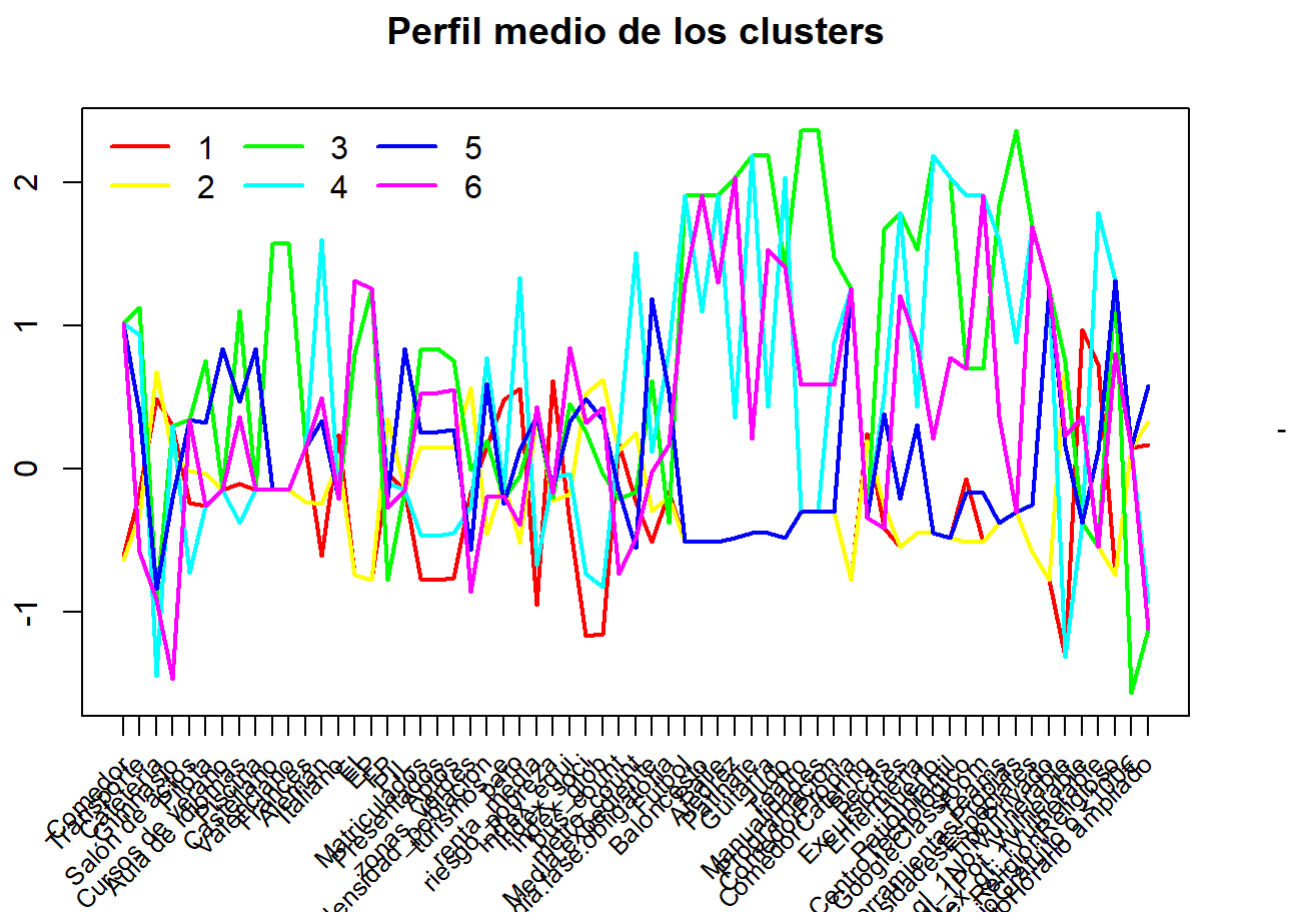
```
misclust = factor(clust3$cluster)

mediasCluster = aggregate(schools_cluster, by = list("cluster" = misclust), mean)[,-1]
rownames(mediasCluster) = paste0("c",1:6)
kable(t(round(mediasCluster,2)))
```

	c1	c2	c3	c4	c5	c6
Comedor	-0.61	-0.64	1.01	1.01	1.01	1.01
Transporte	-0.17	-0.33	1.12	0.93	0.39	-0.58
Cafetería	0.48	0.68	-0.91	-1.44	-0.84	-0.91
Gimnasio	0.30	0.10	0.30	0.30	-0.20	-1.47
Salón de actos	-0.24	-0.02	0.34	-0.73	0.34	0.34
Pilota	-0.26	-0.03	0.75	-0.26	0.32	-0.26
Cursos de verano	-0.15	-0.15	-0.15	-0.15	0.83	-0.15

	c1	c2	c3	c4	c5	c6
Aula de idiomas	-0.11	-0.38	1.10	-0.38	0.47	0.36
Piscina	-0.15	-0.15	-0.15	-0.15	0.83	-0.15
Castellano	-0.15	-0.15	1.57	-0.15	-0.15	-0.15
Valenciano	-0.15	-0.15	1.57	-0.15	-0.15	-0.15
Francés	0.15	-0.24	0.15	0.15	0.15	0.15
Alemán	-0.61	-0.24	0.49	1.60	0.34	0.49
Italiano	0.24	0.06	-0.21	-0.21	-0.21	-0.21
EI	-0.74	-0.74	0.80	1.31	1.31	1.31
EP	-0.78	-0.78	1.26	1.26	1.26	1.26
FP	-0.04	0.35	-0.78	-0.10	-0.20	-0.27
PIL	-0.15	-0.15	-0.15	-0.15	0.83	-0.15
Matriculados	-0.77	0.15	0.83	-0.46	0.25	0.52
Presentados	-0.77	0.15	0.84	-0.47	0.25	0.53
Aptos	-0.77	0.15	0.76	-0.45	0.27	0.55
zonas_verdes	-0.17	0.57	-0.01	-0.27	-0.57	-0.86
densidad_poblacion	0.16	-0.45	0.19	0.77	0.59	-0.20
turismos_e	0.48	-0.08	-0.21	-0.16	-0.24	-0.19
paro	0.56	-0.52	-0.05	1.33	0.13	-0.39
renta_media	-0.95	0.38	0.37	-0.67	0.37	0.43
riesgo_pobreza	0.61	-0.23	-0.20	-0.07	-0.14	-0.17
index_equi	-0.38	-0.18	0.45	-0.04	0.33	0.85
index_soci	-1.17	0.52	0.26	-0.74	0.48	0.32
index_glob	-1.15	0.62	-0.04	-0.83	0.34	0.43
bus_count	0.17	0.12	-0.21	0.23	-0.15	-0.73
metro_count	-0.24	0.25	-0.16	1.51	-0.55	-0.49
Media.expediente	-0.51	-0.30	0.62	0.12	1.18	-0.02
Media.fase.obligatoria	-0.16	-0.20	-0.38	0.83	0.53	0.17
Futbol	-0.51	-0.51	1.90	1.90	-0.51	1.30
Baloncesto	-0.51	-0.51	1.90	1.10	-0.51	1.90
Baile	-0.51	-0.51	1.90	1.90	-0.51	1.30
Ajedrez	-0.48	-0.48	2.03	0.36	-0.48	2.03
Patinaje	-0.45	-0.45	2.18	2.18	-0.45	0.21

	c1	c2	c3	c4	c5	c6
Guitarra	-0.45	-0.45	2.18	0.43	-0.45	1.53
Judo	-0.48	-0.48	1.40	2.03	-0.48	1.40
Teatro	-0.30	-0.30	2.36	-0.30	-0.30	0.58
Manualidades	-0.30	-0.30	2.36	-0.30	-0.30	0.58
Programacion	-0.30	-0.30	1.47	0.88	-0.30	0.58
ComedorPropia	-0.78	-0.78	1.26	1.26	1.26	1.26
ComedorCatering	0.24	0.19	-0.34	-0.34	-0.34	-0.34
Becas	-0.41	-0.26	1.67	0.51	0.38	-0.41
Excursiones	-0.55	-0.55	1.79	1.79	-0.21	1.21
Enfermeria	-0.45	-0.45	1.53	0.43	0.30	0.87
Huerto	-0.45	-0.45	2.18	2.18	-0.45	0.21
PatioInfantil	-0.48	-0.48	2.03	2.03	-0.48	0.78
CentroTecnologico	-0.07	-0.51	0.69	1.90	-0.17	0.69
GoogleClassroom	-0.51	-0.51	0.69	1.90	-0.17	1.90
Teams	-0.38	-0.38	1.84	1.60	-0.38	0.36
HerramientasPropias	-0.30	-0.30	2.36	0.88	-0.30	-0.30
NecesidadesEspeciales	-0.58	-0.58	1.69	1.69	-0.26	1.69
TipoPrivado	-0.78	-0.78	1.26	1.26	1.26	1.26
index_gl_1No Vulnerable	-1.31	0.74	0.74	-1.31	0.16	0.23
index_gl_1Pot. Vulnerable	0.97	-0.38	-0.38	-0.38	-0.38	0.36
index_gl_1Vulnerable	0.73	-0.55	-0.55	1.79	0.12	-0.55
ReligionReligioso	-0.74	-0.74	1.31	1.31	1.31	0.80
PrecioGratuito o <100€	0.15	0.15	-1.57	0.15	0.15	0.15
HorarioHorario ampliado	0.17	0.33	-1.12	-0.93	0.58	-1.12

```
par(mar = c(5, 4, 4, 2) + 0.1)
matplot(t(mediasCluster), type = "l", col = rainbow(6), 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(1:6), col = rainbow(6), lwd = 2, ncol = 3, bty = "n")
```



Cluster 1 (rojo): Generalmente no presentan comedor, pero si cafetería. En cuanto idiomas no presentan Alemán pero si Italiano. Centros de secundaria y bachiller. Numero de alumnado bajo, las medias de bachiller son las mas bajas y en cuanto a la PAU las segundas mas bajas.. Barrios con un paro elevado, renta_media la mas baja, pobreza la mas alta. En cuanto a los indices podemos decir que son bajos lo que hace que sean potencialmente vulnerables. No muchas extraescolares. Comedor de catering. Poco numero de becas, así como numero de instalaciones bajo. Centro públicos y laicos. Precio <100€.

-Cluster 2 (amarillo): Generalmente no presentan comedor, pero si cafeteria. Linea en castellano sin ofrecer otros idiomas (francés, italiano o aleman). Centros de secundaria y bachiller, además cuentan con FP. Se sitúan en barrios poco poblados y con un paro bajo, en cuanto a la renta medio-alto, con indice de pobreza bajo. Indices social y glob altos lo que hace que el barrio fuera no vulnerable. Bien conectados por bus y metro. Media de expediente generalmente baja (segunda peor) parecido en la PAU. Centros sin una gran cantidad de instalaciones. Centros públicos, laicos y con precio <100€.

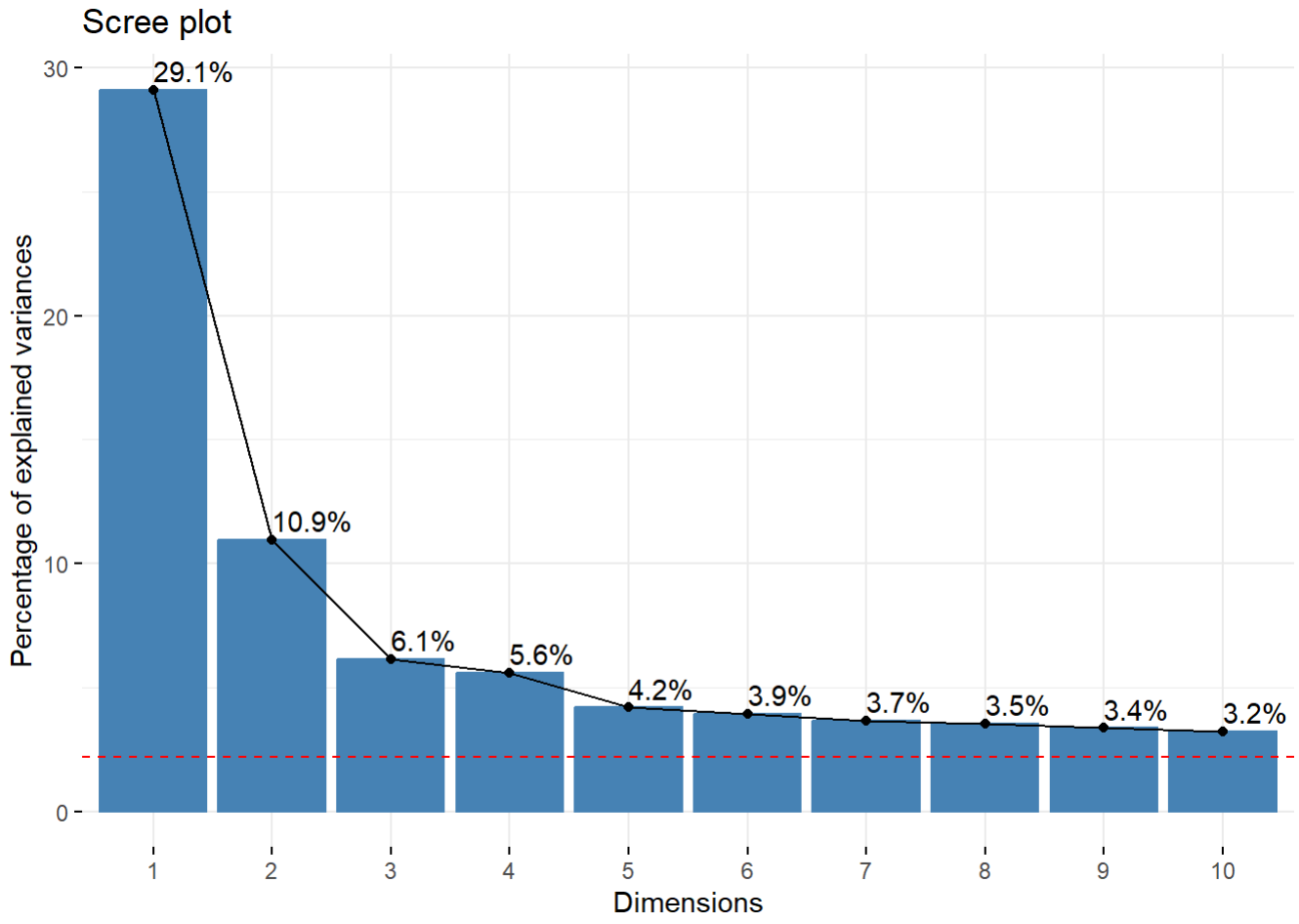
-Cluster 3 (verde): Centros con comedor, pero sin cafeteria. Usualmente con transporte privado. Gran numero de instalaciones (aula de idiomas, salon de actos, pilota…) Con líneas en castellano y valenciano. Generalmente ofrecen idiomas (frances, aleman e italiano). Estos centros abarcan desde EI hasta bachiller, pero carecen de FP. Gran número de matriculados, con una buena nota de bachiller (segunda mejor), pero en la PAU bajan siendo la peor. Gran numero de extraescolares e instalaciones/servicios. En cuanto a los barrios no destacan en nada pero no son vulnerables. Centros privados y religiosos. Capacidad de tratar con alumnos con necesidades especiales. Precios elevados.

-Cluster 4 (azul-claro): Centros con comedor, pero sin cafeteria. Usualmente con transporte privado. Generalmente sin salon de actos ni aula de idiomas entre otras instalaciones. Destacan por ser institutos con alemán. Son centros desde infantil hasta bachiller con un bajo numero de alumnos. Barrios donde la densidad de población es elevada así como el paro. Usualmente un renta media baja. En cuanto a los indices generalmente bajos lo que hace que se sitúen en barrios vulnerables. Media de expediente normal pero la nota de la PAU alta. Ofrece algunas extraescolares. Son centros privaods y religiosos, pero precio inferior a los 100€.

-Cluster 5 (azul-oscuro): Centros con algunas instalaciones (Salon de actos, gimnasio, pilota, piscina…). Se suelen realizar cursos de verano. No destacan en ningun idioma. Linea en castellano. Son centros desde infantil hasta bachiller con FP, media de bachiller y PAU elevadas. Se sitúan en barrios de población alta y buena renta media. En cuanto a los indices generalmente altos lo que hace que los barrios sean vulnerables y no vulnerables. No están muy bien comunicados (cuentan con transporte privado). No ofrecen muchas extraescolares y servicios. Son centros privados y religiosos, pero precio inferior a los 100€.

-Cluster 6 (rosa): Centros con comedor, sin transporte ni gimnasio. Presentan algunas instalaciones como salon de actos, aula de idiomas… Solo línea castellana pero ofrecen generlamente alemán como segunda lengua. Centros desde EI hasta Bachiller, con un elevado numero de alumnos pero su nota de expediente y PAU no muy destacables. Se sitúan en barrios con una densidad de poblacion baja, paro bajo y renta media superior al resto de clusters. Indices generalmente altos lo que hace que sean barrios no vulnerables o potencialmente vulnerables. Se sitúan en zonas mal comunicadas. Ofrecen muchas extraescolares y servicios. Tienen capacidad de tratar con alumnos con necesidades especiales. Son centros privados laicos y religioso, cuyo precio es inferior a los 100€. Horario ampliado.

```
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")
```



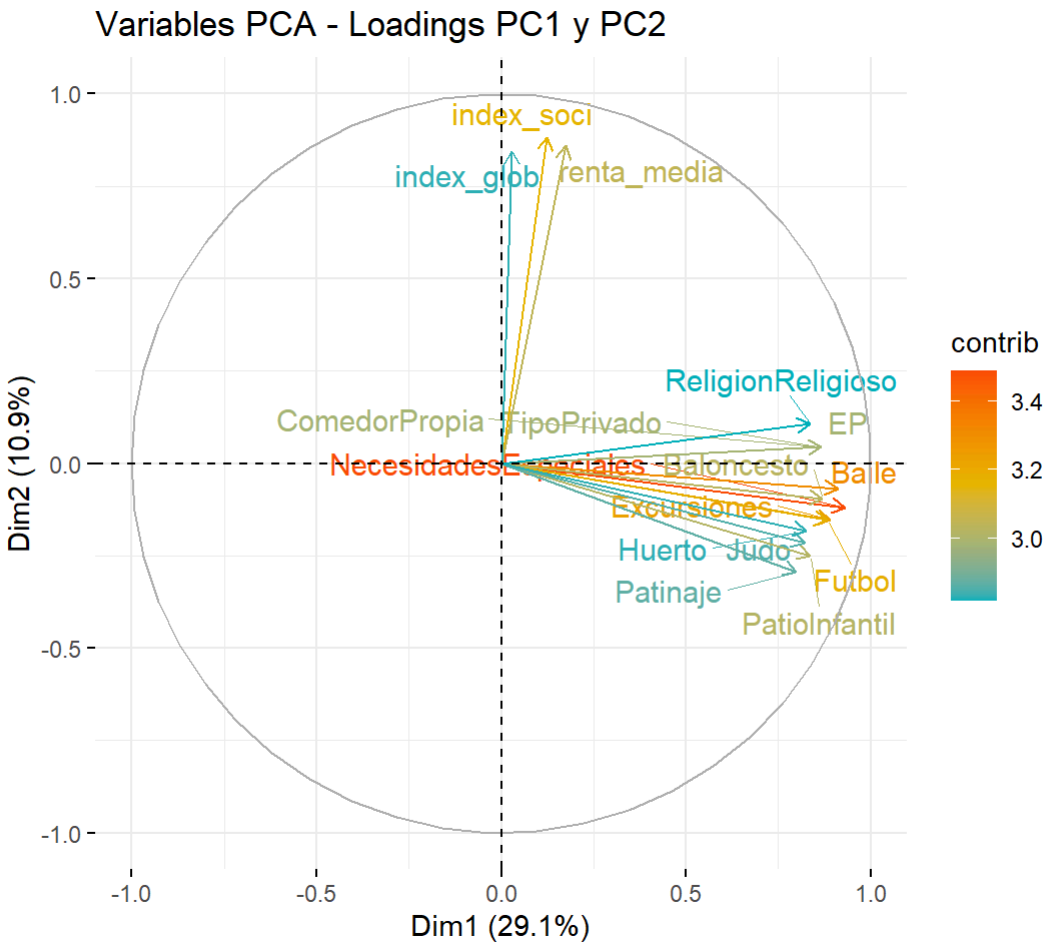

```
kable(eig.val[1:6,])
```

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	18.341894	29.114118	29.11412

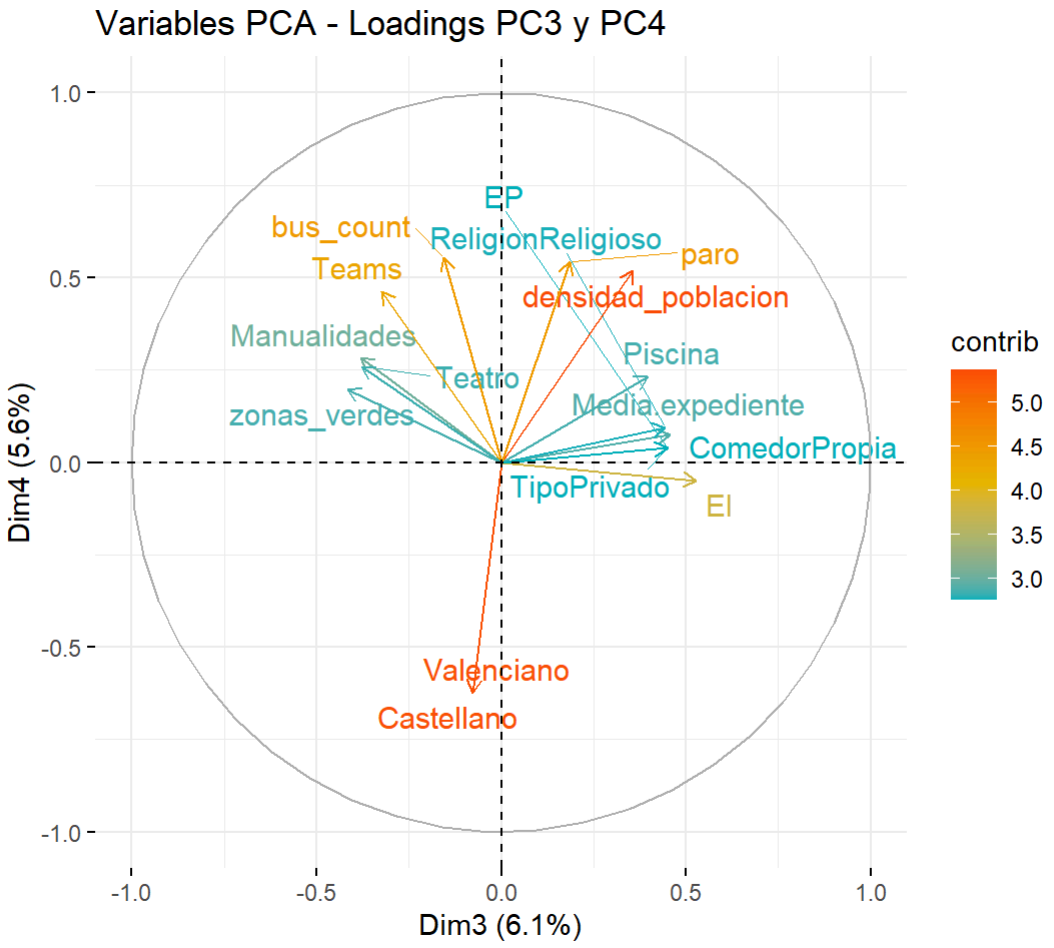
	eigenvalue	variance.percent	cumulative.variance.percent
Dim.2	6.893173	10.941545	40.05566
Dim.3	3.870383	6.143466	46.19913
Dim.4	3.511503	5.573815	51.77294
Dim.5	2.638036	4.187358	55.96030
Dim.6	2.478081	3.933462	59.89376

```
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')
```




```
n_clusters = as.data.frame(clust3$cluster)
```

```
#library(xlsx)
#write.xlsx(prueba_bin, "DatosSecundaria.xlsx", row.names = FALSE)
```