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.

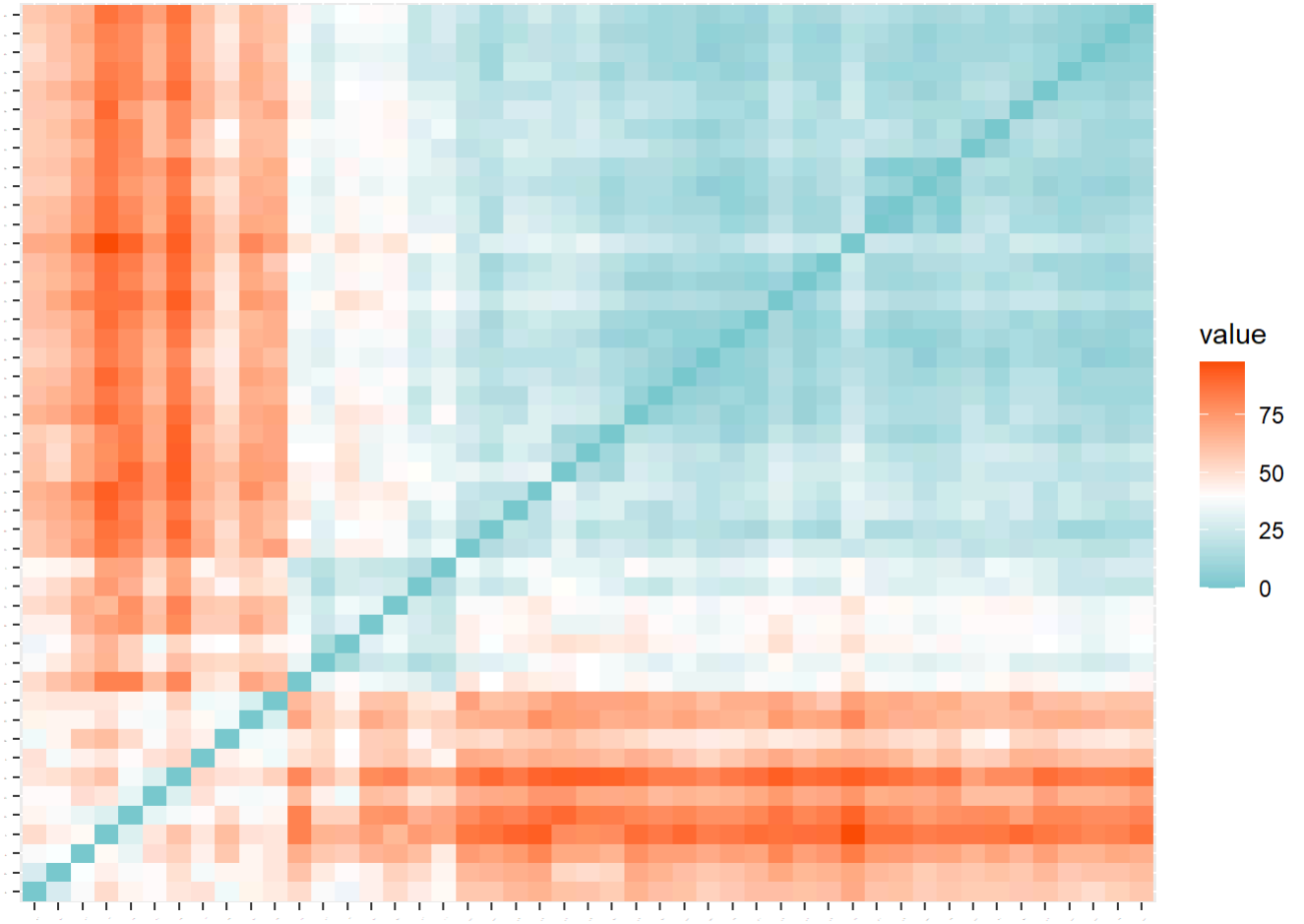
```
prueba_bin = prueba[,-c(2, 59, 63, 65, 68)]
```

Elimino las variables socioeconómicas: paro, renta_media, riesgo_pobreza, index_equi, index_soci, index_glob, index_gl_1 y precio gratuito o < 100€.

```
prueba_bin = prueba_bin[, -c(25:30,58:60, 62)]
```

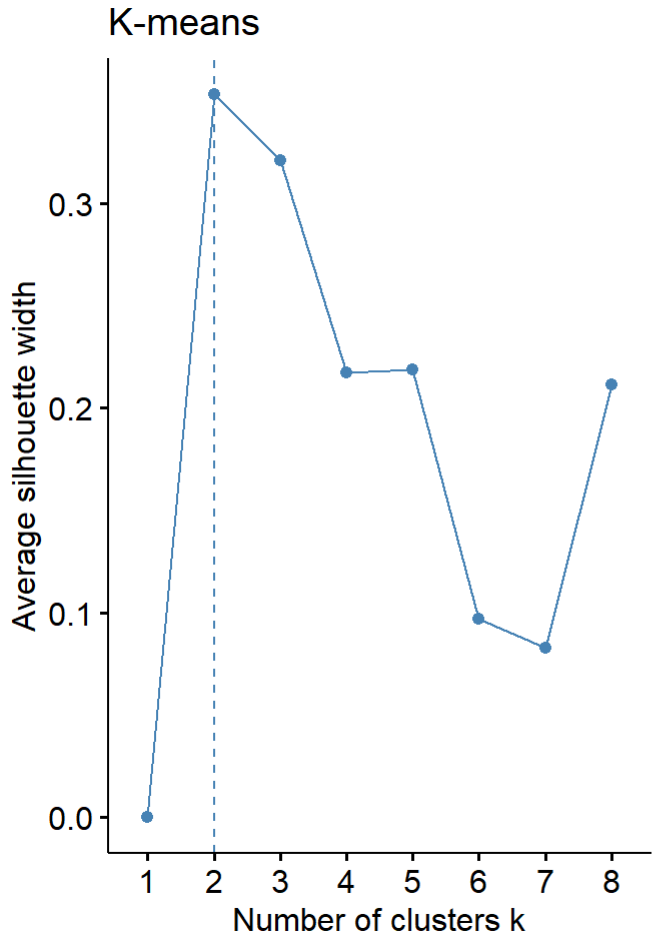
Hago el escalado y centrado de los datos, usare la distancia de Manhattan (ver diferencias con Euclídea). Printeo una matriz de todas las instancias y su distancia.

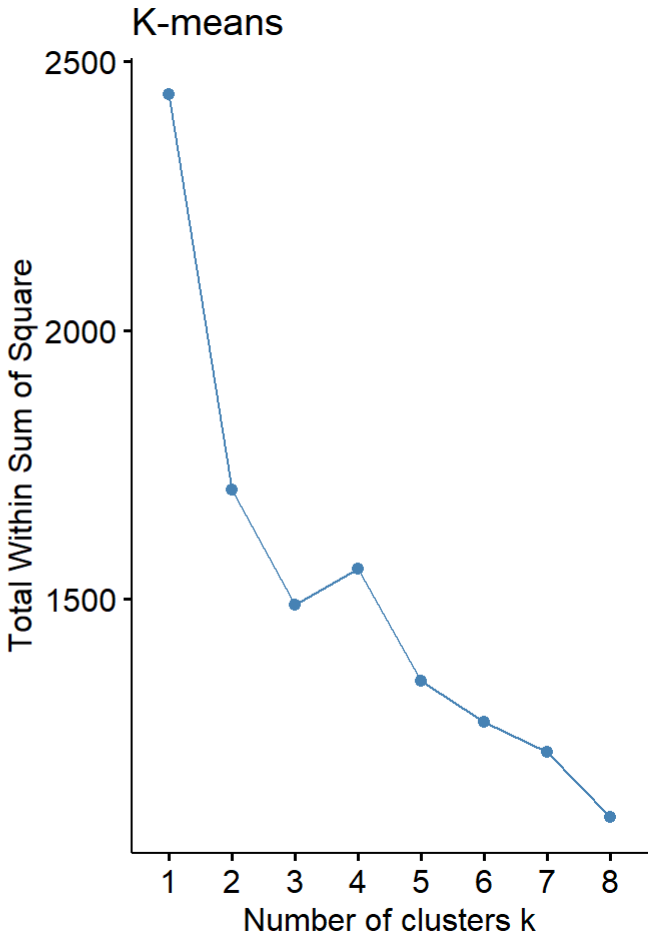
```
schools_cluster = scale(prueba_bin,center=TRUE, scale=TRUE)
midist <- get_dist(schools_cluster, stand = FALSE, method = "manhattan")
fviz_dist(midist, show_labels = TRUE, lab_size = 0.3,
          gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"))
```



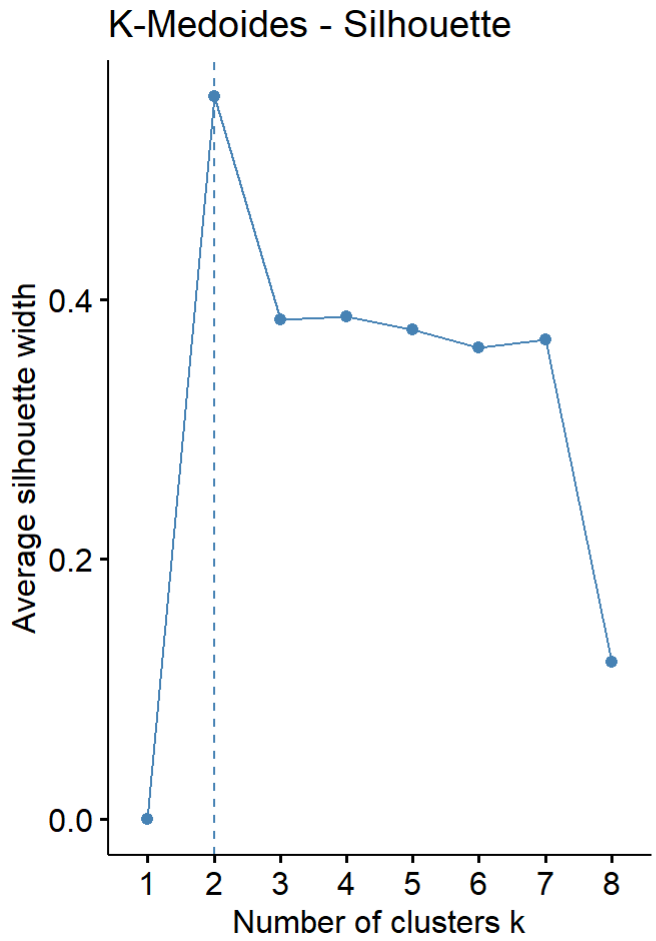
En el gráfico obtenido a partir de las distancias podemos observar posibles agrupaciones entre centros que serían los cuadrados azules que se forman a través de la diagonal principal de la matriz. El siguiente paso será realizar un método de partición, en concreto realizaremos el algoritmo **k-medias**.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

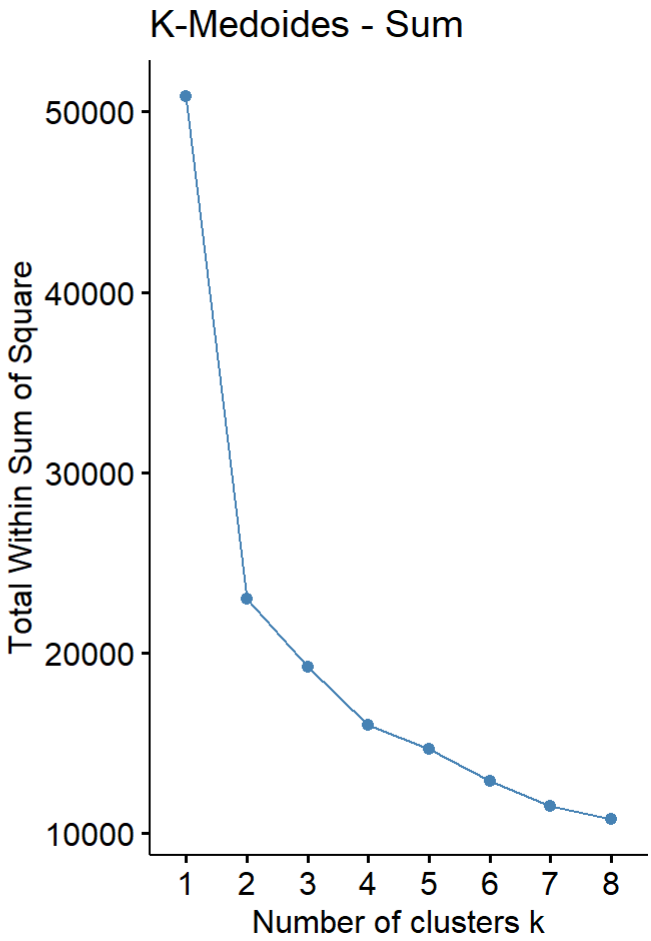
```
p1 = fviz_nbclust(x = schools_cluster, FUNcluster = kmeans, method = "silhouette",
             k.max = 8, verbose = FALSE) +
  labs(title = "K-means")
p2 = fviz_nbclust(x = schools_cluster, FUNcluster = kmeans, method = "wss",
             k.max = 8, verbose = FALSE) +
  labs(title = "K-means")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

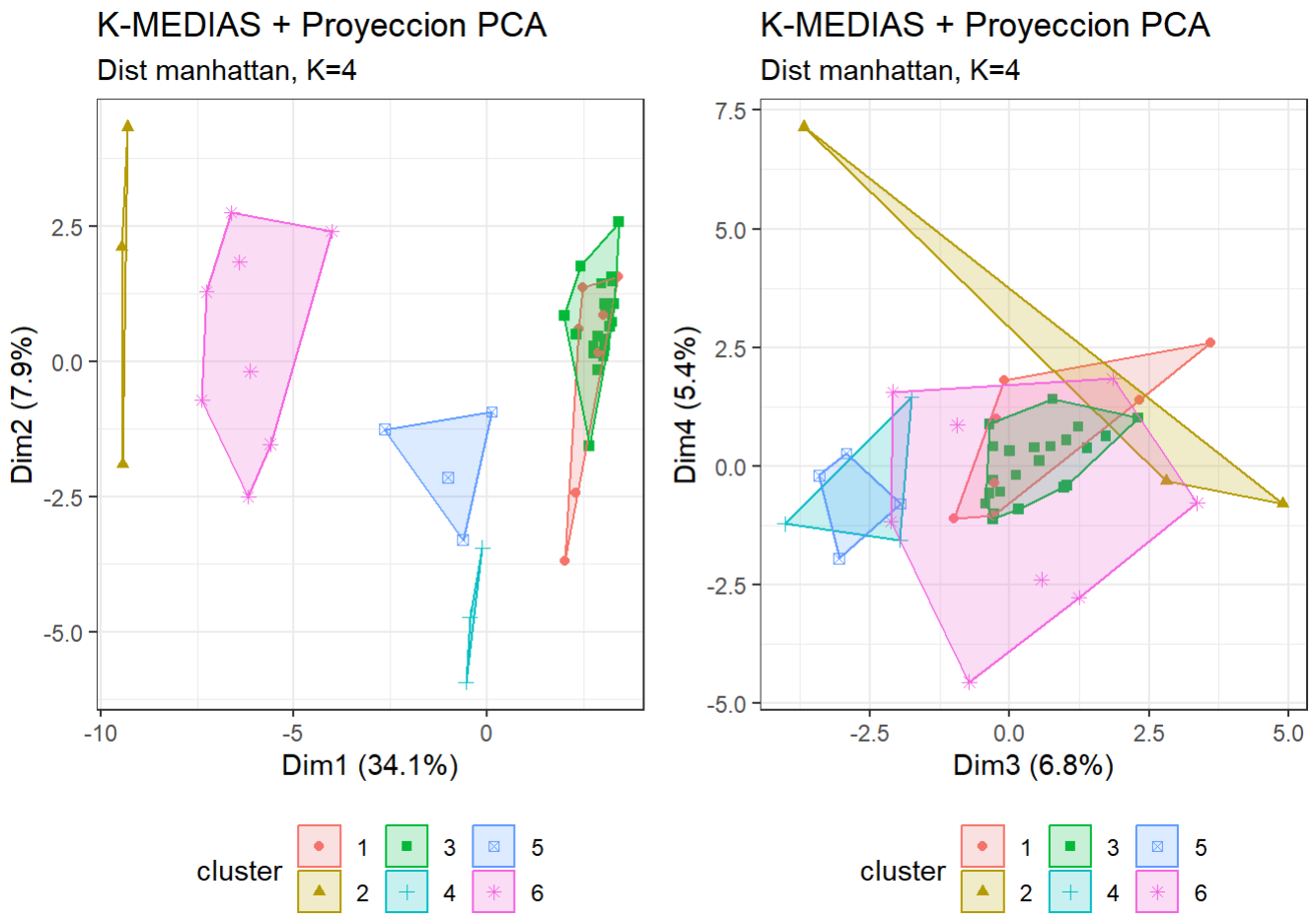




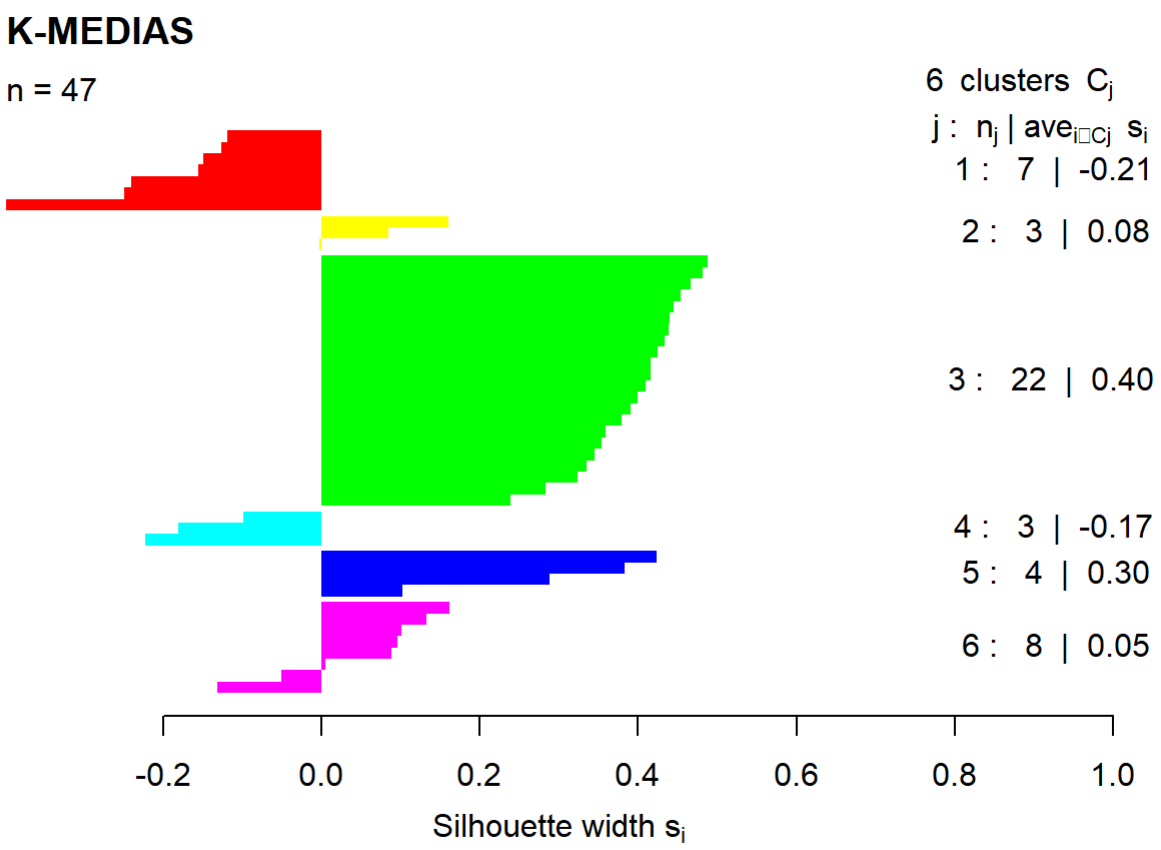
```
set.seed(100)
clust3 <- kmeans(midist, centers = 6, nstart = 20)
table(clust3$cluster)
```

```
##
##  1  2  3  4  5  6
##  7  3 22  3  4  8
```

```
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)
```

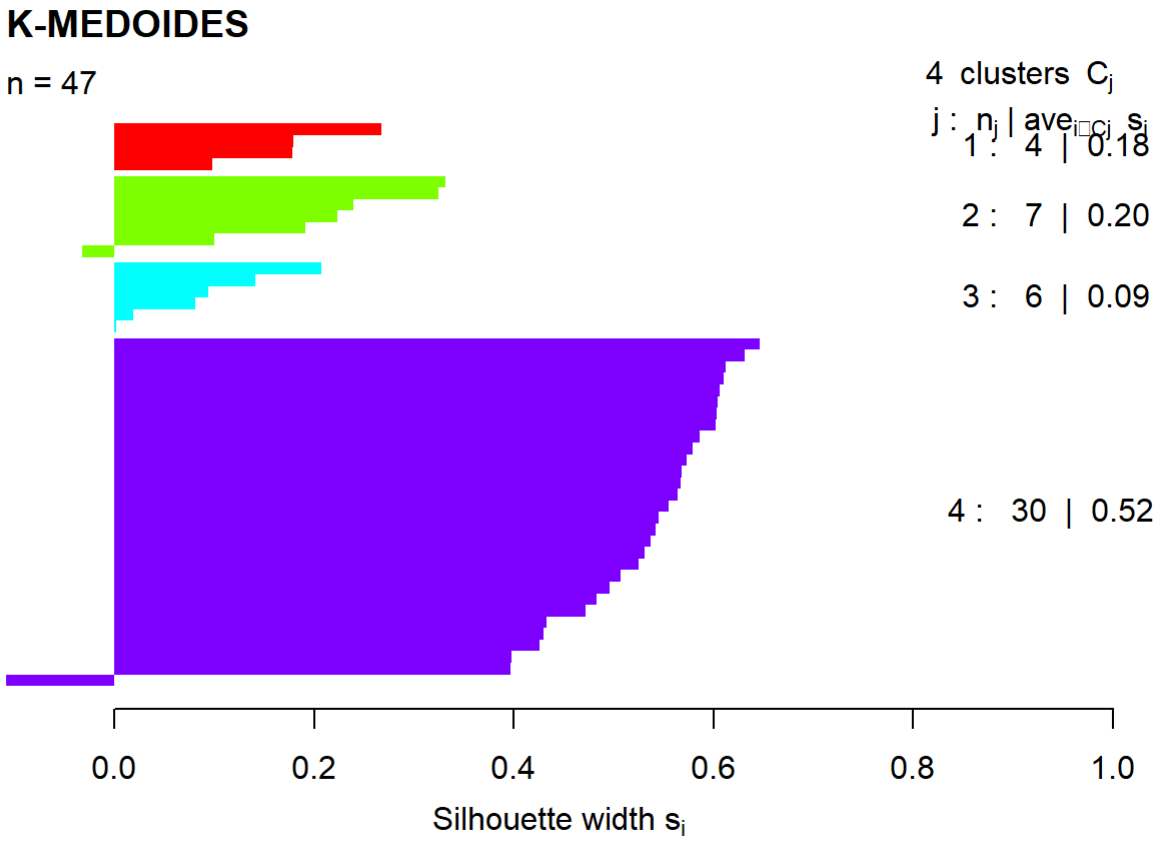



```
plot(silhouette(clust3$cluster, midist), col=rainbow(6), border=NA, main = "K-MEDIAS")
```



Average silhouette width : 0.18

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



Average silhouette width : 0.39

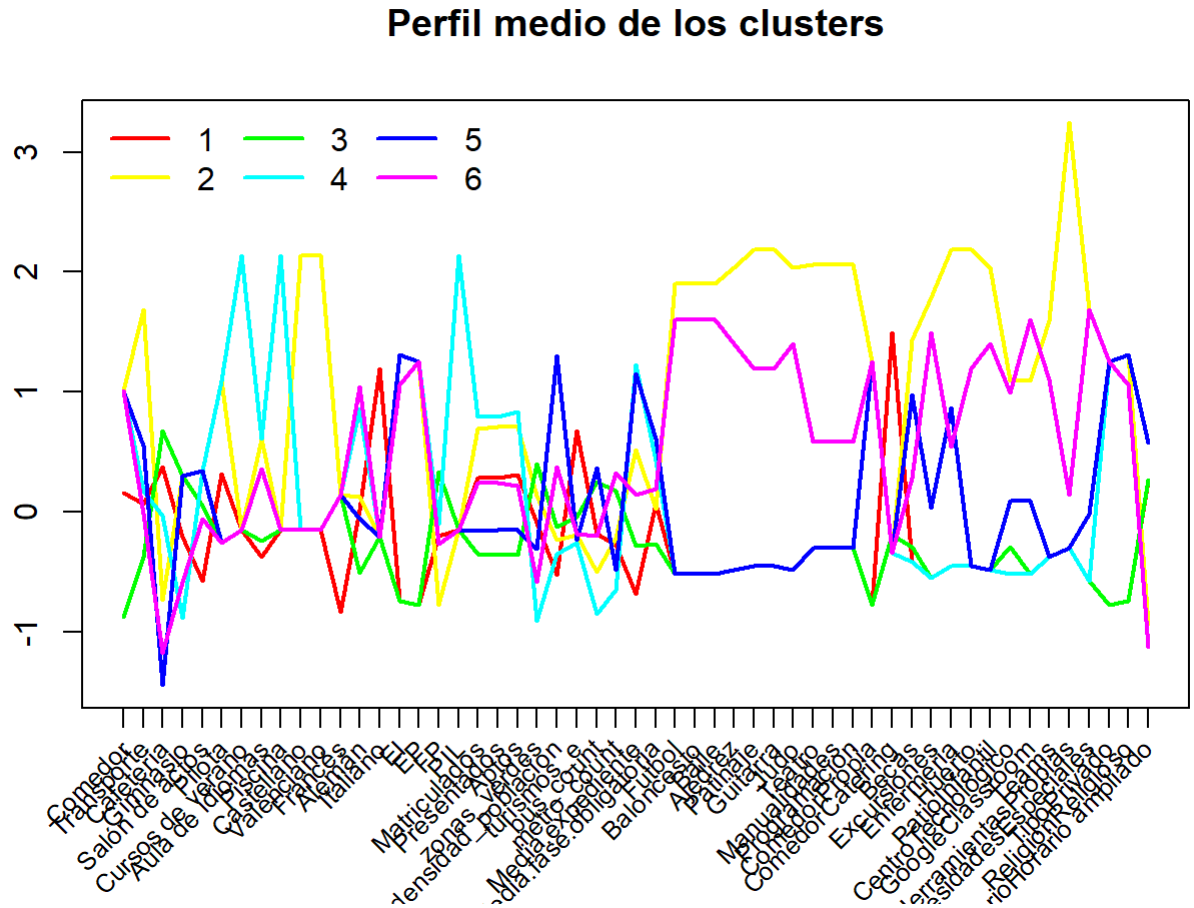
```
sil = data.frame(silhouette(clust3$cluster, midist))

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)))
```

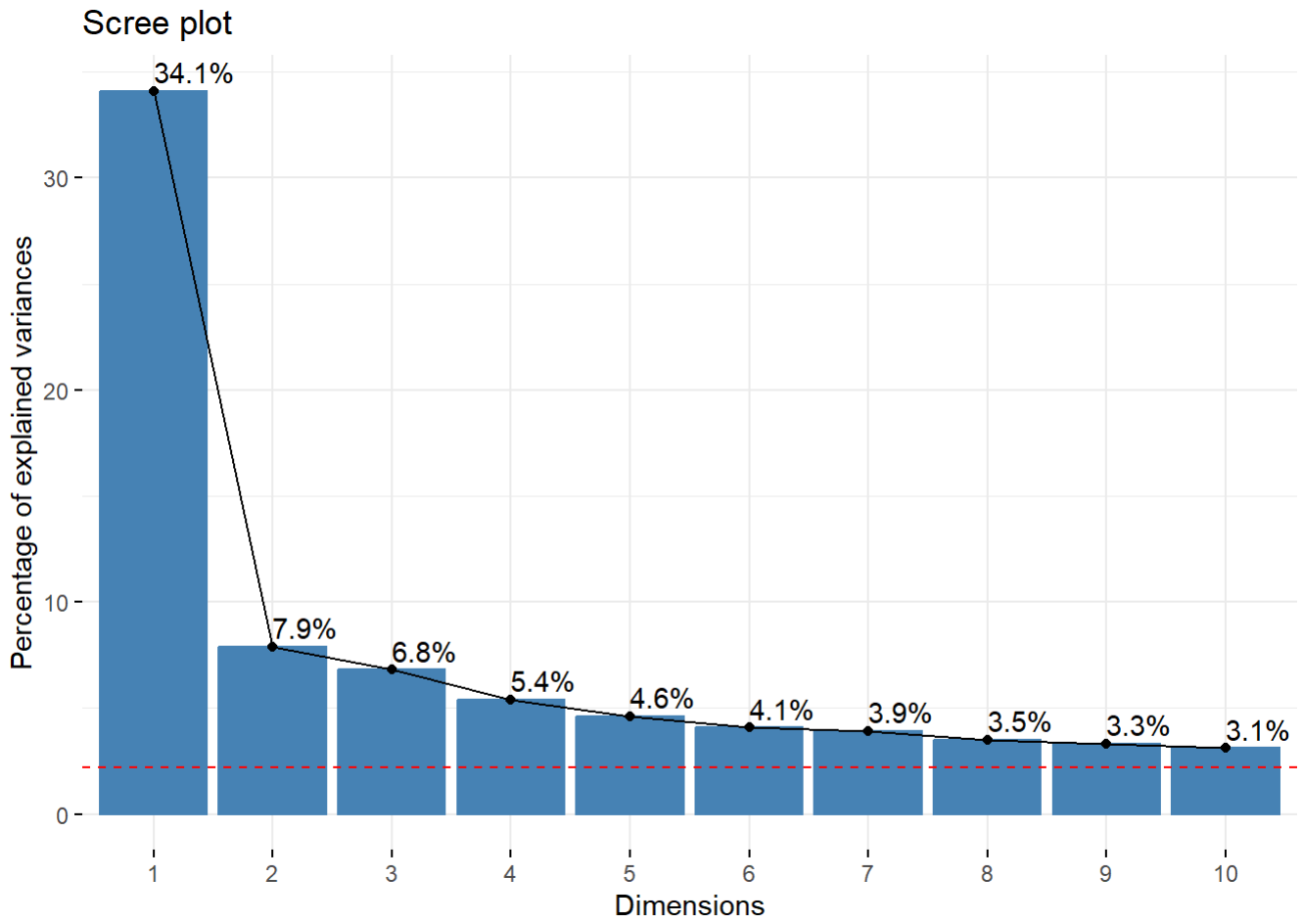
	c1	c2	c3	c4	c5	c6
Comedor	0.16	1.01	-0.88	1.01	1.01	1.01
Transporte	0.07	1.69	-0.37	0.18	0.56	-0.01
Cafetería	0.37	-0.74	0.68	-0.03	-1.44	-1.18
Gimnasio	-0.20	0.30	0.30	-0.88	0.30	-0.58
Salón de actos	-0.58	0.34	0.05	0.34	0.34	-0.06
Pilota	0.32	1.09	-0.26	1.09	-0.26	-0.26
Cursos de verano	-0.15	-0.15	-0.15	2.14	-0.15	-0.15
Aula de idiomas	-0.38	0.61	-0.24	0.61	0.36	0.36
Piscina	-0.15	-0.15	-0.15	2.14	-0.15	-0.15
Castellano	-0.15	2.14	-0.15	-0.15	-0.15	-0.15
Valenciano	-0.15	2.14	-0.15	-0.15	-0.15	-0.15
Francés	-0.83	0.15	0.15	0.15	0.15	0.15
Alemán	0.02	0.13	-0.51	0.86	-0.06	1.05
Italiano	1.19	-0.21	-0.21	-0.21	-0.21	-0.21
EI	-0.74	1.31	-0.74	1.31	1.31	1.06
EP	-0.78	1.26	-0.78	1.26	1.26	1.26
FP	-0.20	-0.78	0.33	-0.10	-0.27	-0.27
PIL	-0.15	-0.15	-0.15	2.14	-0.15	-0.15
Matriculados	0.28	0.70	-0.35	0.80	-0.15	0.24
Presentados	0.29	0.71	-0.36	0.79	-0.15	0.24
Aptos	0.31	0.71	-0.36	0.83	-0.15	0.22
zonas_verdes	-0.09	0.13	0.40	-0.91	-0.31	-0.59
densidad_poblacion	-0.53	-0.23	-0.13	-0.35	1.30	0.37
turismos_e	0.68	-0.19	-0.04	-0.26	-0.23	-0.19
bus_count	-0.19	-0.50	0.25	-0.85	0.37	-0.20

	c1	c2	c3	c4	c5	c6
metro_count	-0.28	-0.22	0.18	-0.65	-0.49	0.32
Media.expediente	-0.68	0.52	-0.28	1.23	1.15	0.14
Media.fase.obligatoria	0.06	0.03	-0.26	0.42	0.61	0.19
Futbol	-0.51	1.90	-0.51	-0.51	-0.51	1.60
Baloncesto	-0.51	1.90	-0.51	-0.51	-0.51	1.60
Baile	-0.51	1.90	-0.51	-0.51	-0.51	1.60
Ajedrez	-0.48	2.03	-0.48	-0.48	-0.48	1.40
Patinaje	-0.45	2.18	-0.45	-0.45	-0.45	1.20
Guitarra	-0.45	2.18	-0.45	-0.45	-0.45	1.20
Judo	-0.48	2.03	-0.48	-0.48	-0.48	1.40
Teatro	-0.30	2.06	-0.30	-0.30	-0.30	0.58
Manualidades	-0.30	2.06	-0.30	-0.30	-0.30	0.58
Programacion	-0.30	2.06	-0.30	-0.30	-0.30	0.58
ComedorPropia	-0.78	1.26	-0.78	1.26	1.26	1.26
ComedorCatering	1.49	-0.34	-0.20	-0.34	-0.34	-0.34
Becas	-0.41	1.44	-0.29	-0.41	0.98	0.28
Excursiones	-0.55	1.79	-0.55	-0.55	0.04	1.50
Enfermeria	-0.45	2.18	-0.45	-0.45	0.87	0.54
Huerto	-0.45	2.18	-0.45	-0.45	-0.45	1.20
PatioInfantil	-0.48	2.03	-0.48	-0.48	-0.48	1.40
CentroTecnologico	-0.51	1.10	-0.29	-0.51	0.09	1.00
GoogleClassroom	-0.51	1.10	-0.51	-0.51	0.09	1.60
Teams	-0.38	1.60	-0.38	-0.38	-0.38	1.10
HerramientasPropias	-0.30	3.24	-0.30	-0.30	-0.30	0.14
NecesidadesEspeciales	-0.58	1.69	-0.58	-0.58	-0.01	1.69
TipoPrivado	-0.78	1.26	-0.78	1.26	1.26	1.26
ReligionReligioso	-0.74	1.31	-0.74	1.31	1.31	1.06
HorarioHorario ampliado	0.26	-0.93	0.27	0.58	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")
```




```
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")
```

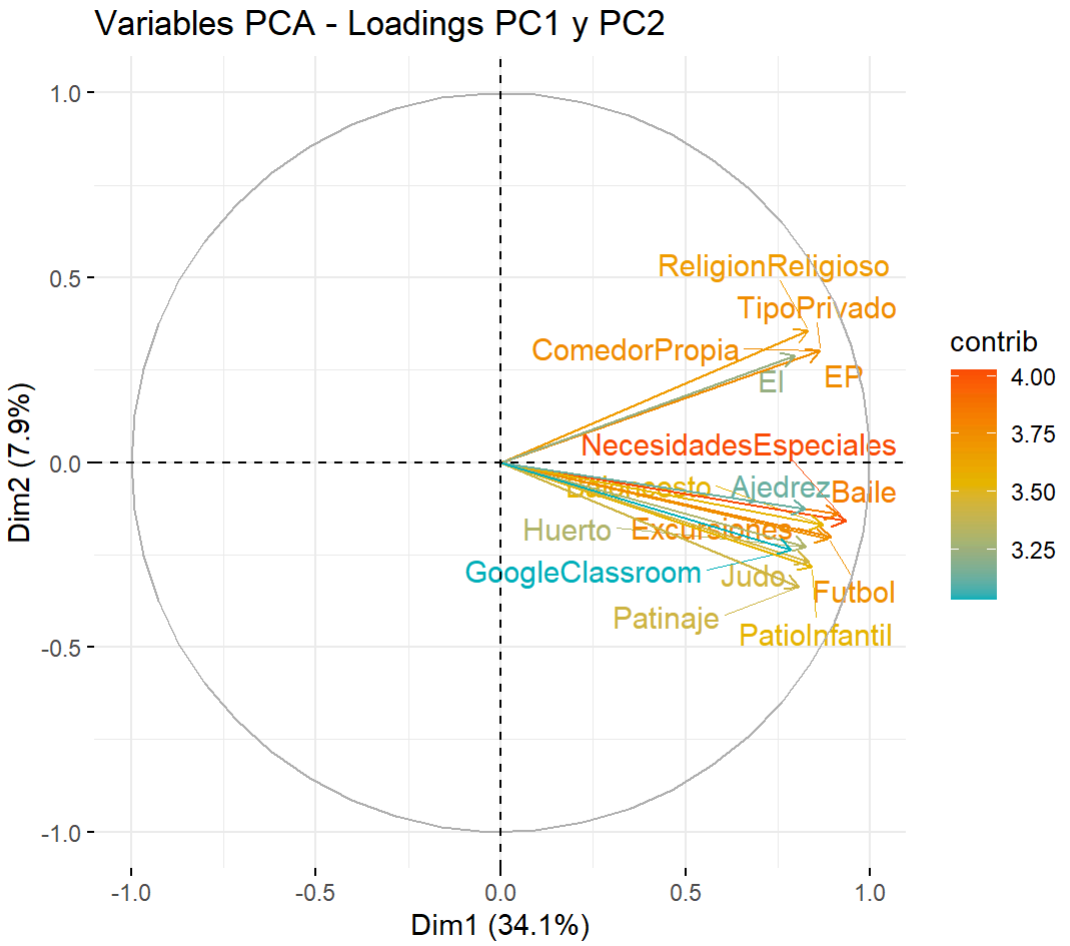



```
kable(eig.val[1:6,])
```

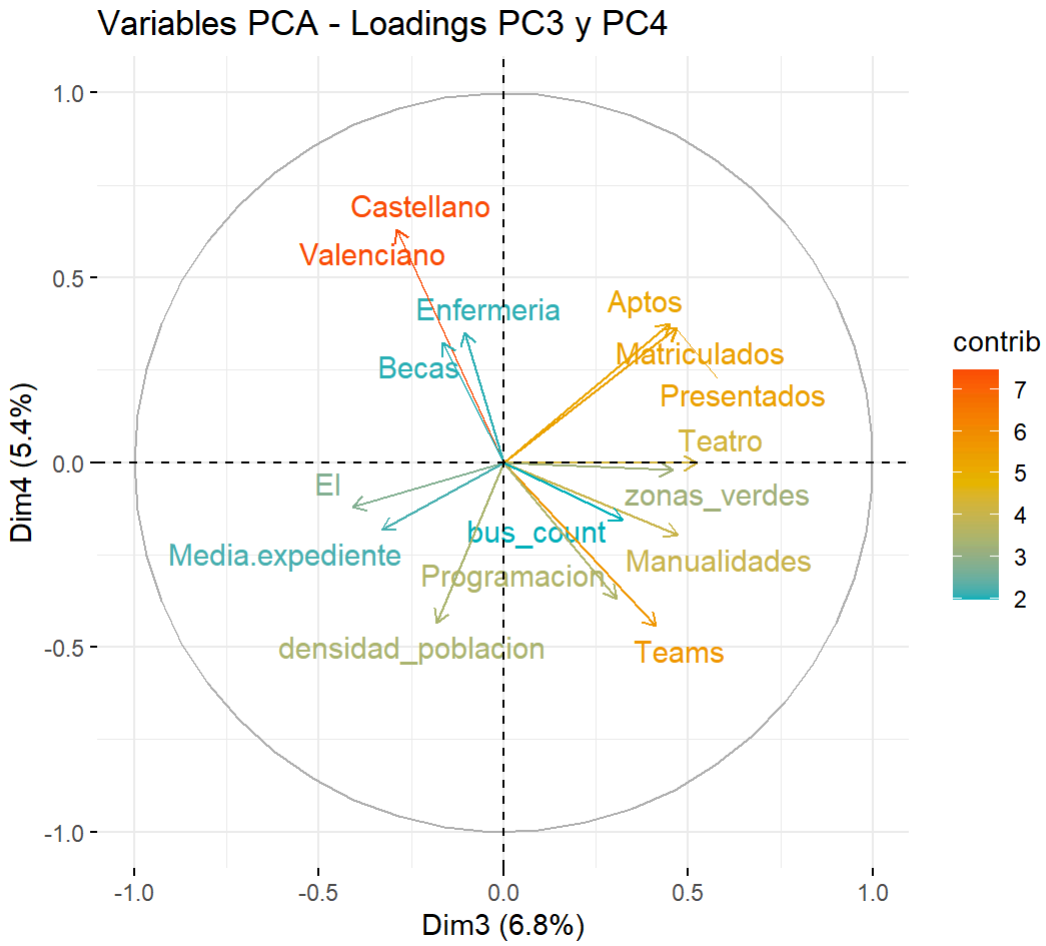
	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	18.075336	34.104408	34.10441
Dim.2	4.172568	7.872770	41.97718
Dim.3	3.604425	6.800802	48.77798
Dim.4	2.850271	5.377869	54.15585
Dim.5	2.432688	4.589978	58.74583
Dim.6	2.160372	4.076174	62.82200

```
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')
```



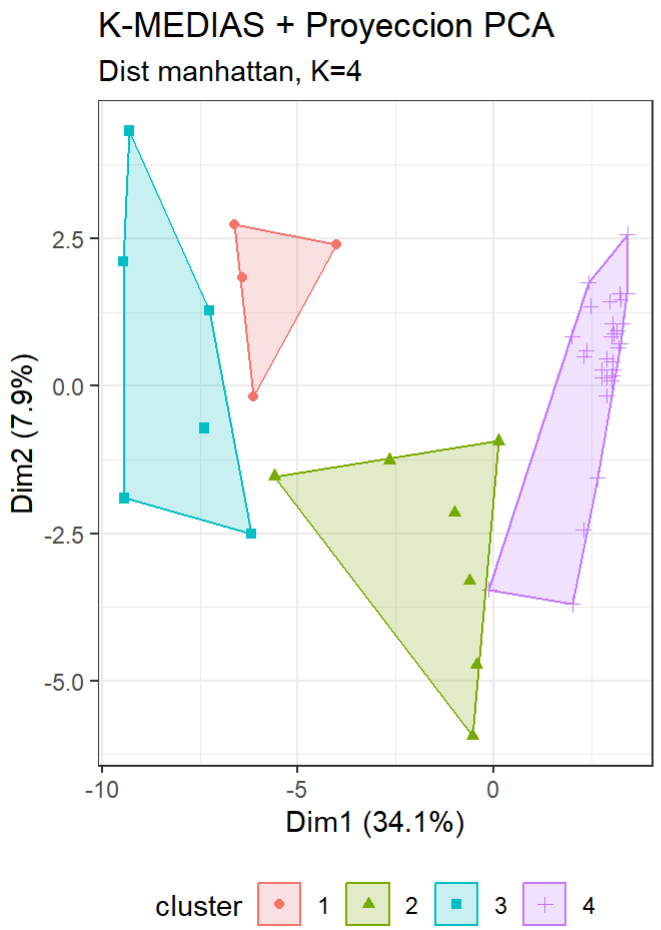

```
n_clusters = as.data.frame(clust3$cluster)
```

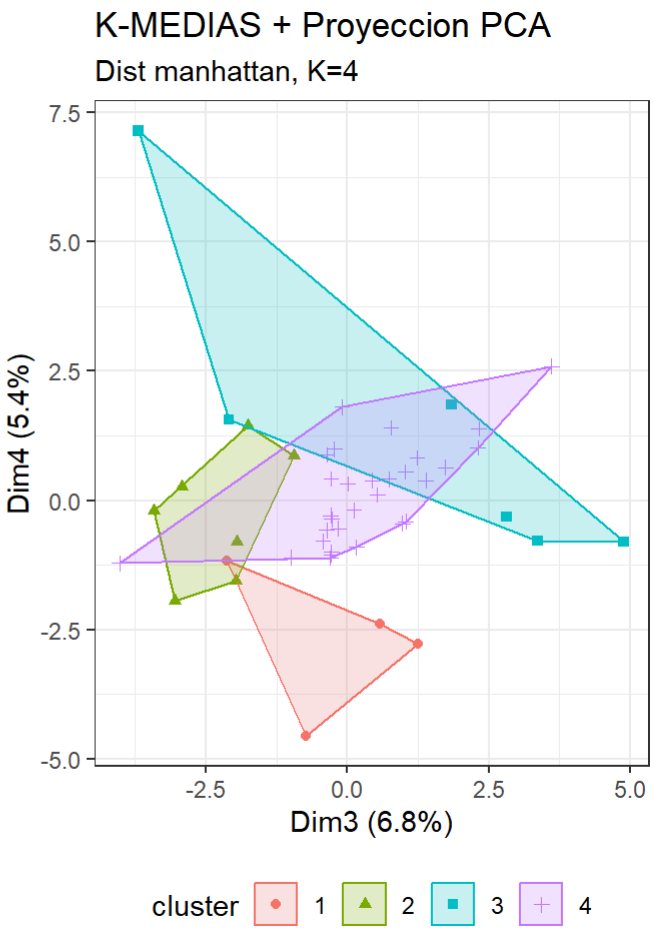
```
#library(xlsx)
#write.xlsx(prueba_bin, "DatosSecundaria.xlsx", row.names = FALSE)
```

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

```
##
##  1  2  3  4
##  4  7  6 30
```

```
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)
```





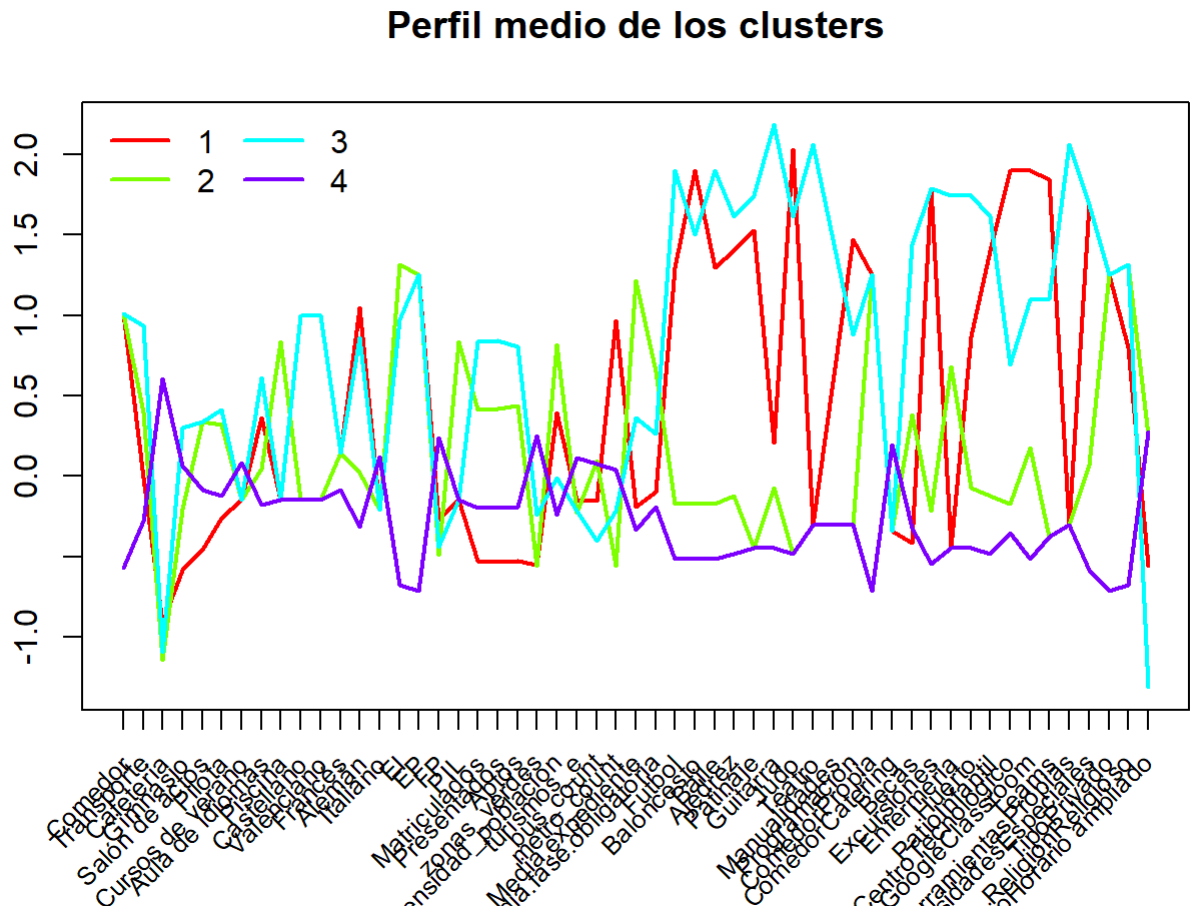

```
misclust = factor(clust4$cluster)

mediasCluster = aggregate(schools_cluster, by = list("cluster" = misclust), mean)[,-1]
rownames(mediasCluster) = paste0("c",1:4)
kable(t(round(mediasCluster,2)))
```

	c1	c2	c3	c4
Comedor	1.01	1.01	1.01	-0.57
Transporte	-0.01	0.39	0.93	-0.28
Cafetería	-0.91	-1.14	-1.09	0.61
Gimnasio	-0.58	-0.20	0.30	0.07
Salón de actos	-0.46	0.34	0.34	-0.09
Pilota	-0.26	0.32	0.42	-0.12
Cursos de verano	-0.15	-0.15	-0.15	0.08
Aula de idiomas	0.36	0.05	0.61	-0.18
Piscina	-0.15	0.83	-0.15	-0.15
Castellano	-0.15	-0.15	1.00	-0.15
Valenciano	-0.15	-0.15	1.00	-0.15
Francés	0.15	0.15	0.15	-0.08

	c1	c2	c3	c4
Alemán	1.05	0.02	0.86	-0.32
Italiano	-0.21	-0.21	-0.21	0.12
EI	1.31	1.31	0.97	-0.68
EP	1.26	1.26	1.26	-0.71
FP	-0.27	-0.49	-0.44	0.24
PIL	-0.15	0.83	-0.15	-0.15
Matriculados	-0.53	0.42	0.84	-0.19
Presentados	-0.53	0.42	0.84	-0.20
Aptos	-0.52	0.44	0.80	-0.19
zonas_verdes	-0.55	-0.55	-0.24	0.25
densidad_poblacion	0.39	0.82	-0.01	-0.24
turismos_e	-0.14	-0.23	-0.21	0.12
bus_count	-0.15	0.10	-0.40	0.08
metro_count	0.97	-0.55	-0.22	0.04
Media.expediente	-0.19	1.22	0.36	-0.33
Media.fase.obligatoria	-0.10	0.65	0.27	-0.19
Futbol	1.30	-0.17	1.90	-0.51
Baloncesto	1.90	-0.17	1.50	-0.51
Baile	1.30	-0.17	1.90	-0.51
Ajedrez	1.40	-0.12	1.61	-0.48
Patinaje	1.53	-0.45	1.75	-0.45
Guitarra	0.21	-0.07	2.18	-0.45
Judo	2.03	-0.48	1.61	-0.48
Teatro	-0.30	-0.30	2.06	-0.30
Manualidades	0.58	-0.30	1.47	-0.30
Programacion	1.47	-0.30	0.88	-0.30
ComedorPropia	1.26	1.26	1.26	-0.71
ComedorCatering	-0.34	-0.34	-0.34	0.19
Becas	-0.41	0.38	1.44	-0.32
Excursiones	1.79	-0.21	1.79	-0.55
Enfermeria	-0.45	0.68	1.75	-0.45
Huerto	0.87	-0.07	1.75	-0.45

	c1	c2	c3	c4
PatioInfantil	1.40	-0.12	1.61	-0.48
CentroTecnologico	1.90	-0.17	0.69	-0.35
GoogleClassroom	1.90	0.18	1.10	-0.51
Teams	1.84	-0.38	1.10	-0.38
HerramientasPropias	-0.30	-0.30	2.06	-0.30
NecesidadesEspeciales	1.69	0.07	1.69	-0.58
TipoPrivado	1.26	1.26	1.26	-0.71
ReligionReligioso	0.80	1.31	1.31	-0.68
HorarioHorario ampliado	-0.56	0.26	-1.31	0.28

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



-Cluster 1 (rojo): Comedor con cocina propia. Usualmente sin gimnasio, salon de actos, pilota ni piscina. Pero si que presenta aula de idiomas ofreciendo frances y aleman (sobretodo). Centros desde infantil hasta bachiller, pero sin FP. Usualmente con pocos alumnos matriculados. Se sitúan con una densidad de población

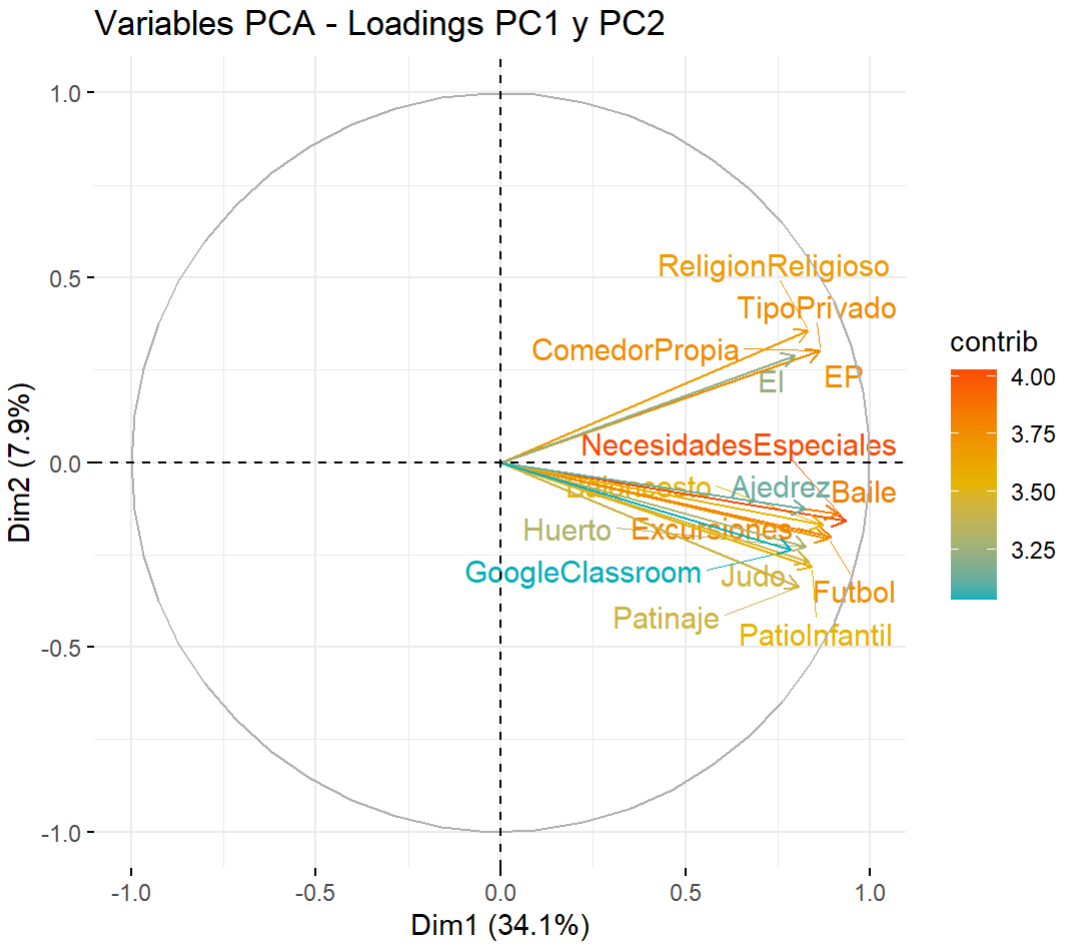
alta (2 más alta). Zonas bien comunicadas por metro y no presenta transporte privado.Media de Bachiller baja (2 peor) pero mejora un poco en la PAU (2 peor). Alto numero de extraescolares donde destacan: Judo, Programación y Baloncesto. Pocas becas, muchas excursiones. No suelen tener enfermería pero si Centro Tecnológico. Además de uso de herramientas tecnológicas como: Teams o Google Classroom. Capacidad para tratar alumnos con necesidades especiales. Generalmente privados (laicos+religiosos). Horario de mañana generalmente.

-Cluster 2 (verde): Comedor con cocina propia. Instalaciones como: salón de actos, pilota, piscina (no cursos de verano). Unicamente línea en castellano y en valenciano. Ofreciendo francés como 3/4 idioma. Centros desde infantil hasta bachiller. Presenta PIL. Cluster con el segundo número de matriculados más elevado. Se sitúan en barrios con la densidad de población más alta. Comunicado por bus aunque algunos también disponen de transporte privado. Media de expediente más alta de los clústers pero disminuye en la nota PAU (2 mejor). No presenta un gran número de extraescolares. Algunos centros presentan becas. Centros privados y religiosos, con horarios tanto ampliados como solo de mañanas.

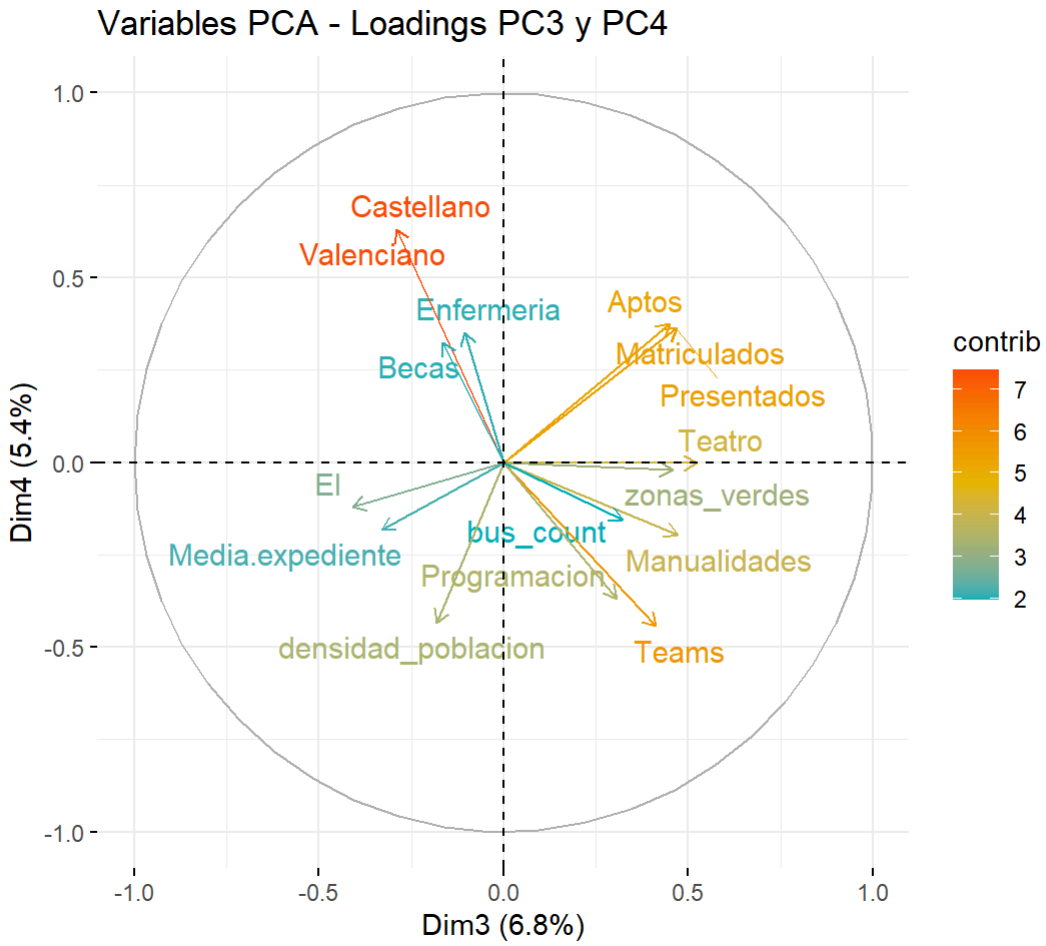
-Cluster 3 (azul): La mayoría de centros de este clúster presentan gimmnasio, salon de actos y aula de idiomas (ofrecen linea en valenciano y castellano además de francés o alemán como 3/4 idioma). Son centros desde infantil hasta bachiller, con un elevado número de matriculados. Además en cuanto a nota media de bachiller es la segunda más alta como en la PAU (pero baja). Barrios con poca población (2 más baja). Mal comunicado (ni bus ni metro), pero presentan transporte privado. Comedor con cocina propia. Gran número de extraescolares destacando: futbol, ajedrez o guitarra. Centros que presentan herramientas propias y con capacidad de tratar alumnos con necesidades especiales. Centros privados y religiosos con horarios de mañana.

-Cluster 4 (morado):No presentan comedor pero si cafetería (comida de catering). Suelen presentar gimnasio y realizar cursos de verano (no suelen tener piscina). Destacan en cuanto idiomas por haber centros que ofrecen italiano como 3/4 idioma. Son centros de secundaria y bachiller unicamente, además de FP. Numero de matriculados segundo más bajo, las notas de bachiller y PAU (se mejora) son las peores. Se sitúan barrios poco poblados y con algunas conexiones de bus y metro (no transporte privado). No muchas extraescolares ni servicios. Además de ser centros públicos y laicos con horarios tanto ampliados como solo de mañanas.

```
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')
```


Variables PCA - Loadings PC3 y PC4
Castellano
Valenciano
Enfermeria
Becas
Aptos
Matriculados
Presentados
Teatro
El
zonas_verdes
bus_count
Media.expediente
Programacion
Manualidades
densidad_poblacion
Teams
contrib
7
6
5
4
3
2
1.0
0.5
0.0
-0.5
-1.0
Dim4 (5.4%)
Dim3 (6.8%)