Barrios_RMCword

Juana M

```
knitr::opts_chunk$set(echo = TRUE)
```

Importar los datos

```
BARRIOS <- read_excel("C:/Users/reven/OneDrive/Desktop/Master Big data/Clases/Cluster/Cluster/BARRIOS.xlsx")
datos <- as.data.frame(BARRIOS)
rownames(datos) <- datos[,1]
dat_B <- datos[,-1]
```

Un primer resumen

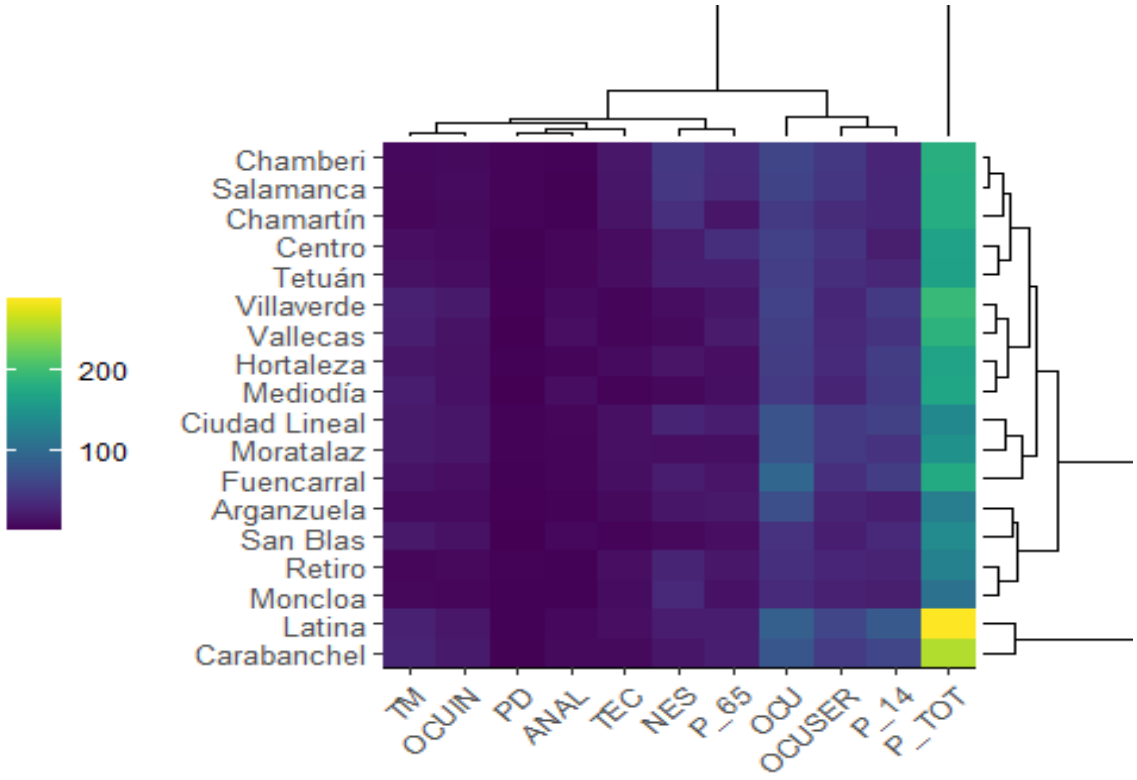
```
tabla1<-summary(dat_B)
knitr::kable(tabla1, caption = "Tabla resumen de las variables")
```

Tabla resumen de las variables

P_TOT	P_14	P_65	ANAL	NES	OCU	OCUIN	OCUSER	TEC	PD	TM
Min. :108.4	Min. :23.30	Min. :10.10	Min. : 0.800	Min. : 5.50	Min. :34.60	Min. : 5.30	Min. :24.10	Min. : 2.900	Min. :0.200	Min. : 4.700
1st Qu.:139.8	1st Qu.:29.68	1st Qu.:13.95	1st Qu.: 1.850	1st Qu.:11.45	1st Qu.:49.27	1st Qu.: 7.60	1st Qu.:30.25	1st Qu.: 7.275	1st Qu.:0.650	1st Qu.: 6.675
Median :169.8	Median :36.45	Median :17.60	Median : 4.100	Median :21.50	Median :55.35	Median :11.25	Median :36.85	Median : 9.200	Median :1.350	Median :15.050
Mean :171.7	Mean :40.56	Mean :19.90	Mean : 4.472	Mean :22.07	Mean :59.67	Mean :11.66	Mean :38.18	Mean : 9.172	Mean :1.356	Mean :15.150
3rd Qu.:182.0	3rd Qu.:50.80	3rd Qu.:23.40	3rd Qu.: 6.000	3rd Qu.:28.18	3rd Qu.:72.62	3rd Qu.:15.78	3rd Qu.:46.08	3rd Qu.:11.475	3rd Qu.:2.050	3rd Qu.:22.100
Max. :289.5	Max. :79.50	Max. :38.10	Max. :10.300	Max. :47.20	Max. :95.60	Max. :19.40	Max. :59.80	Max. :17.100	Max. :2.800	Max. :28.200

Creamos un mapa de calor interactivo con las filas ordenadas de forma que estén juntas las mas parecidas. Tenemos una primera aproximación

```
ggheatmap(dat_B, seriate = "mean" )
```



#Calculamos las distancias con los valores sin estandarizar #Mostramos las primeras seis filas dela matriz de distancias

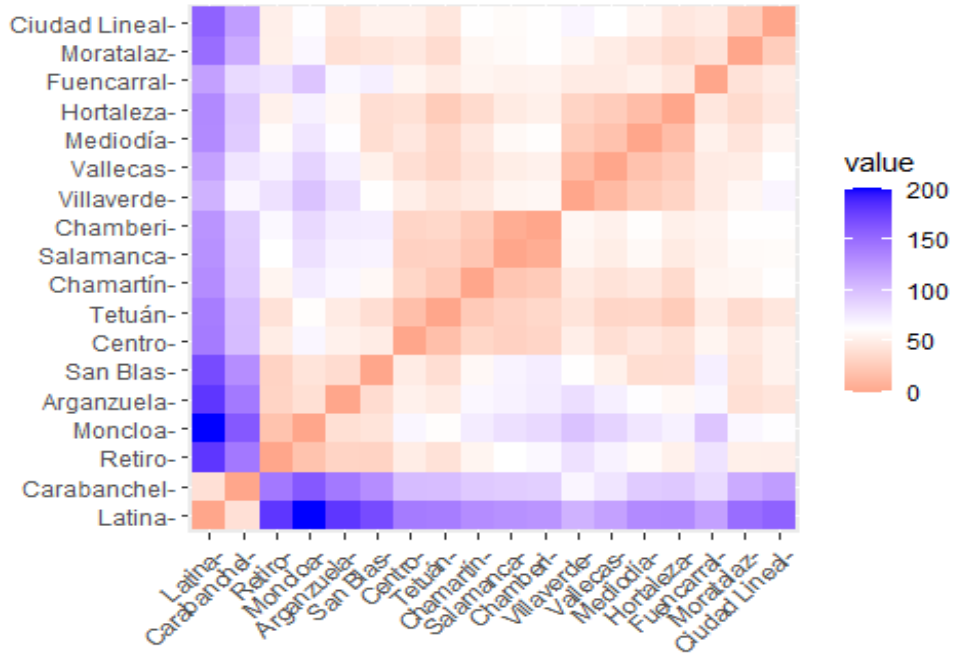
```
d <- dist(dat_B, method = "euclidean") # distance matrix
d6<-as.matrix(d)[1:6, 1:6]
knitr::kable(d6, digits =2,caption = "Distancias")
```

Distancias

	Centro	Arganzuela	Retiro	Salamanca	Chamartín	Tetuán
Centro	0.00	53.76	50.36	30.70	34.42	17.79
Arganzuela	53.76	0.00	32.79	71.28	67.98	49.11
Retiro	50.36	32.79	0.00	63.78	56.31	42.76
Salamanca	30.70	71.28	63.78	0.00	22.66	31.28
Chamartín	34.42	67.98	56.31	22.66	0.00	25.51
Tetuán	17.79	49.11	42.76	31.28	25.51	0.00

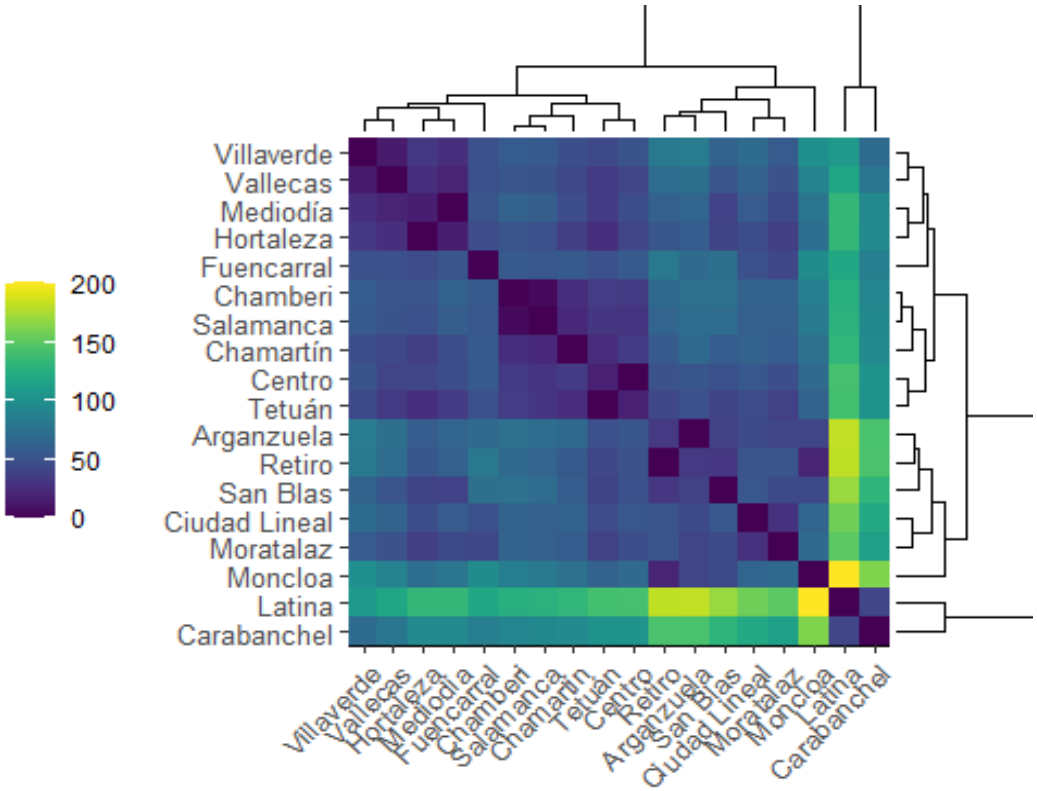
#Representamos gráficamente la matriz de distancias

```
fviz_dist(d, show_labels = TRUE)
```



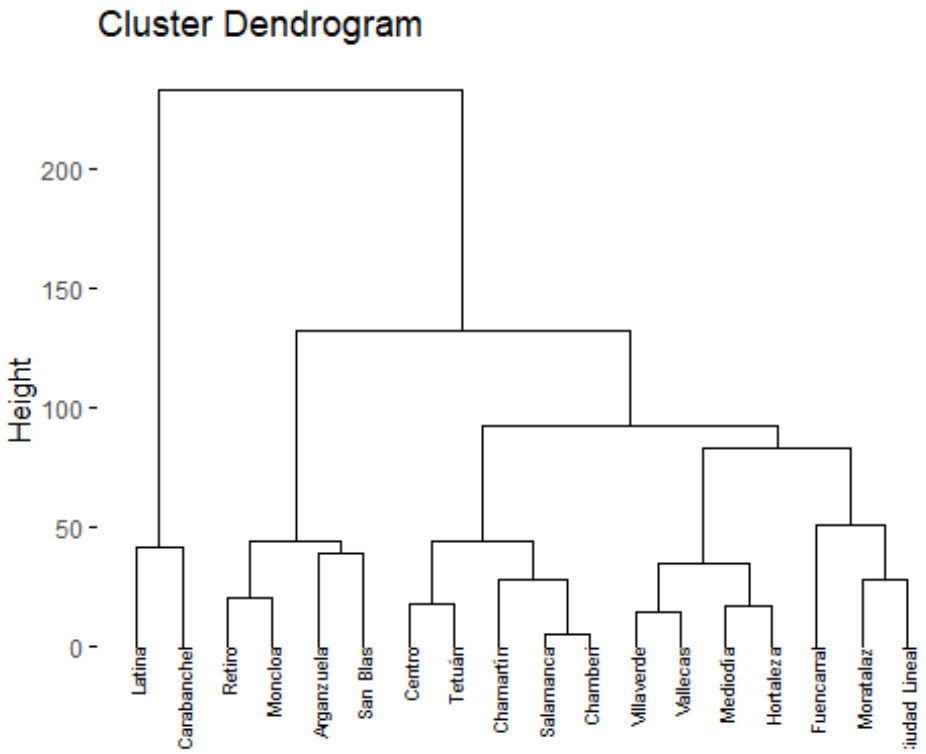
#Reordenamos para agrupar las observaciones que están más próximas y visualizar los posibles clusters

```
ggheatmap(as.matrix(d), seriate="mean")
```



#Agrupamos las observaciones según el criterio de ward #Dibujamos el dendograma correspondiente

```
res.hc <- hclust(d, method="ward.D2")

fviz_dend(res.hc, cex = 0.5)
```



Standardize the data # Show the first 6 rows

```
datos_ST <- scale(dat_B)
```

#Calculamos las distancias con los valores estandarizados

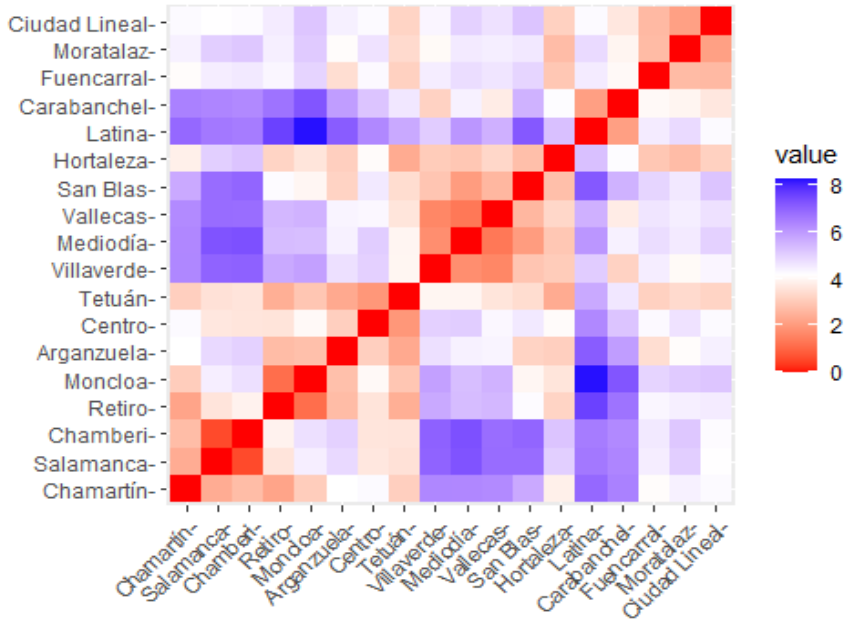
```
d_st <- dist(datos_ST, method = "euclidean") # distance matrix
d_st6<-as.matrix(d_st)[1:6, 1:6]
knitr::kable(d_st6, digits =2,caption = "Distancias")
```

Distancias

	Centro	Arganzuela	Retiro	Salamanca	Chamartín	Tetuán
Centro	0.00	3.15	3.60	3.68	4.30	1.95
Arganzuela	3.15	0.00	2.74	4.88	4.20	2.34
Retiro	3.60	2.74	0.00	3.62	2.23	2.45
Salamanca	3.68	4.88	3.62	0.00	2.40	3.55
Chamartín	4.30	4.20	2.23	2.40	0.00	3.15
Tetuán	1.95	2.34	2.45	3.55	3.15	0.00

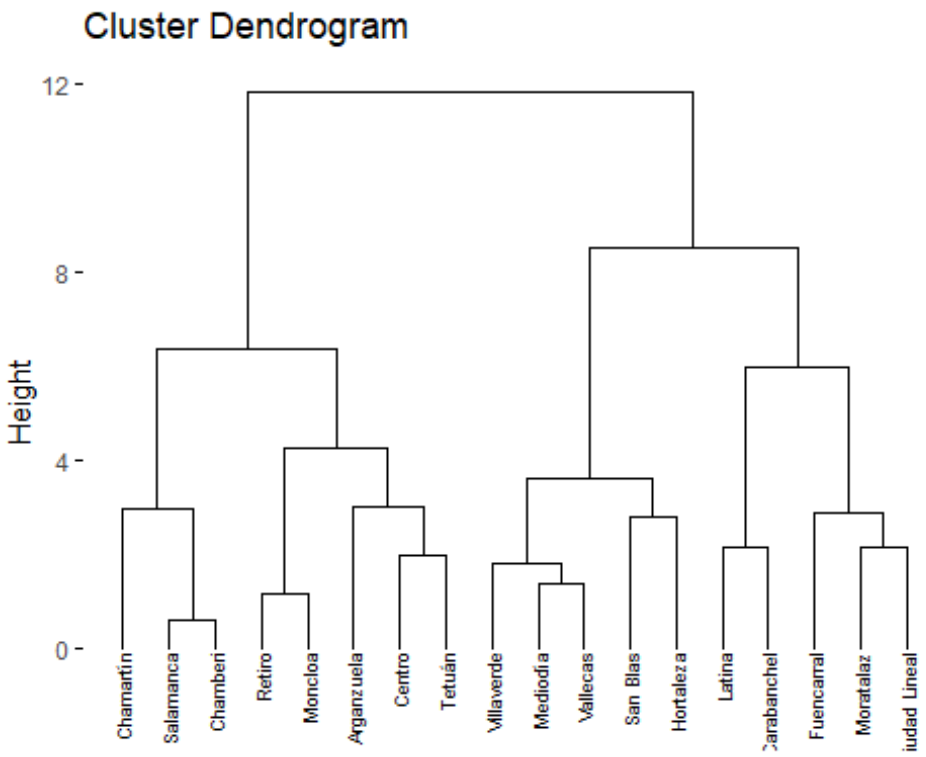
#Visualizamos

```
fviz_dist(d_st)
```



#Hacemos el cluster jerárquico con las distancias entre los datos estandarizados

```
res.hc_st <- hclust(d_st, method="ward.D2")
fviz_dend(res.hc_st, cex = 0.5)
```



Decidimos hacer cuatro clusters

```
grp <- cutree(res.hc_st, k = 4)
head(grp, n = 4)

##     Centro Arganzuela     Retiro  Salamanca
##          1          1          1          2
```

Número de miembros en cada cluster

```
knitr::kable(table(grp), caption = "Número de individuos por cluster")
```

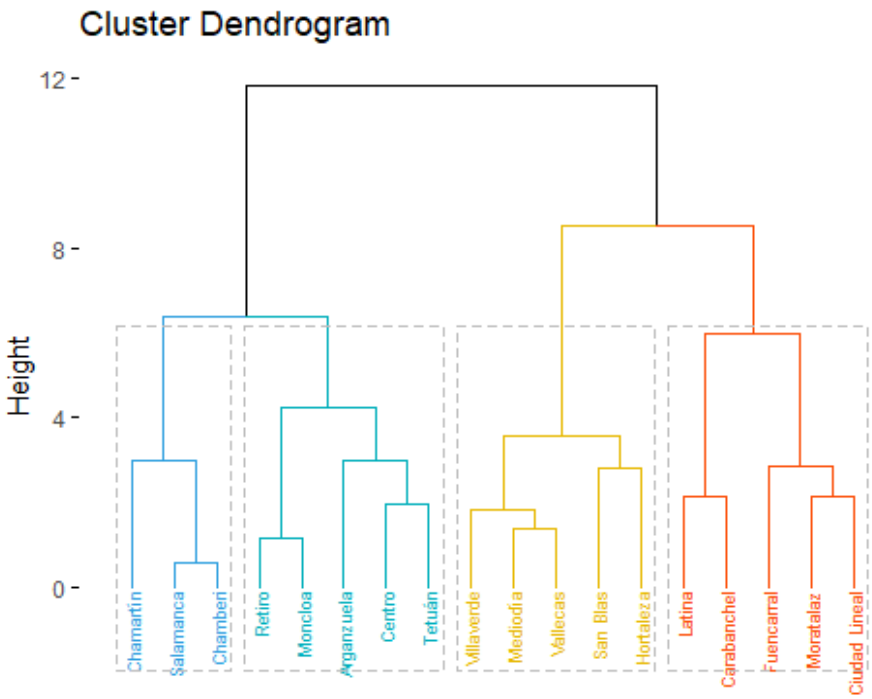
Número de individuos por cluster

grp	Freq
1	5
2	3
3	5
4	5

Mostramos los nombres de los individuos en el cluster 1

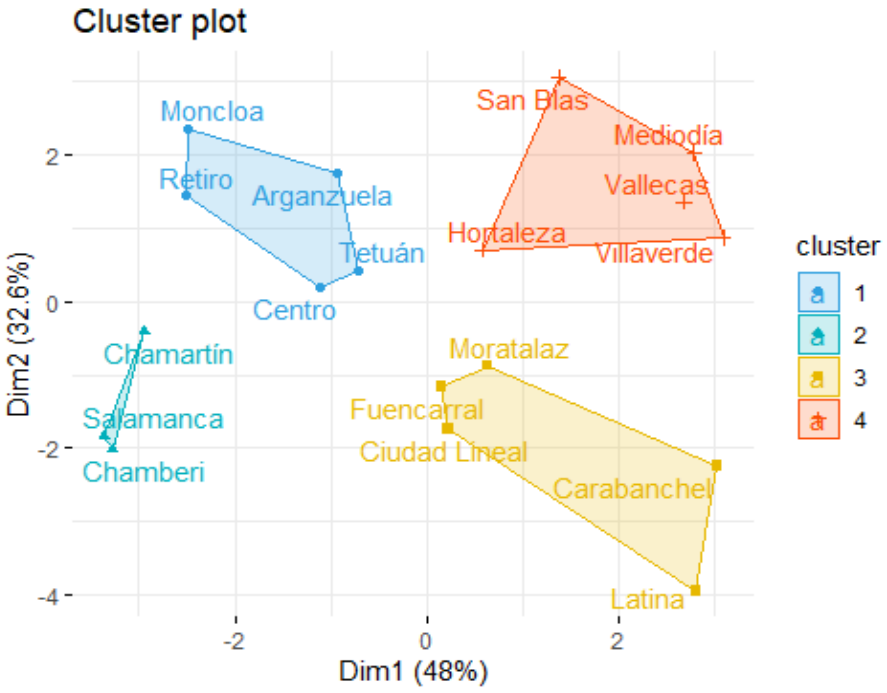
Representamos el dendrograma marcando los cluster con color

```
rownames(dat_B)[grp == 1]

## [1] "Centro"     "Arganzuela" "Retiro"     "Tetuán"     "Moncloa"


fviz_dend(res.hc_st, k = 4, # Cut in four groups
          cex = 0.5, # label size
          k_colors = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
          color_labels_by_k = TRUE, # color labels by groups
          rect = TRUE) # Add rectangle around groups
```



Visualizamos los clusters

```
fviz_cluster(list(data = datos_ST, cluster = grp), palette = c("#2E9FDF",
"#00AFBB", "#E7B800", "#FC4E07"), ellipse.type = "convex", repel = TRUE,
show.clust.cent = FALSE, ggtheme = theme_minimal())
```

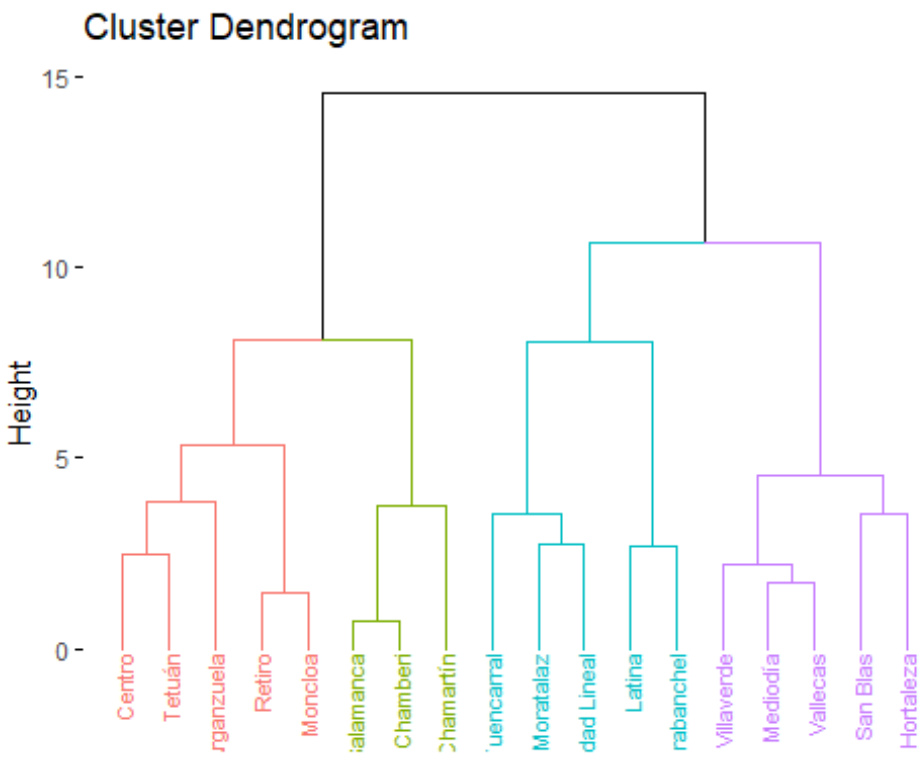


Podemos realizar los pasos anteriores a las representaciones con la siguiente función:

```
res.agnes <- agnes(x =dat_B, stand = TRUE, metric ="euclidean", #
distancia entre individuos   method = "ward") # distancia entre clusters
```

Representamos los resultados

```
fviz_dend(res.agnes, cex = 0.6, k = 4)
```

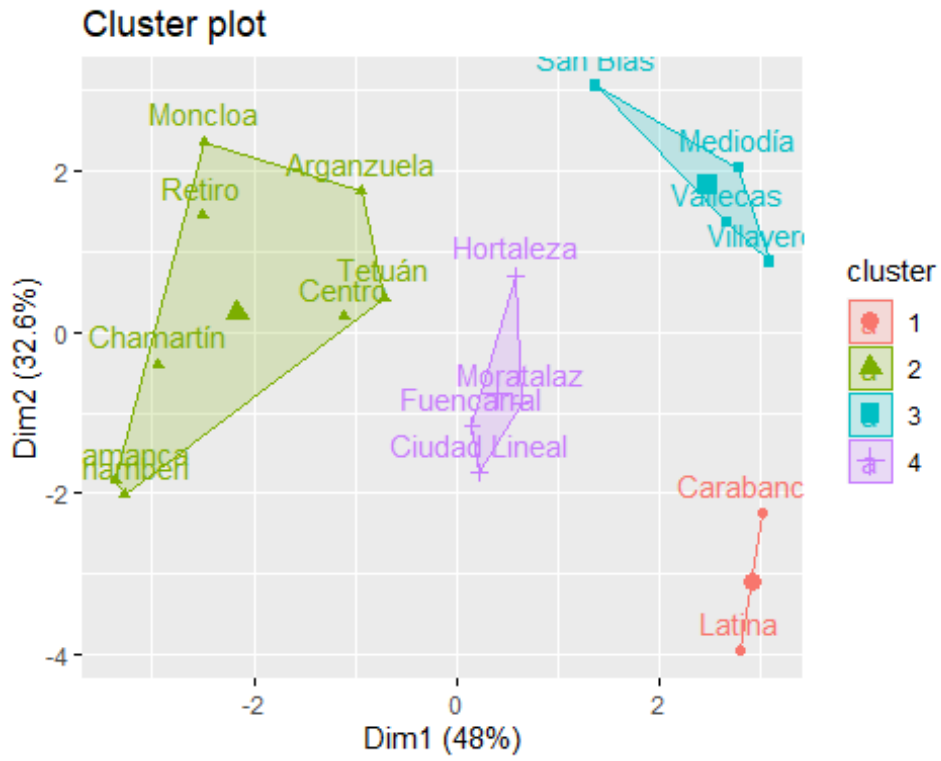


Vamos a hacer un cluster no jerárquico con el mismo número de clusters. Nos aseguramos que tenemos todos la misma semilla

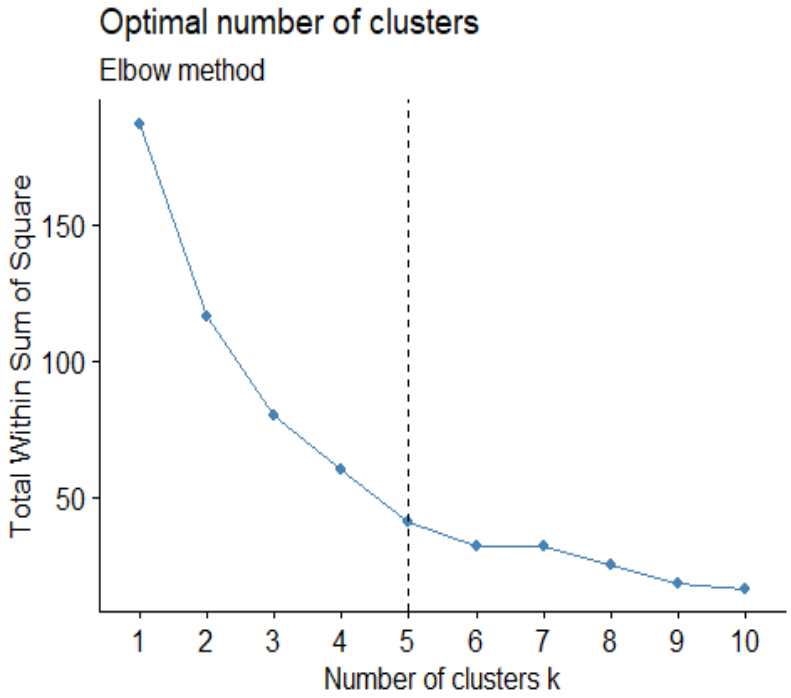
```
RNGkind(sample.kind = "Rejection")
set.seed(1234)
```

Compute k-means

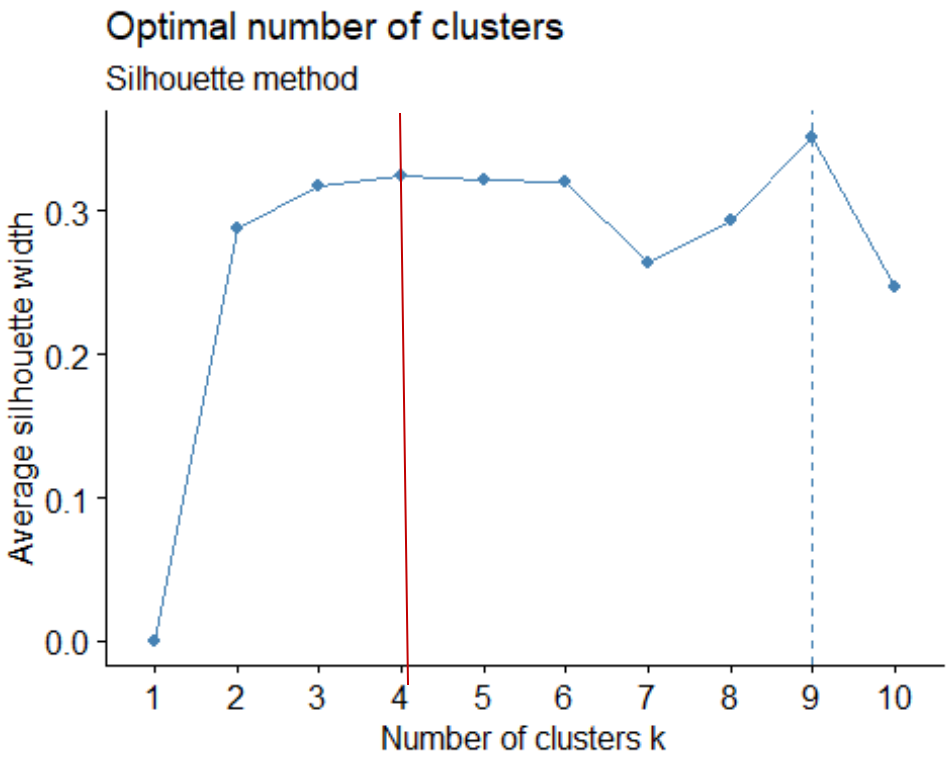
```
km.res <- kmeans(datos_ST, 4)


fviz_cluster(km.res, datos_ST)
```




```
#Determinación del número óptimo de clusters
# Elbow method
fviz_nbclust(datos_ST, kmeans, method = "wss") + geom_vline(xintercept
=5, linetype = 2)+ labs(subtitle = "Elbow method")
```



```r
# Silhouette method
fviz_nbclust(datos_ST, kmeans, method = "silhouette")+
  labs(subtitle = "Silhouette method")
```

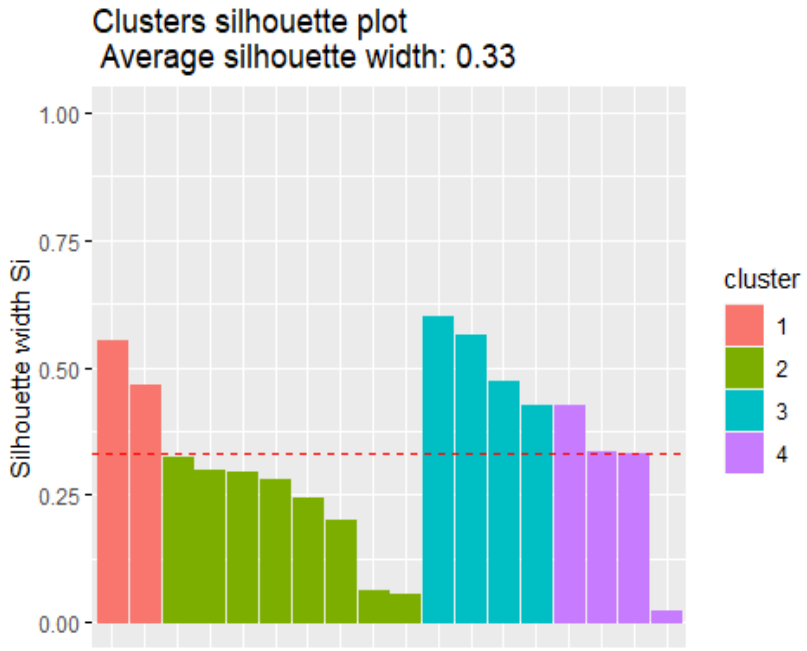



```r
#Evaluación de la calidad de los clusters
sil <- silhouette(km.res$cluster, dist(datos_ST))
rownames(sil) <- rownames(datos)
head(sil[, 1:3])
```

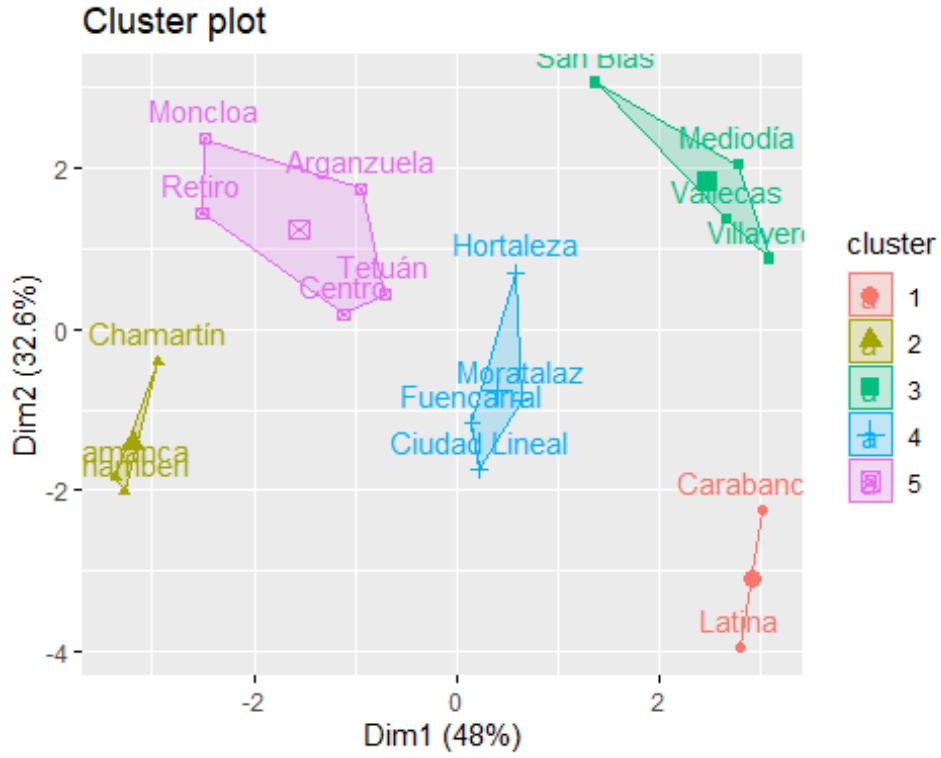
```
##            cluster neighbor  sil_width
## Centro           2        4 0.20080571
## Arganzuela       2        4 0.05674150
## Retiro           2        4 0.32326445
## Salamanca        2        4 0.29819371
## Chamartín        2        4 0.24459304
## Tetuán           2        4 0.06325847
```

```r
fviz_silhouette(sil)
```

```
##   cluster size ave.sil.width
## 1       1    2          0.51
## 2       2    8          0.22
## 3       3    4          0.52
## 4       4    4          0.28
```




```
# Probamos con 5 Clusters que es lo que nos recomienda el criterio Elbow
RNGkind(sample.kind = "Rejection")
set.seed(1234)
km.res5 <- kmeans(datos_ST, 5)
fviz_cluster(km.res5, datos_ST)
```

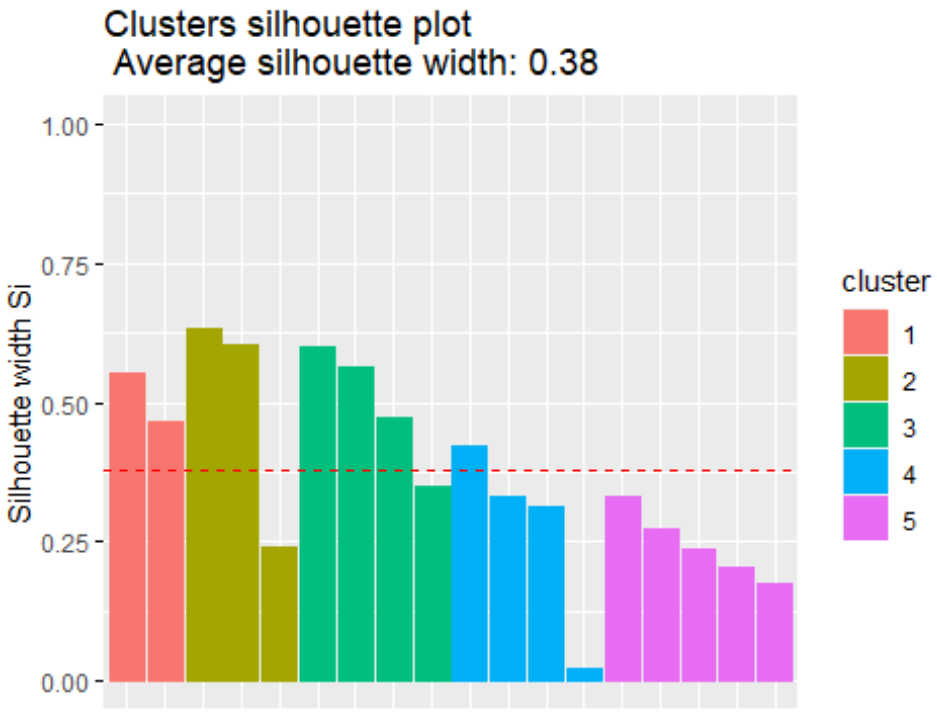


```r
sil <- silhouette(km.res5$cluster, dist(datos_ST))
rownames(sil) <- rownames(datos)
head(sil[, 1:3])
```

```
##            cluster neighbor sil_width
## Centro           5        2 0.1740834
## Arganzuela       5        4 0.2747446
## Retiro           5        2 0.2355824
## Salamanca        2        5 0.6318150
## Chamartín        2        5 0.2419290
## Tetuán           5        4 0.2055077
```

```r
fviz_silhouette(sil)
```

```
##   cluster size ave.sil.width
## 1       1    2          0.51
## 2       2    3          0.49
## 3       3    4          0.50
## 4       4    4          0.27
## 5       5    5          0.24
```




```r
ordenado<-sort(km.res5$cluster)
knitr::kable(ordenado, digits =2, caption = "Barrio y cluster")
```

Barrio y cluster

	x
Latina	1
Carabanchel	1
Salamanca	2
Chamartín	2
Chamberi	2
Villaverde	3
Mediodía	3
Vallecas	3
San Blas	3
Fuencarral	4
Moratalaz	4
Ciudad Lineal	4
Hortaleza	4
Centro	5
Arganzuela	5
Retiro	5
Tetuán	5
Moncloa	5

```
knitr::kable(km.res5$centers, digits =2,caption = "Estadísticos de los
clusters, datos STD")
```

Estadísticos de los clusters, datos STD

P_TOT	P_14	P_65	ANAL	NES	OCU	OCUIN	OCUSER	TEC	PD	TM
2.26	1.91	0.43	0.74	-0.19	1.35	1.48	1.68	-0.05	-0.25	1.46
0.21	-0.68	0.97	-1.08	1.70	-0.28	-0.90	0.43	1.57	1.51	-1.17
0.02	0.16	-0.64	1.40	-1.17	-0.58	0.63	-0.94	-1.41	-1.25	0.92
-0.34	0.59	-0.62	-0.18	-0.25	0.85	0.48	0.44	0.32	0.54	0.27
-0.77	-0.96	0.25	-0.63	0.19	-0.58	-0.94	-0.53	-0.05	-0.24	-0.83

```
#Se puede calcular las medias de las variables originales
EsT_Clus<-aggregate(dat_B, by=list(km.res5$cluster),mean)
knitr::kable(EsT_Clus, digits =2,caption = "Estadísticos de los
clusters")
```

Estadísticos de los clusters

Group.1	P_TOT	P_14	P_65	ANAL	NES	OCU	OCUIN	OCUSER	TEC	PD	TM
1	272.70	70.00	23.60	6.65	19.65	82.00	18.55	54.95	8.95	1.15	27.30
2	180.90	30.13	28.17	1.30	43.93	55.00	7.43	42.43	15.90	2.60	5.43
3	172.65	43.02	14.45	8.60	7.00	50.05	14.60	28.85	3.15	0.32	22.80
4	156.28	49.70	14.62	3.95	18.85	73.65	13.90	42.60	10.55	1.80	17.38
5	137.24	25.74	22.04	2.62	24.54	50.06	7.28	32.86	8.94	1.16	8.22