Punto 4

a) Genere un conjunto de datos simulados con 20 observaciones en cada una de tres clases (es decir, 60 observaciones en total) y 50 variables. Sugerencia: hay una serie de funciones en R que puede utilizar para generar datos. Un ejemplo es la función rnorm(); runif() es otra opción. Asegúrese de agregar un cambio en la media en las observaciones de cada clase a fin de obtener tres clases distintas.

b) Realice PCA en las 60 observaciones y grafique las observaciones en términos de las 2 primeras variables principales Z1 y Z2. Use un color diferente para indicar las observaciones en cada una de las tres clases. Si las tres clases aparecen separados en esta gráfica, solo entonces continúe con la parte (c). Si no, vuelva al inciso a) y modifique la simulación para que haya una mayor separación entre las tres clases. No continúe con la parte (c) hasta que las tres clases muestren al menos algún grado de separación en los dos primeros vectores de scores de componentes principales.

c) Desarrolle agrupación de K-medias de las observaciones con K = 3. ¿Qué tan bien funcionan los clústeres que obtuvo con el algoritmo de K-medias comparado con las verdaderas etiquetas de clase? Sugerencia: puede usar la función table() en R para comparar las verdaderas etiquetas de clase con las etiquetas de clase obtenidas por agrupamiento. Tener cuidado cómo se interpretan los resultados: el agrupamiento de K-medias numera los grupos arbitrariamente, por lo que no puede simplemente comprobar si las verdaderas etiquetas de clase y las etiquetas de agrupación son las mismas.

d) Realice agrupamiento de K-medias con K = 2. Describa sus resultados.

e) Ahora realice agrupamiento de K-medias con K = 4 describa su resultados.

f)Ahora realice agrupamiento de K-medias con K = 3 en los dos primeros vectores de scores de componentes principales, en lugar de los datos en las variables originales. Es decir, realice la agrupación de K-medias en la matriz de 60 O 2, cuya primera columna es la coordenada zi1 en la primera componente principal Z1 y la segunda columna es la coordenada zi2 en la segunda componente principal Z2. Comente los resultados.

g) Con la función scale(), realice agrupamiento de K-medias con K = 3 en los datos después de escalar cada variable para tener una desviación estándar de uno. ¿Cómo se comparan estos resultados con los obtenidos? en (b)? Explique.

Solución

```
library(stats)#prcomp
library(factoextra)
```

```
## Loading required package: ggplot2
```

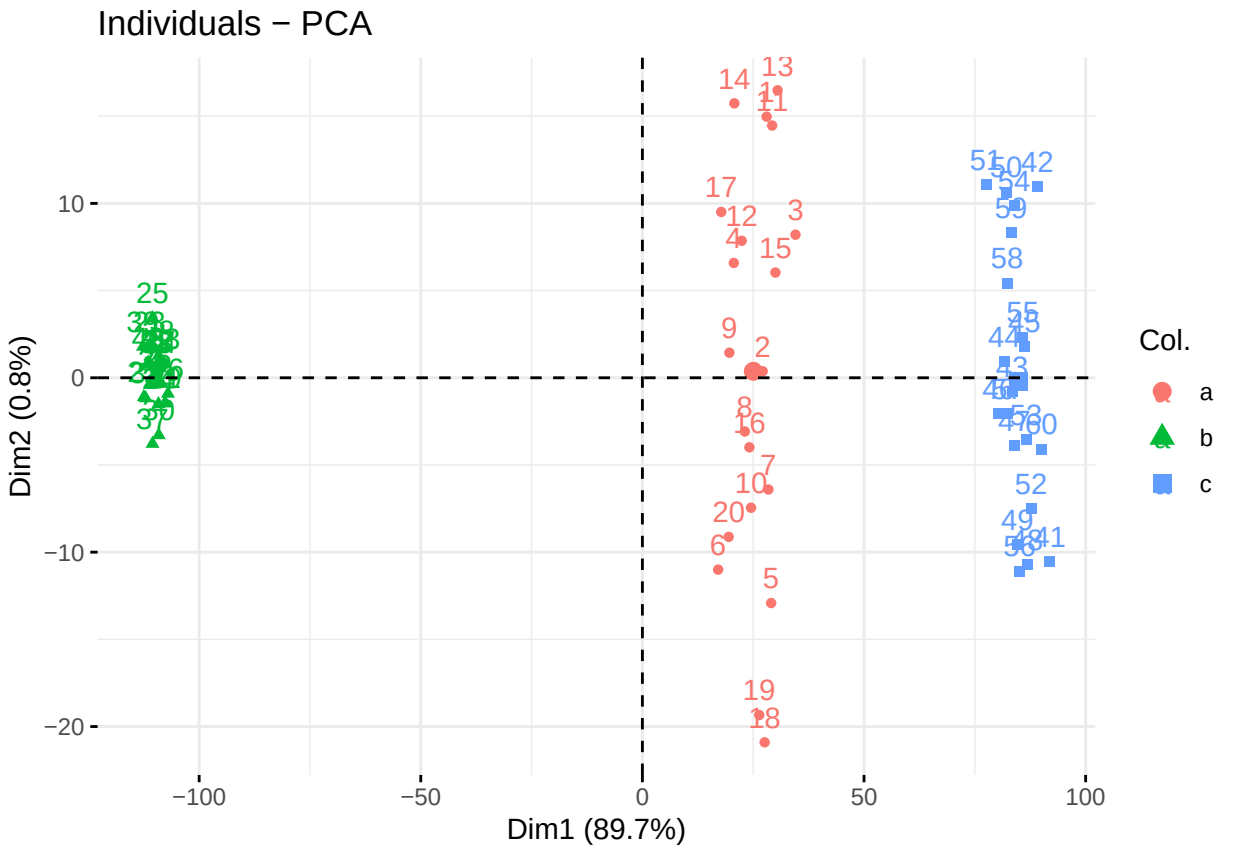
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

a) Haciendo uso de algunas funciones en R para generar datos aleatorios de una distribución como rpois(), rnorm() y runif() se generará una base de datos con 60 observaciones, 50 variables numéricas y una categórica con los tres difrentes grupos.

```
set.seed(1016)
datos <- data.frame("grupo" = c(rep("a",20),rep("b",20),rep("c",20)))
for(i in 2:51){
  datos[,i] <- c(rpois(n = 20,lambda = 24),
                 rnorm(n = 20, mean = 5 ,sd = 2),
                 runif(n = 20, min = 25, max = 40))
}
```

b) Haciendo un ajuste de PCA con ayuda de la función princomp() se procede a gráficar los datos en términos de las 2 primeras variables principales Z1 y Z2, en las cuales se resaltan los grupos con los colores rosado, verde y azul. Además cabe mencionar que la primera componente es la que mayor variabilidad logra explicar del conjunto de datos:

```
res.pca <- princomp(datos[,-1])
fviz_pca_ind(res.pca,col.ind = datos$grupo)
```



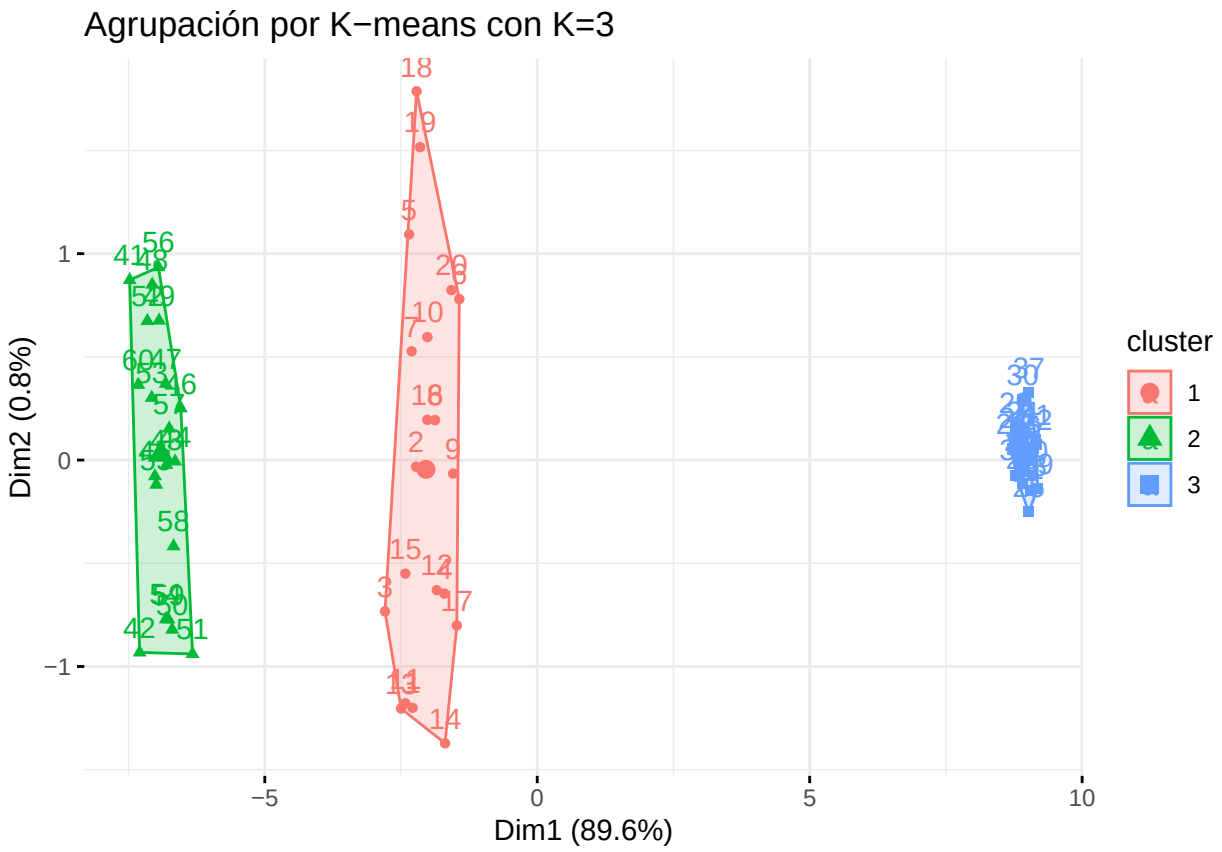
c) Se realiza a continuación agrupación por k medias con K = 3, y con 20 asignaciones iniciales de clúster. En comparación con el gráfico de literal b, se puede concluir que, a pesar que la agrupación por k-means aleatoriza las etiquetas, la clústerización distingue correctamente las 3 diferentes poblaciones. Además, la escala de medida tuvo una reducción.

```
set.seed(1016)
km.out =kmeans (datos[,-1],3, nstart =20)
km.out$tot.withinss
```

```
## [1] 45635.42
```

```
fviz_cluster(km.out, data = datos[,-1],ggtheme = theme_minimal(base_size = 11),
             title = "Agrupación por K-means con K=3")
```

```
## Warning: argument title is deprecated; please use main instead.
```



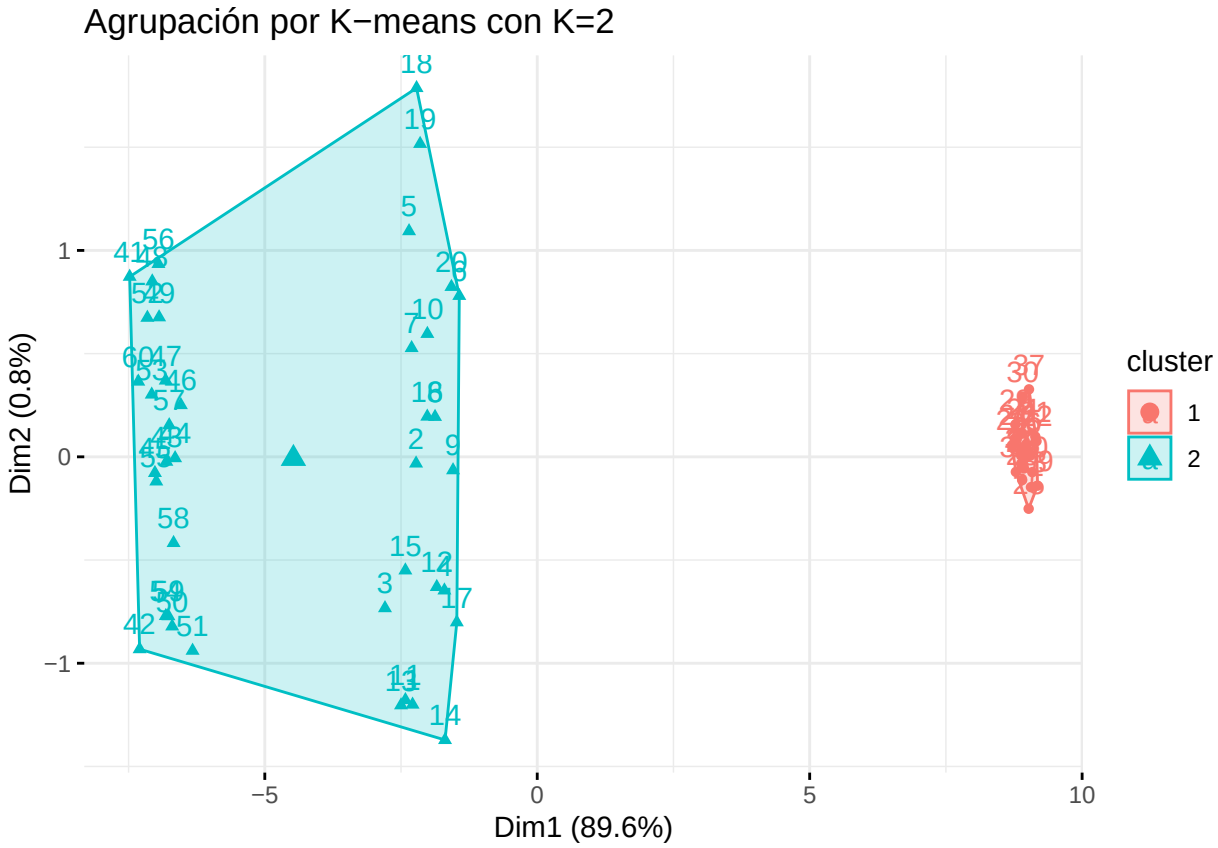
d) Realizando agrupamiento de K-medias con K = 2 y con 20 asignaciones iniciales de clúster, se puede observar que el algoritmo clasificó 2 poblaciones. Dos de las poblaciones originales, al parecer las más cercanas, las convirtió en una sola, la cual corresponde al cluster número 2 en el siguiente gráfico:

```
set.seed(1016)
km.out =kmeans (datos[,-1],2, nstart =20)
km.out$tot.withinss
```

```
## [1] 82198.7
```

```
fviz_cluster(km.out, data = datos[,-1],ggtheme = theme_minimal(base_size = 11),
             title = "Agrupación por K-means con K=2")
```

```
## Warning: argument title is deprecated; please use main instead.
```



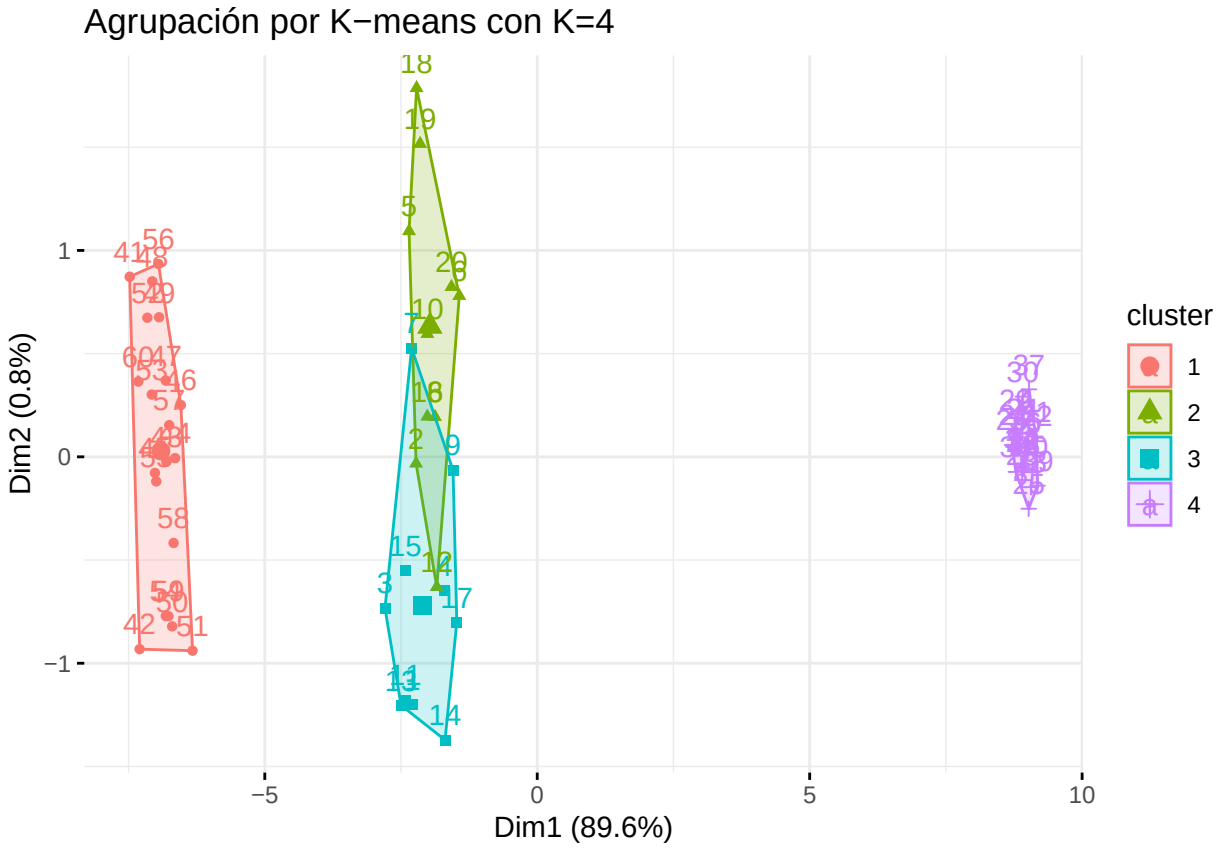
e) Realizando agrupamiento de K-medias con K = 4 y con 20 asignaciones iniciales de clúster, se puede observar que el algoritmo clasificó 4 poblaciones, dos de las poblaciones originales se matuvieron igual (cluster 1 y 4) y una de ellas la dividió en dos, las cuales corresponden en el siguiente gráfico a los clusters 2 y 3.

```
set.seed(1016)
km.out =kmeans (datos[,-1],4, nstart =20)
km.out$tot.withinss
```

```
## [1] 43144.57
```

```
fviz_cluster(km.out, data = datos[,-1],ggtheme = theme_minimal(base_size = 11),
             title = "Agrupación por K-means con K=4")
```

```
## Warning: argument title is deprecated; please use main instead.
```

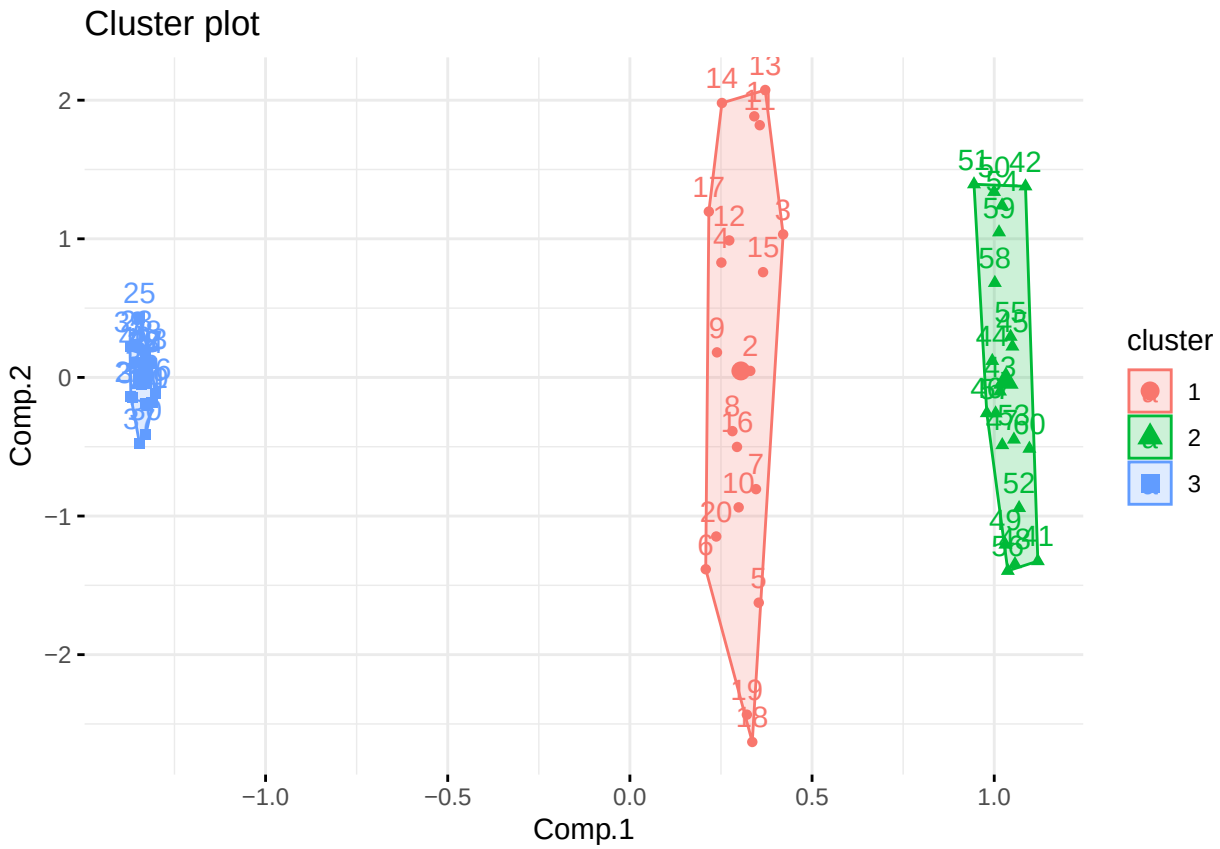


f) Por medio del siguiente gráfico se análizan los scores obtenidos del ajuste de componentes principales con la técnica de agrupamiento por K-means, con K=3 y con 20 asignaciones iniciales de clúster. En primer lugar, es notorio el cambio en la escala de medida, que pasó al intervalo (-2 , -1.5) para la primera componente y de (-3, 2) para la segunda. En cuanto a la clasificación, el algoritmo obtiene una correcta clasificación para los individuos de cada población. Por ejemplo, los individuos de uno al veinte, que corresponden a los generados por la distribución poison, quedaron en el cluster número 1.

```
set.seed(1016)
scores12 <- data.frame(res.pca$scores[,c(1,2)])
km.out <- kmeans(scores12, 3, nstart =20)
km.out$tot.withinss
```

```
## [1] 4434.877
```

```
fviz_cluster(km.out, data = scores12, ggtheme = theme_minimal(base_size = 11))
```

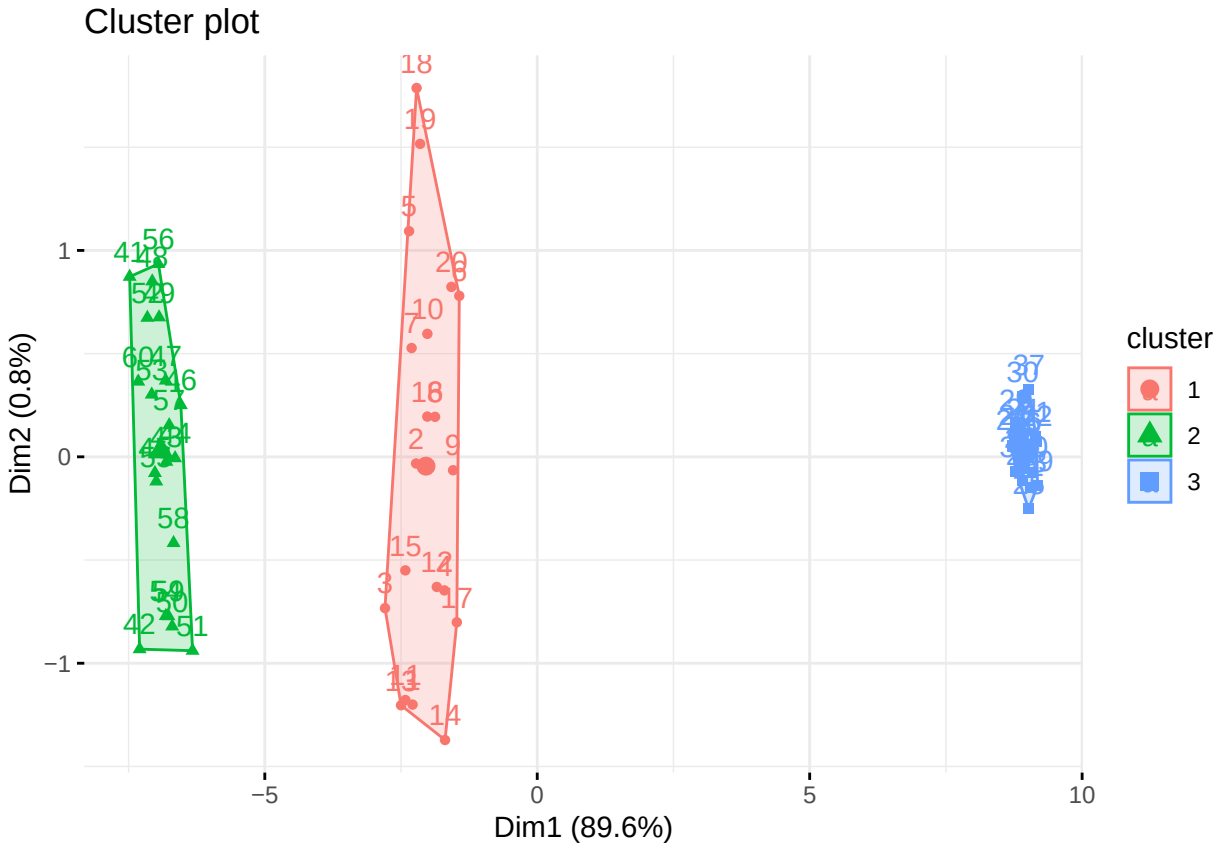


g) A contunuación se ajusta el agrupamiento por K-means con los datos escalados y con 20 asignaciones iniciales de clúster. En el gráfico a continuación, es notorio el cambio en la escala de medida, que pasó al intervalo aproximadamente (-7, 10) para la primera componente y de aproximadamente (-2, 3) para la segunda. Por otro lado, con los datos escalados se obtienen buenas clasificaciones para los diferentes grupos de población de la base de datos.

```
set.seed(1016)
data_scal <- scale(datos[,-1])
km.out <- kmeans (data_scal,3, nstart =20)
km.out$tot.withinss
```

```
## [1] 303.9718
```

```
fviz_cluster(km.out, data = data_scal,  ggtheme = theme_minimal(base_size = 11))
```


Cluster plot
Dim2 (0.8%)
Dim1 (89.6%)
cluster
1
2
3