Segmentación de clientes de un banco

Ruben Serrano Hernández

2023-04-25

```
datos <- read.csv("C:\\Users\\Cosas\\OneDrive\\Escritorio\\ESTADISTICA\\MINERIA DE DATOS\\Credit_card.cs
head(datos,10)
```

```
##    Sl_No Customer.Key Avg_Credit_Limit Total_Credit_Cards Total_visits_bank
## 1      1        87073           100000                  2                 1
## 2      2        38414            50000                  3                 0
## 3      3        17341            50000                  7                 1
## 4      4        40496            30000                  5                 1
## 5      5        47437           100000                  6                 0
## 6      6        58634            20000                  3                 0
## 7      7        48370           100000                  5                 0
## 8      8        37376            15000                  3                 0
## 9      9        82490             5000                  2                 0
## 10    10        44770             3000                  4                 0
##    Total_visits_online Total_calls_made
## 1                    1                0
## 2                   10                9
## 3                    3                4
## 4                    1                4
## 5                   12                3
## 6                    1                8
## 7                   11                2
## 8                    1                1
## 9                    2                2
## 10                   1                7
```

```
colSums(is.na(datos))
```

```
##               Sl_No        Customer.Key    Avg_Credit_Limit  Total_Credit_Cards
##                  0                   0                   0                   0
##   Total_visits_bank Total_visits_online    Total_calls_made
##                   0                   0                   0
```

```
apply(datos,2,function(x) any(x==-99))
```

```
##               Sl_No        Customer.Key    Avg_Credit_Limit  Total_Credit_Cards
##               FALSE               FALSE                TRUE                TRUE
##   Total_visits_bank Total_visits_online    Total_calls_made
##                TRUE                TRUE                TRUE
```

```
datos[,c(3:7)] <- replace(datos[,c(3:7)], datos[,c(3:7)] == -99, NA)
```

```
colSums(is.na(datos))
```

```
##               Sl_No        Customer.Key    Avg_Credit_Limit  Total_Credit_Cards
```

```
##                    0                    0                    1                    1
##   Total_visits_bank Total_visits_online    Total_calls_made
##                    1                    1                    1
```

```
apply(is.na(datos),2,which)
```

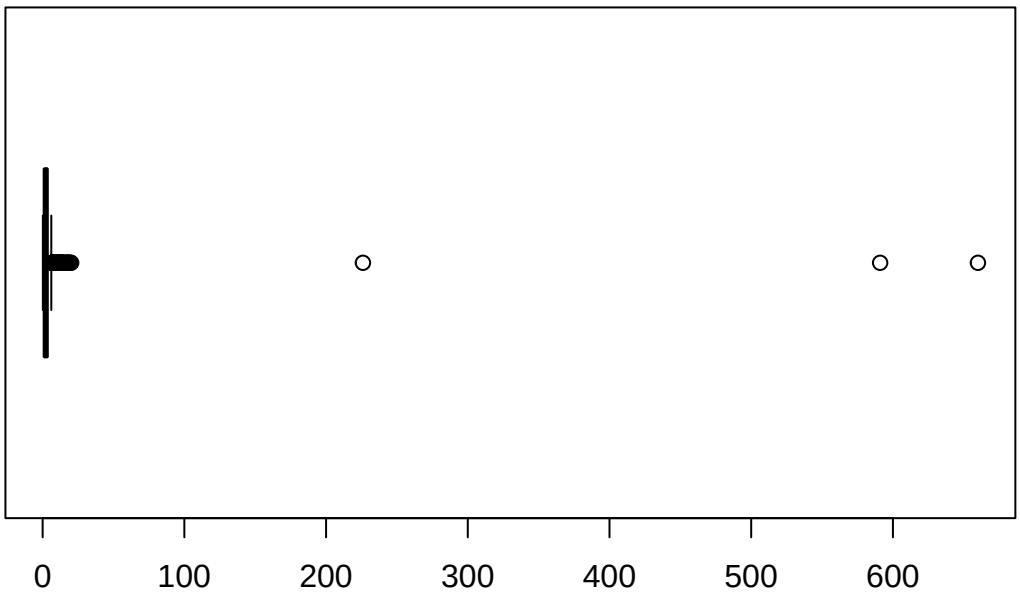
```
## $Sl_No
## integer(0)
##
## $Customer.Key
## integer(0)
##
## $Avg_Credit_Limit
## [1] 661
##
## $Total_Credit_Cards
## [1] 661
##
## $Total_visits_bank
## [1] 661
##
## $Total_visits_online
## [1] 661
##
## $Total_calls_made
## [1] 661
```

Hemos investigado que datos faltantes hay y llegado a la conclusión de que sólo el individuo 661 tiene datos faltantes, no tiene sentido hacer ningún tipo de imputación ya que hay sólo un individuo con 'missing values' y están en la mayoría de columnas. Por lo tanto lo más óptimo sería eliminarlo.

```
datos<-datos[-661,]
rownames(datos) <- c(1:(nrow(datos)))
```

Una vez hecha la "limpieza" de los datos, procedo a realizar un análisis de detección de valores anómalos (outliers) para identificar observaciones que se desvían significativamente de la distribución típica de la variable en cuestión. Como se trata de un caso multidimensional, calculo la distancia de mahalanobis y confirmamos con el análisis gráfico.

```
library("outliers")
mu <- colMeans(datos[,3:7])
sigma <- cov(datos[,3:7], use='p')
distancias <- mahalanobis(datos[,3:7], center = mu, cov = sigma)
boxplot(distancias, horizontal =T)
```




```
dim(datos)
```

```
## [1] 662   7
```

```
#Correción por Bonferroni. alpha_n = alpha / 662. Hemos calculado la distancia con 5 variables así que
qchisq(1 - 0.05/662, 5)
```

```
## [1] 26.37316
```

Por lo tanto, todos los individuos que tengan una distancia de Mahalanobis superior a 26.37 son considerados anómalos según el test.

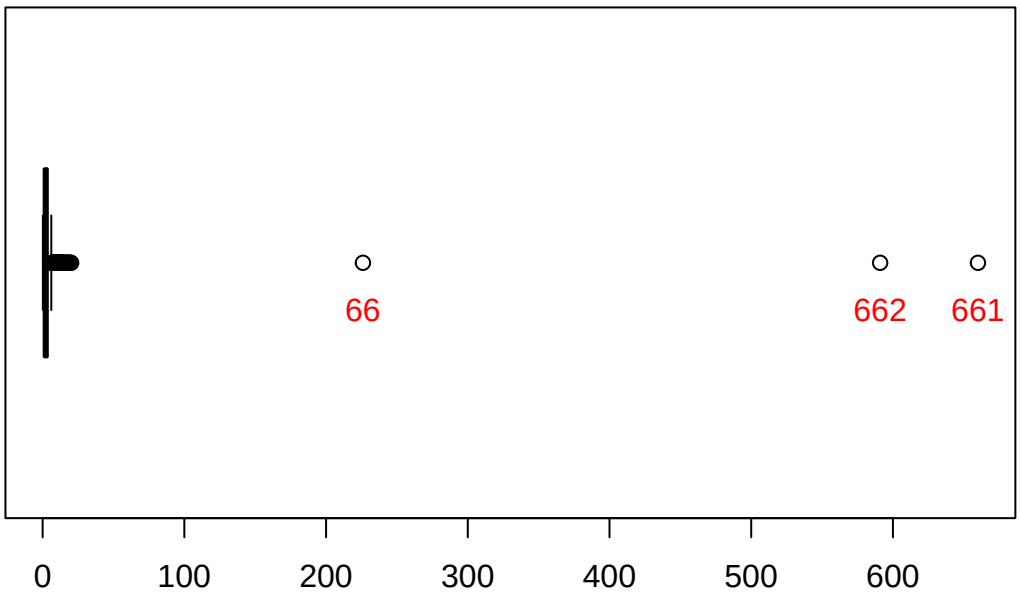
```
out<-which(distancias>26.37)
print("Los individuos cuya distancia de Mahalanobis es superior a 26.37 son:")
```

```
## [1] "Los individuos cuya distancia de Mahalanobis es superior a 26.37 son:"
```

```
print(out)
```

```
##  66 661 662
##  66 661 662
```

```
distancias <- as.vector(distancias)
boxplot(distancias, horizontal =T)
text(x = distancias[out], y = (rep(0.9, length(out))), labels = out, col = "red")
```




```
colnames(datos)
```

```
## [1] "Sl_No"                "Customer.Key"         "Avg_Credit_Limit"
## [4] "Total_Credit_Cards"   "Total_visits_bank"    "Total_visits_online"
## [7] "Total_calls_made"
```

```
library(ggplot2)
library(dplyr)
```

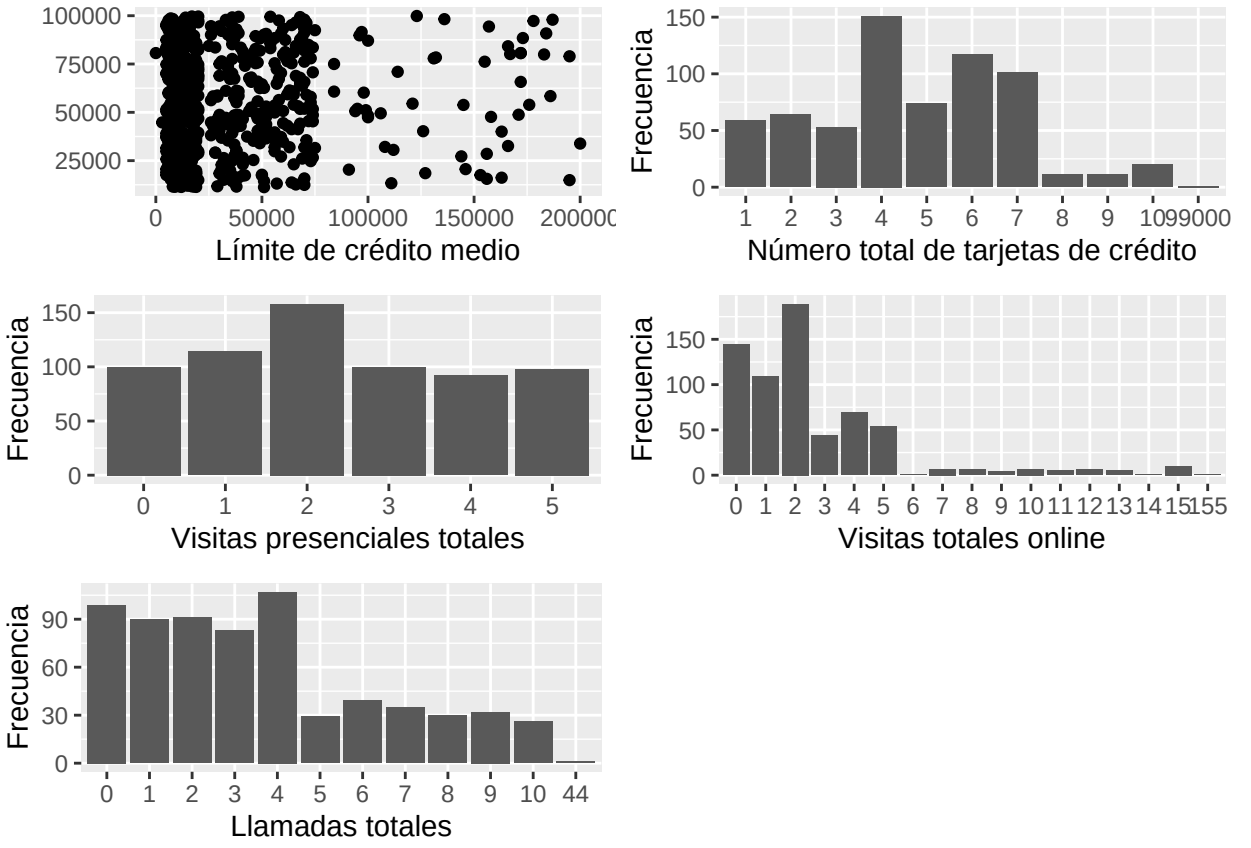
```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggpubr)

g1 <- ggplot(datos, aes(x = Avg_Credit_Limit, y = Customer.Key)) +
  geom_point() +
  xlab("Límite de crédito medio") +
  ylab("")
g2 <- ggplot(datos, aes(x = as.factor(Total_Credit_Cards))) +
  geom_bar(stat = "count") +
  xlab("Número total de tarjetas de crédito") +
```

```
  ylab("Frecuencia")
g3 <- ggplot(datos, aes(x = as.factor(Total_visits_bank))) +
  geom_bar(stat="count") +
  xlab("Visitas presenciales totales") +
  ylab("Frecuencia")
g4 <- ggplot(datos, aes(x = as.factor(Total_visits_online))) +
  geom_bar(stat="count") +
  xlab("Visitas totales online") +
  ylab("Frecuencia")
g5 <- ggplot(datos, aes(x = as.factor(Total_calls_made))) +
  geom_bar(stat="count") +
  xlab("Llamadas totales") +
  ylab("Frecuencia")

ggarrange(g1,g2,g3,g4,g5, ncol=2, nrow=3, size = c(5,5,5,5,5))
```

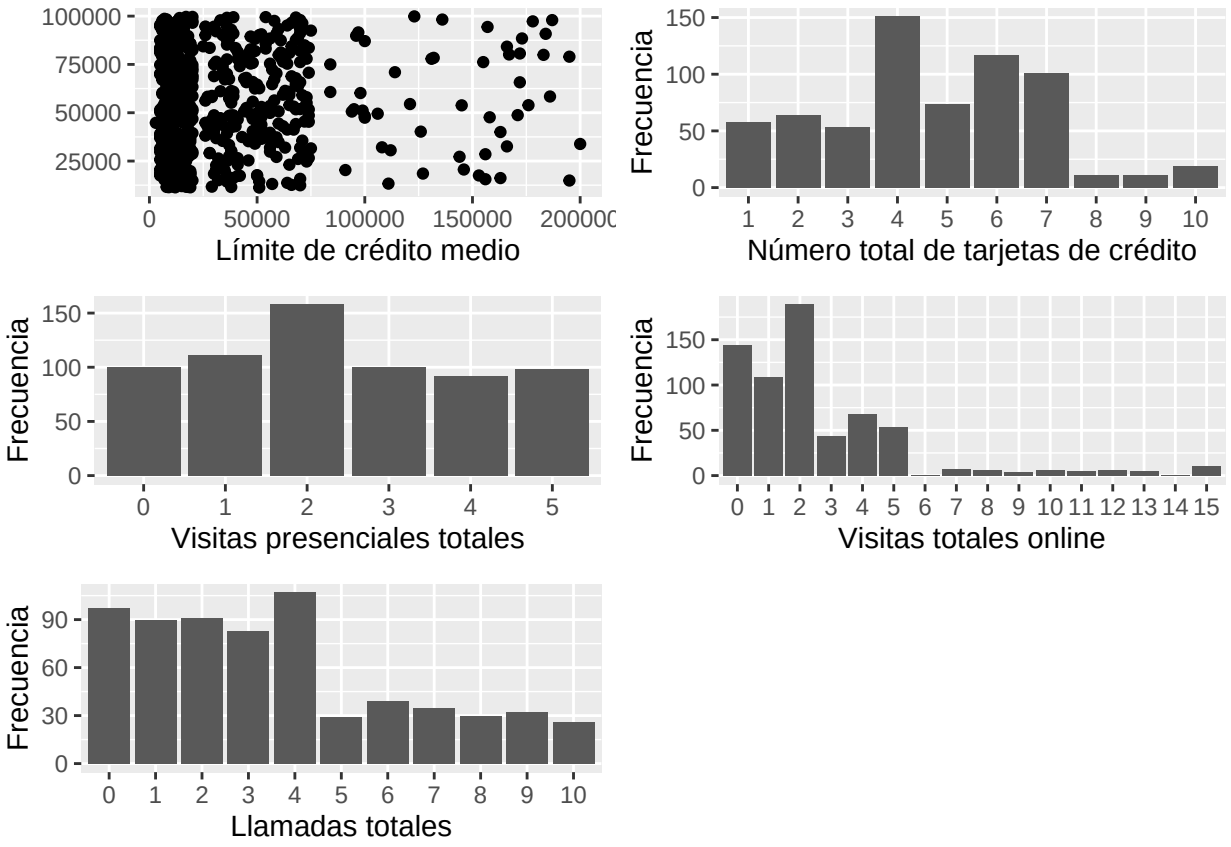



```
outliers = c(which(datos$Total_Credit_Cards==99000), which(datos$Total_visits_online==155), which(datos$
print(outliers)
```

```
## [1] 661 662  66
```

```
datos <- datos[-c(66,661,662),]

write.csv(datos, 'datos_banco_final.csv')

g1 <- ggplot(datos, aes(x = Avg_Credit_Limit, y = Customer.Key)) +
  geom_point() +
```

```
  xlab("Límite de crédito medio") +
  ylab("")
g2 <- ggplot(datos, aes(x = as.factor(Total_Credit_Cards))) +
  geom_bar(stat = "count") +
  xlab("Número total de tarjetas de crédito") +
  ylab("Frecuencia")
g3 <- ggplot(datos, aes(x = as.factor(Total_visits_bank))) +
  geom_bar(stat="count") +
  xlab("Visitas presenciales totales") +
  ylab("Frecuencia")
g4 <- ggplot(datos, aes(x = as.factor(Total_visits_online))) +
  geom_bar(stat="count") +
  xlab("Visitas totales online") +
  ylab("Frecuencia")
g5 <- ggplot(datos, aes(x = as.factor(Total_calls_made))) +
  geom_bar(stat="count") +
  xlab("Llamadas totales") +
  ylab("Frecuencia")

ggarrange(g1,g2,g3,g4,g5, ncol=2, nrow=3, size = c(5,5,5,5,5))
```




```
library(cluster)
dist_euclidea <- dist(datos)

dist_gower <- daisy(datos, metric="gower")
```

```r
fit <- cmdscale(dist_euclidea, k=2, eig=TRUE)
nuevo_dataset <-fit$points

head(nuevo_dataset, 4)
```

```
##         [,1]      [,2]
## 1 -67912.043  26167.37
## 2 -13893.557 -17987.71
## 3 -12072.083 -38981.84
## 4   5851.505 -14184.78
```

```r
fit$GOF
```

```
## [1] 0.9999905 0.9999905
```

```r
library('MASS')
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

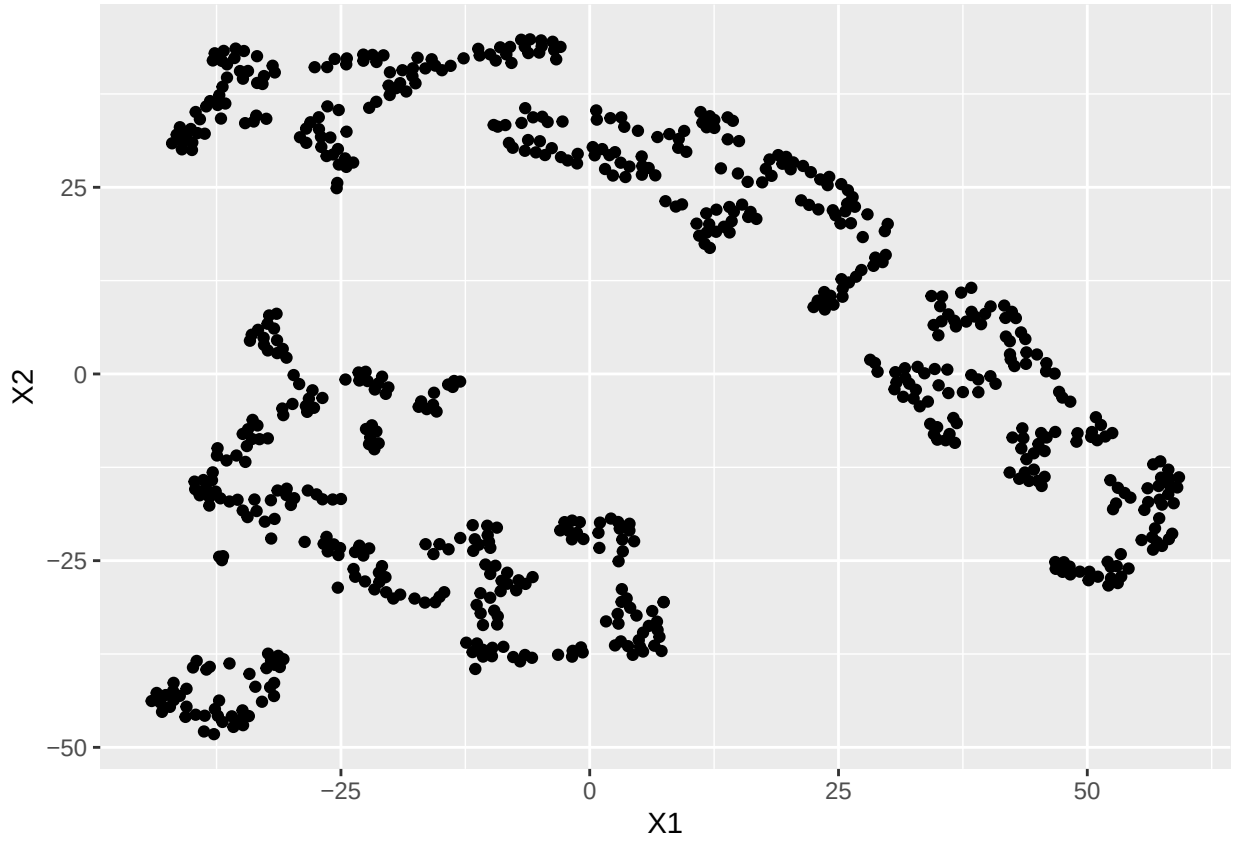
```r
fit2 <- sammon(dist_euclidea, k=2, trace=F)
nuevo_dataset2<- fit2$points

fit2$stress
```

```
## [1] 8.355459e-08
```

```r
library("Rtsne")
set.seed(9202)
tsne <- Rtsne(datos, dims=2, perplexity=16, theta = 0, pca = FALSE)
nuevos_datos <- tsne$Y
head(nuevos_datos)
```

```
##            [,1]      [,2]
## [1,] -36.962637 -24.93320
## [2,]  -8.852506 -27.69262
## [3,]  -0.706113 -37.28392
## [4,]   3.273944 -22.17757
## [5,] -31.140004 -39.23805
## [6,]   7.612065  23.13635
```

```r
ggplot(data.frame(nuevos_datos),
       aes(x = X1, y = X2)) +
  geom_point()
```




```
t <- ggplot(data.frame(nuevos_datos),
            aes(x = X1, y = X2)) +
            geom_point() +
            ggtitle("tsne")
```

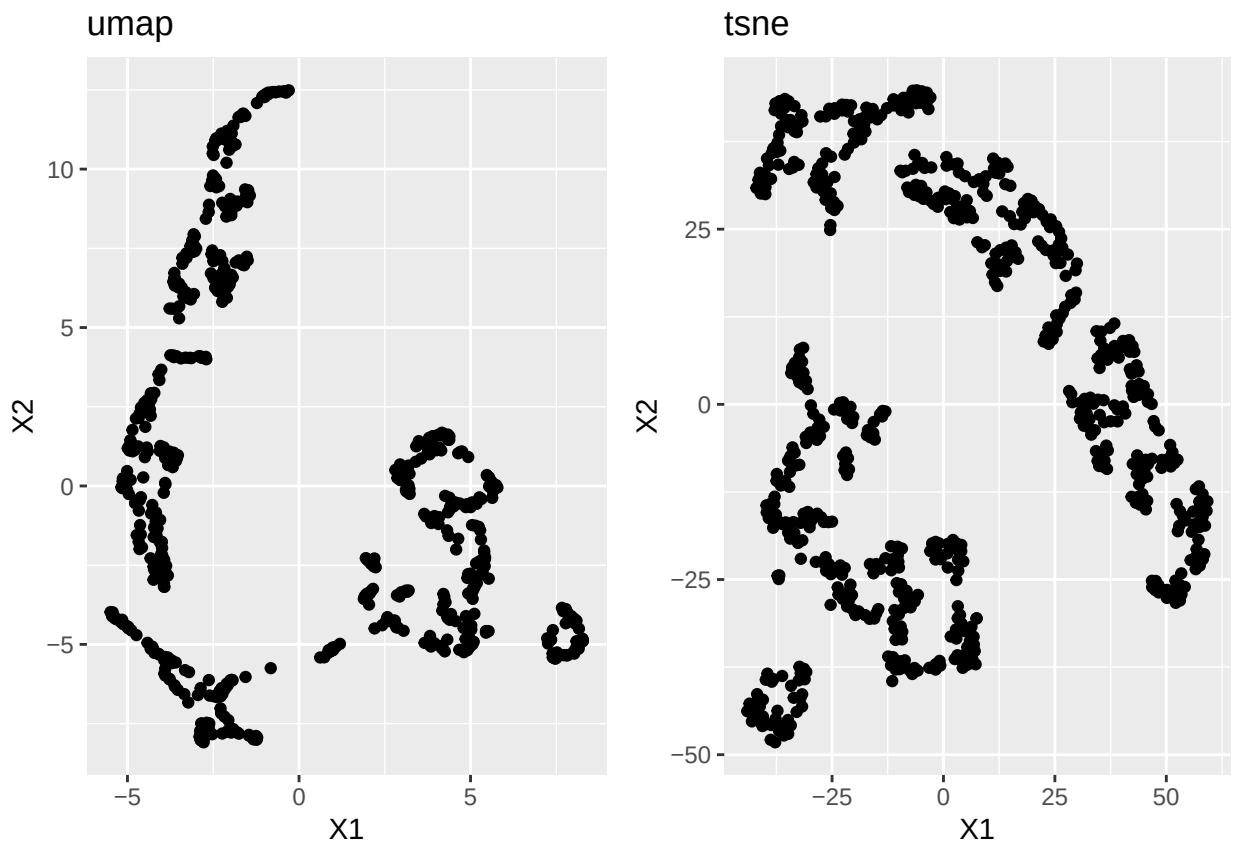
```
library("umap")
umap <- umap(datos)
```

```
set.seed(9202)
umap_dataset <- umap$layout
u <- ggplot(data.frame(umap$layout), aes(x = X1, y = X2)) +
          geom_point() +
          ggtitle("umap")
```

```
ggarrange(u, t, ncol=2, nrow=1, size = c(5,5))[1]
```

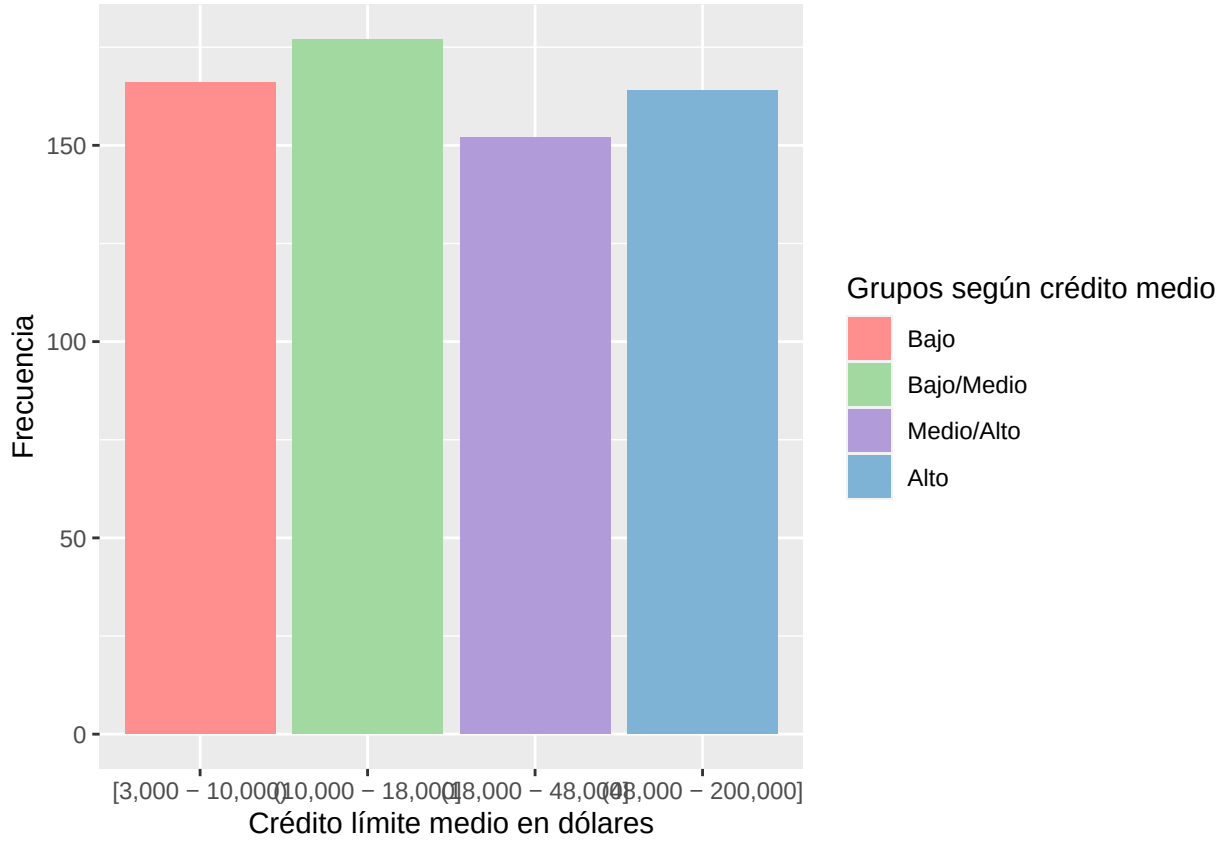
```
## Warning in as_grob.default(plot): Cannot convert object of class numeric into a
## grob.
```

```
## $`1`
```



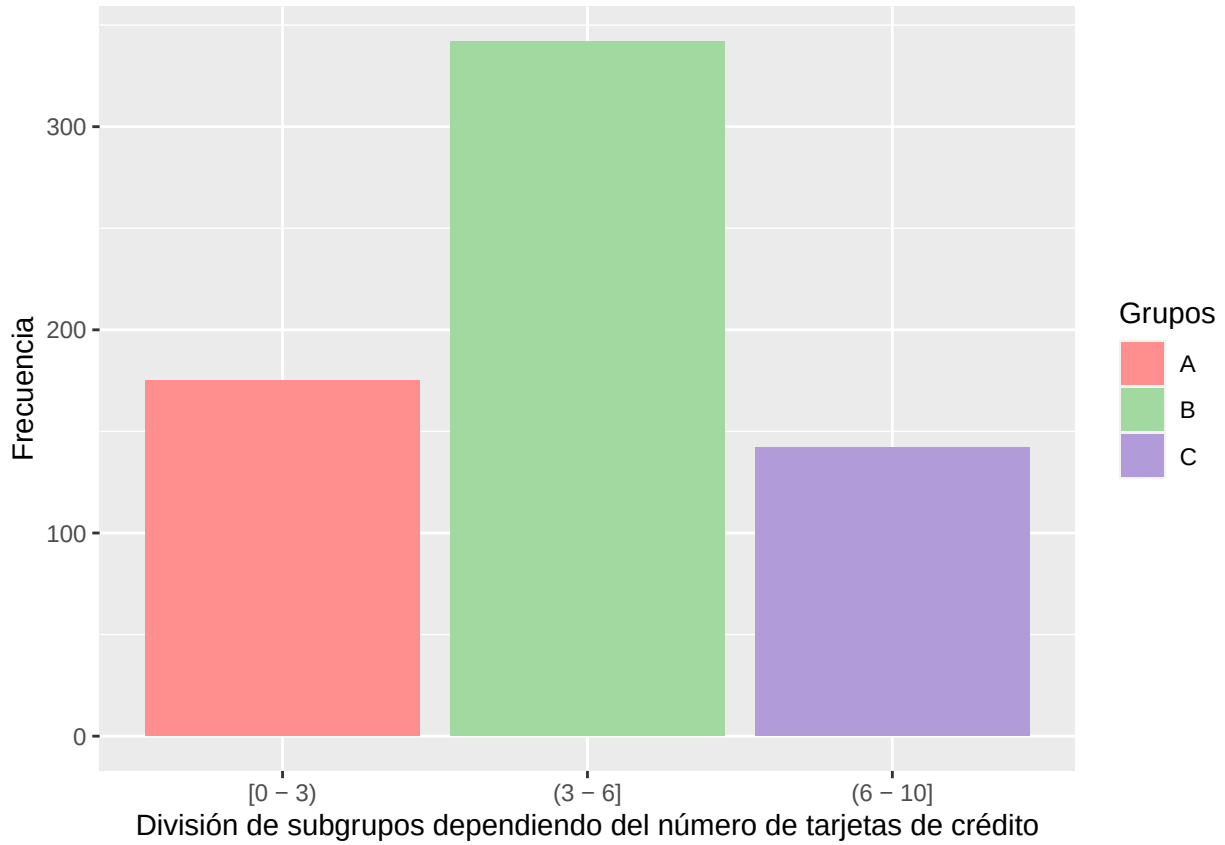

```
datos_disc <- datos[,-c(3,4,6,7)]
```

```
datos_disc$disc_avg_credit <- cut(datos$Avg_Credit_Limit, quantile(datos$Avg_Credit_Limit,probs = seq(0

levels(datos_disc$disc_avg_credit) <- c("[3,000 - 10,000)", "(10,000 - 18,000]", "(18,000 - 48,000]", "

ggplot(datos_disc, aes(x = as.factor(disc_avg_credit))) +
  geom_bar(aes(fill = as.factor(disc_avg_credit)), stat = "count") +
  xlab("Crédito límite medio en dólares") +
  ylab("Frecuencia") +
  scale_fill_manual(name = "Grupos según crédito medio", values = c("#FF8F8F", "#A2D9A1", "#B19CD9", "#
```



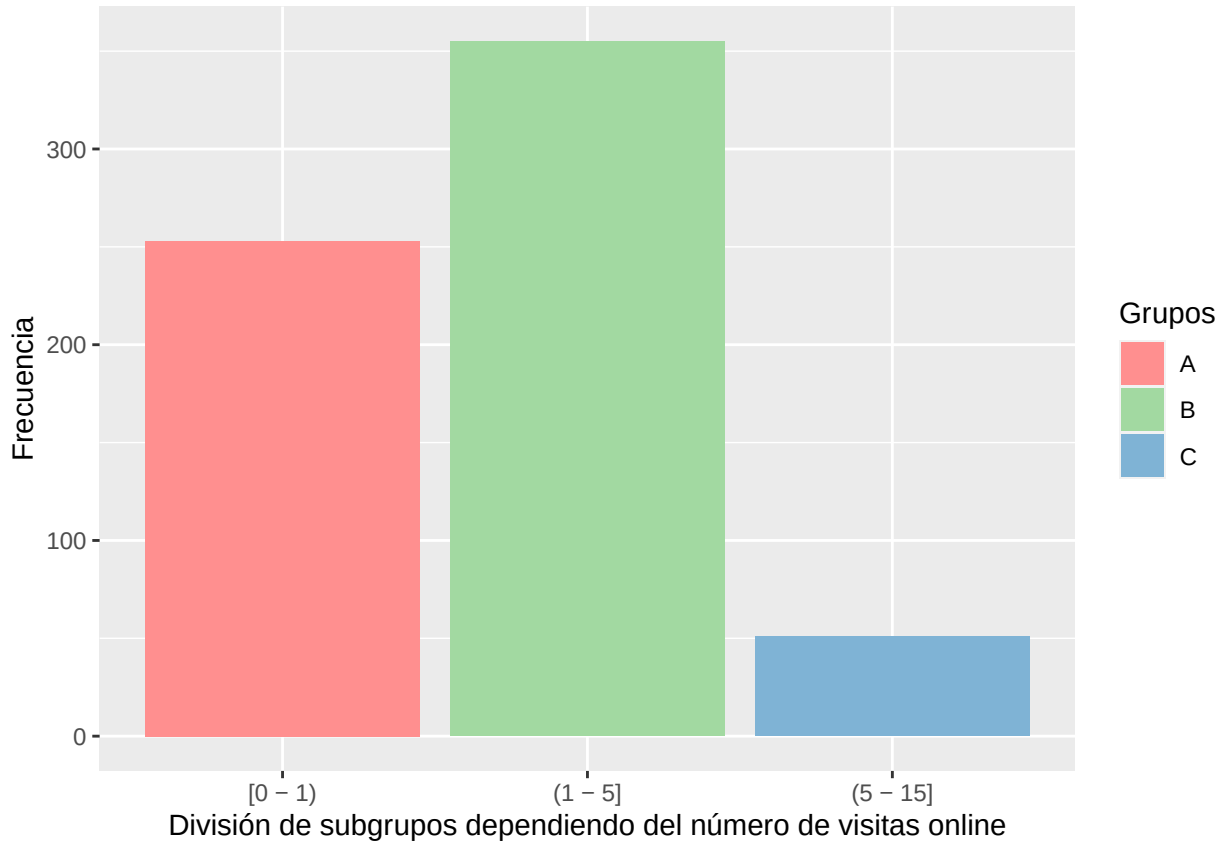

```
levels(datos_disc$disc_avg_credit) <- c('Bajo','Bajo/Medio','Medio/Alto','Alto')
```

```
datos_disc$disc_total_credit <- cut(datos$Total_Credit_Cards, c(0,3,6,10), include.lowest = T)

levels(datos_disc$disc_total_credit)  <- c("[0 - 3)", "(3 - 6]", "(6 - 10]")

ggplot(datos_disc, aes(x = as.factor(disc_total_credit))) +
  geom_bar(aes(fill = as.factor(disc_total_credit)), stat = "count") +
  xlab("División de subgrupos dependiendo del número de tarjetas de crédito") +
  ylab("Frecuencia") +
  scale_fill_manual(name = "Grupos", values = c("#FF8F8F", "#A2D9A1", "#B19CD9"), labels = c("A", "B",
```



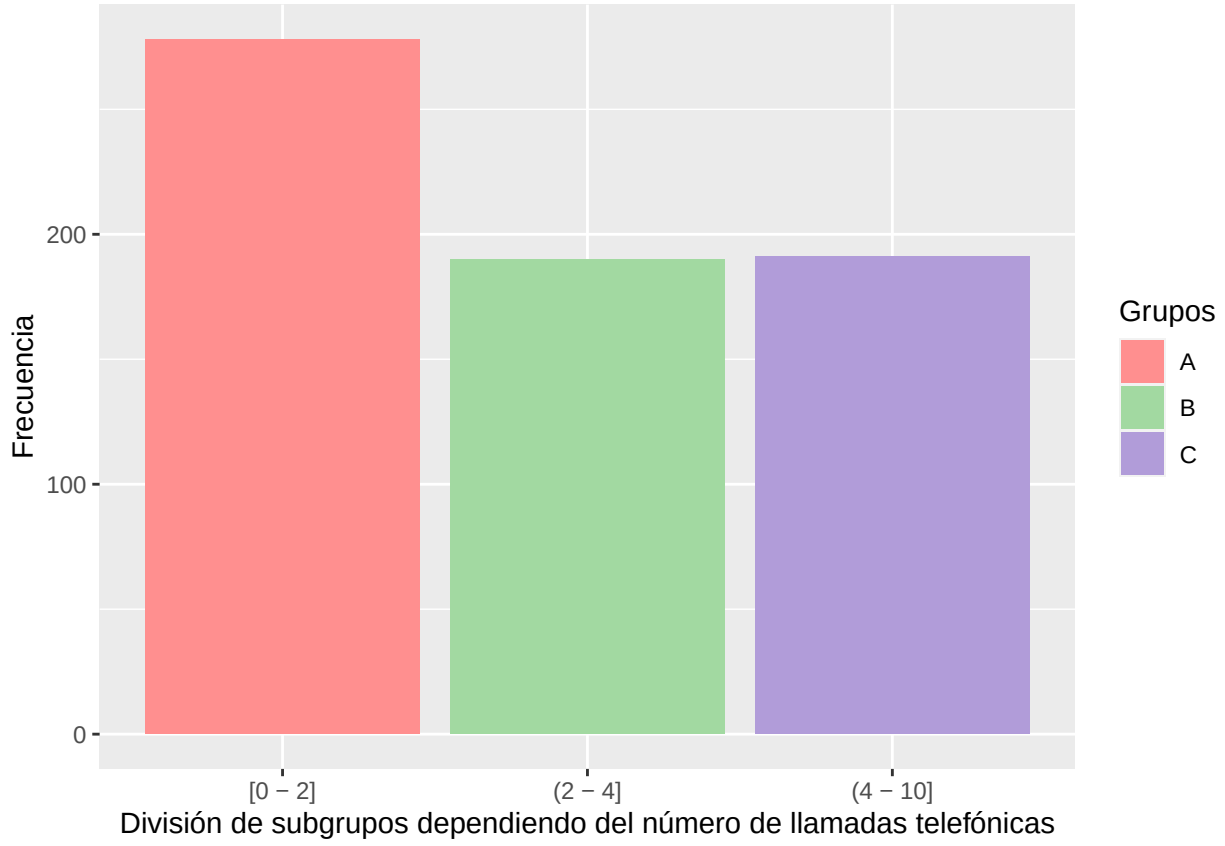

```
levels(datos_disc$disc_total_credit) <-  c("A","B","C")
```

```
datos_disc$disc_online_visits <- cut(datos$Total_visits_online, c(0,1,5,15), include.lowest = T)

levels(datos_disc$disc_online_visits) <- c("[0 - 1)", "(1 - 5]", "(5 - 15]")

ggplot(datos_disc, aes(x = as.factor(disc_online_visits))) +
  geom_bar(aes(fill = as.factor(disc_online_visits)), stat = "count") +
  xlab("División de subgrupos dependiendo del número de visitas online") +
  ylab("Frecuencia") +
  scale_fill_manual(name = "Grupos", values = c("#FF8F8F", "#A2D9A1", "#7FB3D5"), labels = c("A", "B",
```




```
levels(datos_disc$disc_online_visits) <- c("A", "B", "C")
```

```
datos_disc$disc_total_calls <- cut(datos$Total_calls_made, quantile(datos$Total_calls_made,probs = seq(0

levels(datos_disc$disc_total_calls) <- c("[0 - 2]", "(2 - 4]", "(4 - 10]")

ggplot(datos_disc, aes(x = as.factor(disc_total_calls))) +
  geom_bar(aes(fill = as.factor(disc_total_calls)), stat = "count") +
  xlab("División de subgrupos dependiendo del número de llamadas telefónicas") +
  ylab("Frecuencia") +
  scale_fill_manual(name = "Grupos", values = c("#FF8F8F", "#A2D9A1", "#B19CD9"), labels = c("A", "B",
```




```
levels(datos_disc$disc_total_calls) <-  c("A","B","C")
```

```
head(datos_disc,10)
```

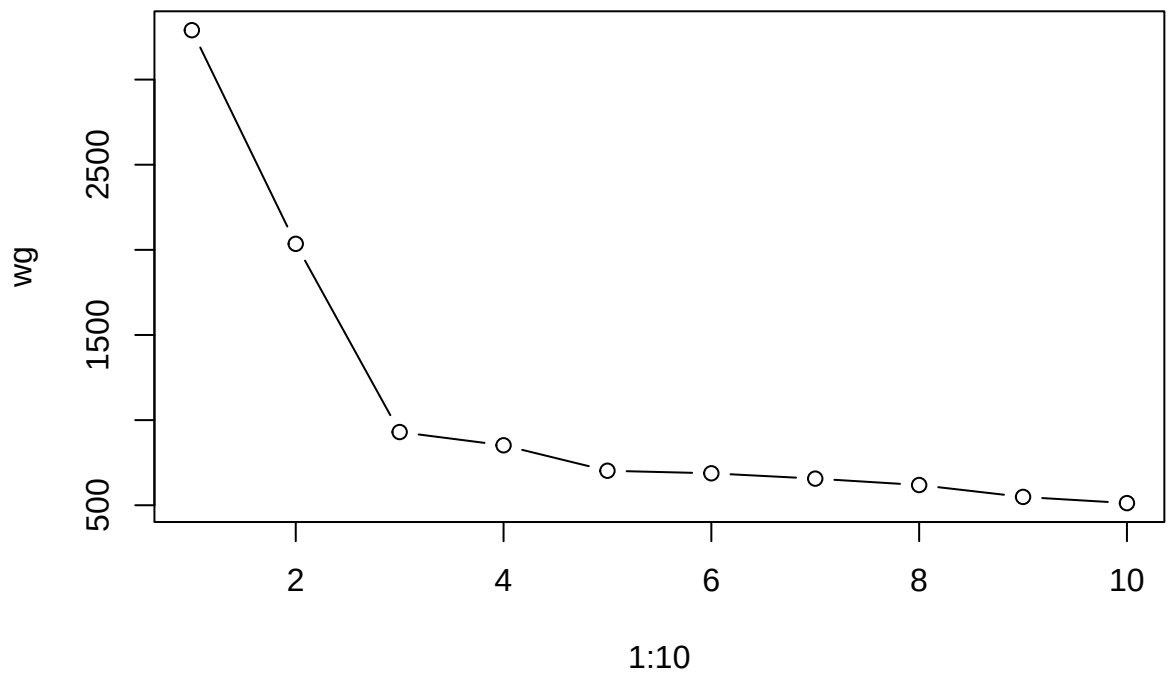
```
##    Sl_No Customer.Key Total_visits_bank disc_avg_credit disc_total_credit
## 1      1        87073                 1            Alto                 A
## 2      2        38414                 0            Alto                 A
## 3      3        17341                 1            Alto                 C
## 4      4        40496                 1      Medio/Alto                 B
## 5      5        47437                 0            Alto                 B
## 6      6        58634                 0      Medio/Alto                 A
## 7      7        48370                 0            Alto                 B
## 8      8        37376                 0      Bajo/Medio                 A
## 9      9        82490                 0            Bajo                 A
## 10    10        44770                 0            Bajo                 B
##    disc_online_visits disc_total_calls
## 1                   A                A
## 2                   C                C
## 3                   B                B
## 4                   A                B
## 5                   C                B
## 6                   A                C
## 7                   C                A
## 8                   A                A
## 9                   B                A
## 10                  A                C
```

```r
datos_sc <- scale(datos[,3:7])
```

```r
set.seed(9202)
library("cluster")
wg <- sapply(1:10, function(i) kmeans(datos_sc, i)$tot.withinss)
print(wg)
```

```
##  [1] 3290.0000 2035.4293  929.9424  852.1141  702.7257  687.7147  656.3964
##  [8]  618.8303  548.9226  512.6172
```

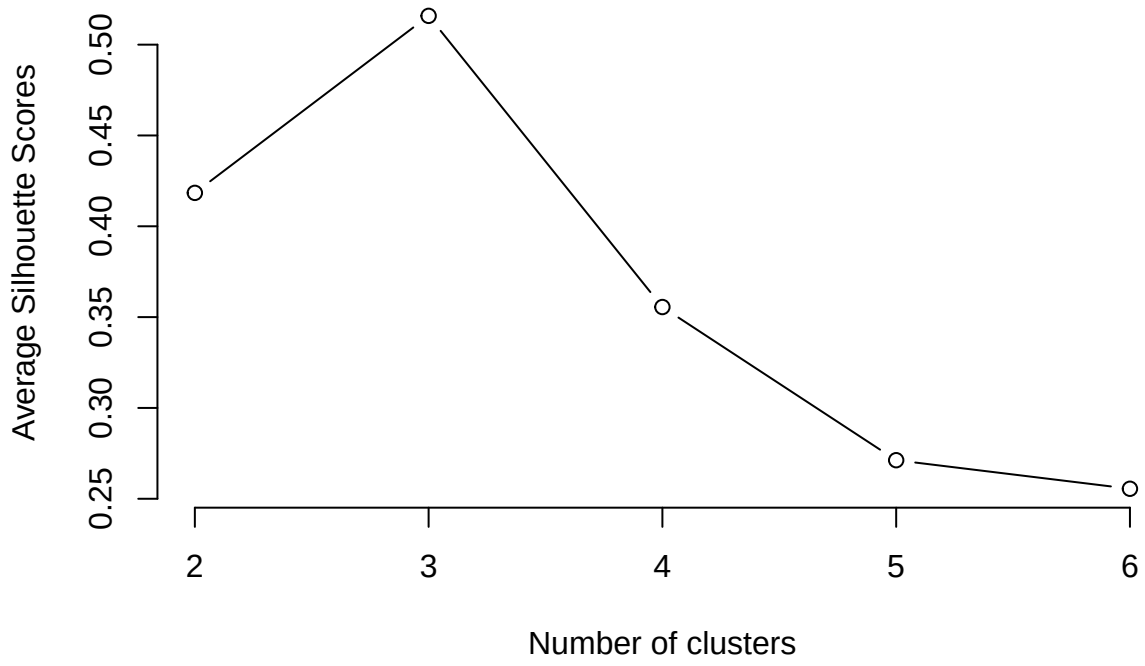
```r
plot(1:10, wg, type="b")
```



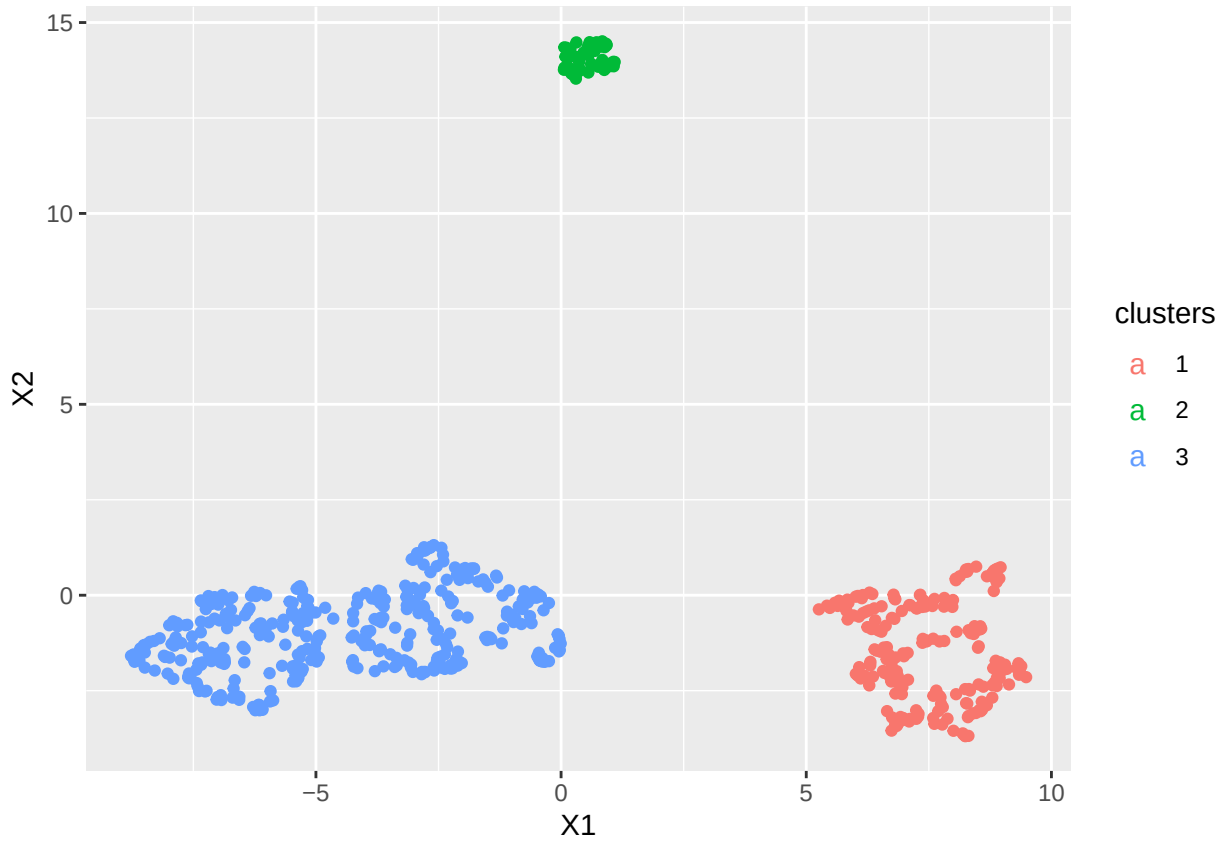

```r
silhouette_score <- function(k){
  km <- kmeans(datos_sc, centers = k, nstart=10)
  ss <- silhouette(km$cluster, dist(datos_sc))
  mean(ss[,3])
}
set.seed(9202)
k <- 2:6
avg_sil <- sapply(k, silhouette_score)
avg_sil
```

```
## [1] 0.4184053 0.5158449 0.3555452 0.2712299 0.2555066
```

```
plot(k, type='b', avg_sil, xlab='Number of clusters', ylab='Average Silhouette Scores', frame=FALSE)
```

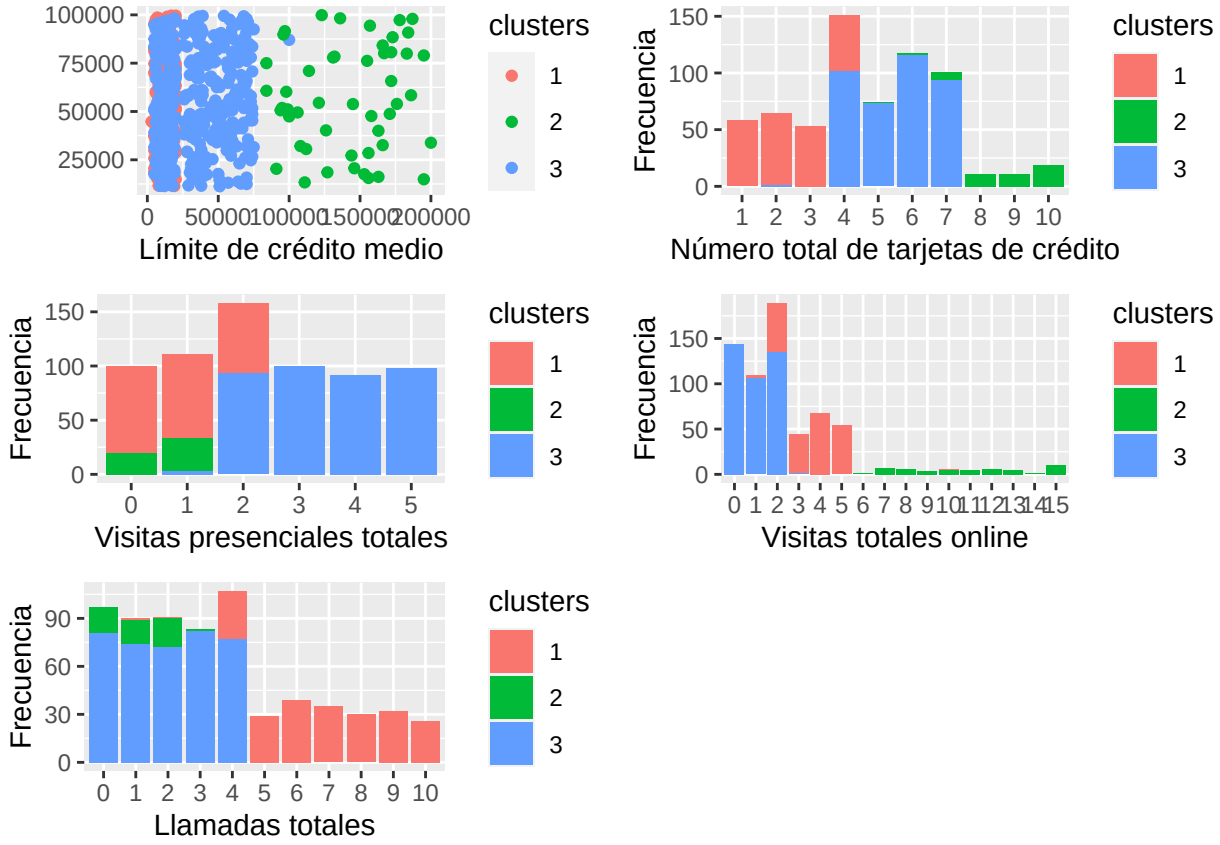



```
umap2 <- umap(datos_sc)
km.bank <- kmeans(datos_sc, 3, nstart = 10)

clusters = as.factor(km.bank$cluster)

ggplot(data.frame(umap2$layout), aes(x = X1, y = X2, color = clusters, label = rownames(datos_sc))) +
  geom_point() + ggrepel::geom_label_repel()
```



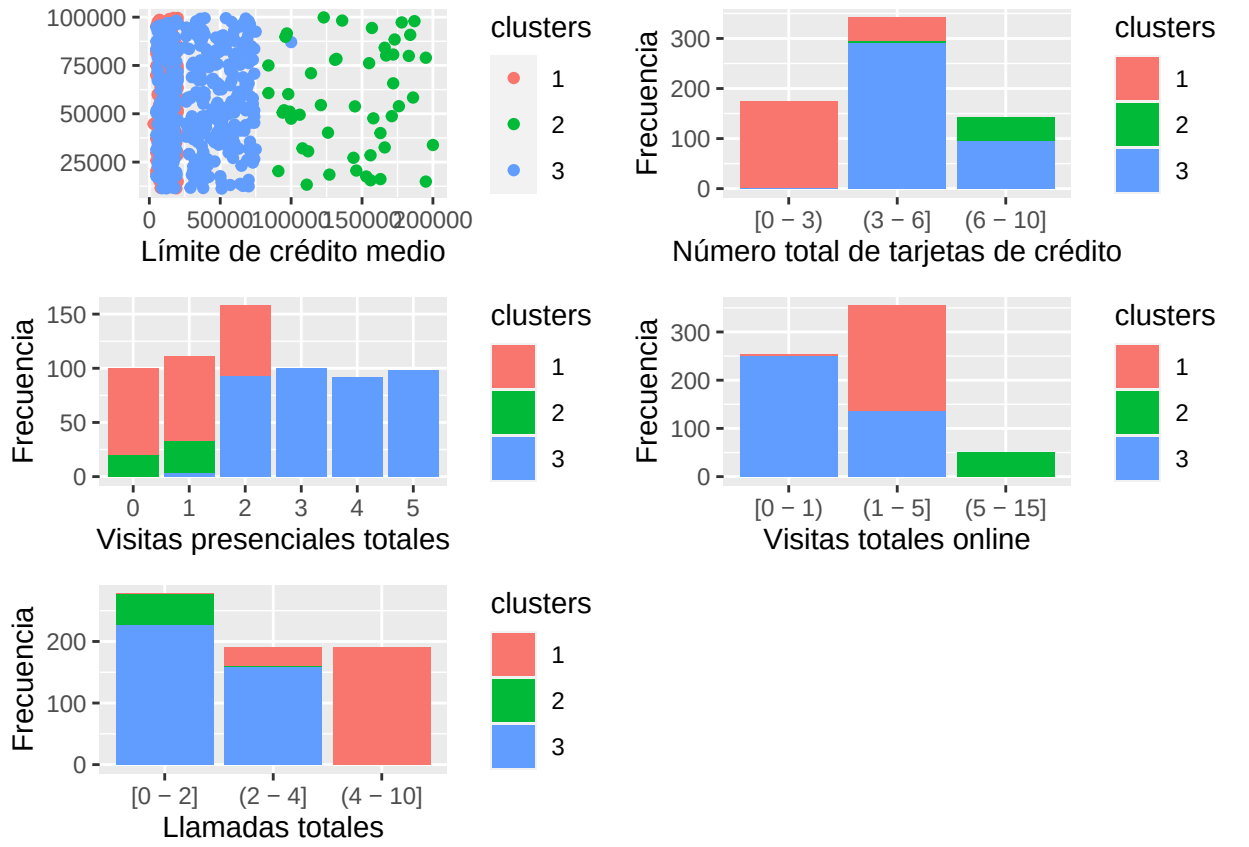

```
clusters_df <- data.frame(clusters)

datos_cl <- cbind(datos_disc, clusters_df)
```

```
gc1 <- ggplot(datos, aes(x = Avg_Credit_Limit, y = Customer.Key, color=clusters)) +
  geom_point() +
  xlab("Límite de crédito medio") +
  ylab("")
gc2 <- ggplot(datos, aes(x = as.factor(Total_Credit_Cards), fill = clusters)) +
  geom_bar(stat = "count") +
  xlab("Número total de tarjetas de crédito") +
  ylab("Frecuencia")
gc3 <- ggplot(datos, aes(x = as.factor(Total_visits_bank), fill = clusters)) +
  geom_bar(stat="count") +
  xlab("Visitas presenciales totales") +
  ylab("Frecuencia")
gc4 <- ggplot(datos, aes(x = as.factor(Total_visits_online), fill = clusters)) +
  geom_bar(stat="count") +
  xlab("Visitas totales online") +
  ylab("Frecuencia")
gc5 <- ggplot(datos, aes(x = as.factor(Total_calls_made), fill = clusters)) +
  geom_bar(stat="count") +
  xlab("Llamadas totales") +
  ylab("Frecuencia")

ggarrange(gc1,gc2,gc3,gc4,gc5, ncol=2, nrow=3, size = c(5,5,5,5,5))
```




```
gc1 <- ggplot(datos, aes(x = Avg_Credit_Limit, y = Customer.Key, color=clusters)) +
  geom_point() +
  xlab("Límite de crédito medio") +
  ylab("")

levels(datos_disc$disc_total_credit) <- c("[0 - 3)", "(3 - 6]", "(6 - 10]")
gc2 <- ggplot(datos_disc, aes(x = as.factor(disc_total_credit), fill = clusters)) +
  geom_bar(stat = "count") +
  xlab("Número total de tarjetas de crédito") +
  ylab("Frecuencia")

gc3 <- ggplot(datos_disc, aes(x = as.factor(Total_visits_bank), fill = clusters)) +
  geom_bar(stat="count") +
  xlab("Visitas presenciales totales") +
  ylab("Frecuencia")

levels(datos_disc$disc_online_visits) <- c("[0 - 1)", "(1 - 5]", "(5 - 15]")
gc4 <- ggplot(datos_disc, aes(x = as.factor(disc_online_visits), fill = clusters)) +
  geom_bar(stat="count") +
  xlab("Visitas totales online") +
  ylab("Frecuencia")

levels(datos_disc$disc_total_calls) <- c("[0 - 2]", "(2 - 4]", "(4 - 10]")
gc5 <- ggplot(datos_disc, aes(x = as.factor(disc_total_calls), fill = clusters)) +
  geom_bar(stat="count") +
  xlab("Llamadas totales") +
```

```
  ylab("Frecuencia")

ggarrange(gc1,gc2,gc3,gc4,gc5, ncol=2, nrow=3)
```




```
# ggarrange(gc1,gc2,gc3,gc4,gc5, ncol=2, nrow=3, common.legend = TRUE, theme = theme(axis.text = elemen
```

Podemos observar que existen tres grupos bien diferenciados, el rojo, verde y azul. El grupo representado con valor rojo son los individuos con menor límite de crédito medio (de 0 a 18000), son los que menos tarjetas de crédito poseen y los que menos visitas presenciales hacen, estan en la media en cuanto a visitas online (entre 1 y 5 veces) y son los que más llamadas hacen.

El grupo azul son los individuos cuyo límite de crédito medio está entre los 0 y los 50000, tienen entre 4 y 7 tarjetas de crédito, son los que menos visitas online y menos llamadas hacen los que más acuden presencialmente al banco.

El grupo verde, por último son los que más crédito límite medio tiene (entre 100000 y 200000), los que más tarjetas de crédito tienen y los que más visitas online hacen (esto último con mucha diferencia) y los que menos visitas presenciales y llamadas hacen.