Tarea R #7 (sol)

Diego Alonso Alfaro Bergueiro

6 de noviembre de 2016

Nota Aclaratoria

Esta representa solo una posible solución. No hay expectativa alguna de que su trabajo haya sido exáctamente igual a este, lo importante es que hayan desarrollado el código para hacer lo que se les pidió y que hayan podido describir lo que estaban haciendo y los resultados.

Análisis del Problema

Entre los problemas más comunes que afectan a nuestra sociedad actualmente, están el sobrepeso y el colesterol alto. Estos problemas suelen generar otros problemas, como por ejemplo problemas cardiacos, hipertensión y diabetes. Debido a la alta propensidad de sufrir enfermedades, las personas con sobrepeso u obesas suelen ser un público "menos atractivo" para compañías de seguros, pues estos representan un riesgo mayor. Sin embargo, para compañías que se dediquen a vender productos para bajar de peso o para cadenas de gimnasios, pueden representar un público meta muy atractivo.

Debido a estas consideraciones, un análisis de agrupamiento para encontrar personas con características de peso y colesterol similares puede ser útil desde ambas perspectivas: para una aseguradora puede representar cierto nivel de riesgo y puede preferir mantenerse alejado de algunos clientes, mientras que para gimnasios o compañías que ofrecen productos para mejorar la salud y bajar de peso pueden representar esos clientes sobre los cuales enfocarse.

Entendimiento de los Datos

Para intentar resolver este problema, se cuenta con un conjunto de datos con tres variables:

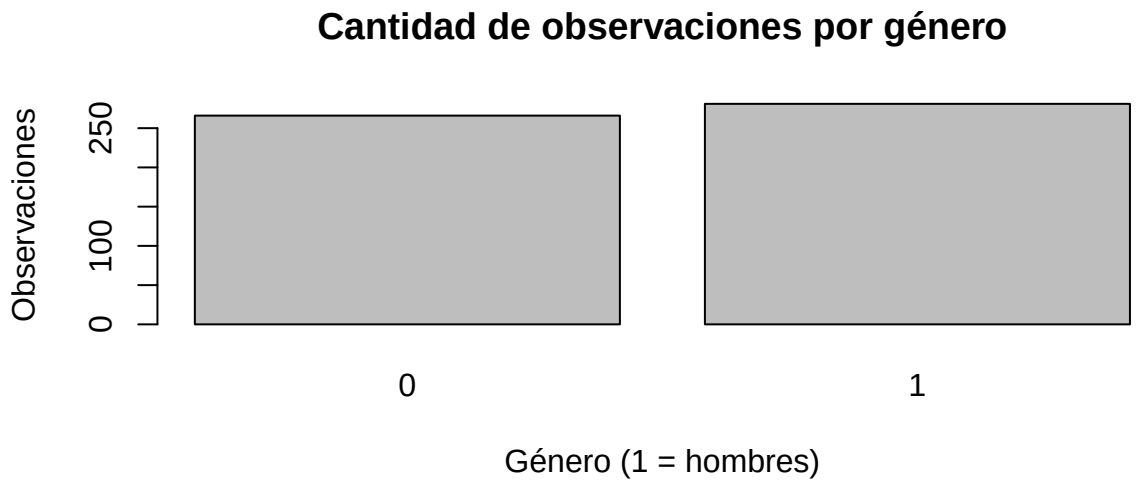
- Peso: cuantitativa, mide el peso de cada observación.
- Colesterol: cuantitativa, mide el nivel de colesterol de cada observación.
- Género: cualitativa, 0 para representar a las mujeres y 1 para los hombres.

Exploración de los Datos

```
library(cluster)

setwd('K:/CENFOTEC/Maestría - Introducción a Minería de Datos/Materiales de las Clases/Semana 11/dataset

datos <- read.csv('Chapter06DataSet.csv')

##Normalizar las primeras dos columnas entre 0 y 1:
datos$Weight <- (datos$Weight - min(datos$Weight)) / (max(datos$Weight) - min(datos$Weight))
datos$Cholesterol <- (datos$Cholesterol - min(datos$Cholesterol)) / (max(datos$Cholesterol) - min(datos$
```

Luego de cargar los datos, podemos comparar la cantidad de observaciones que hay por género:

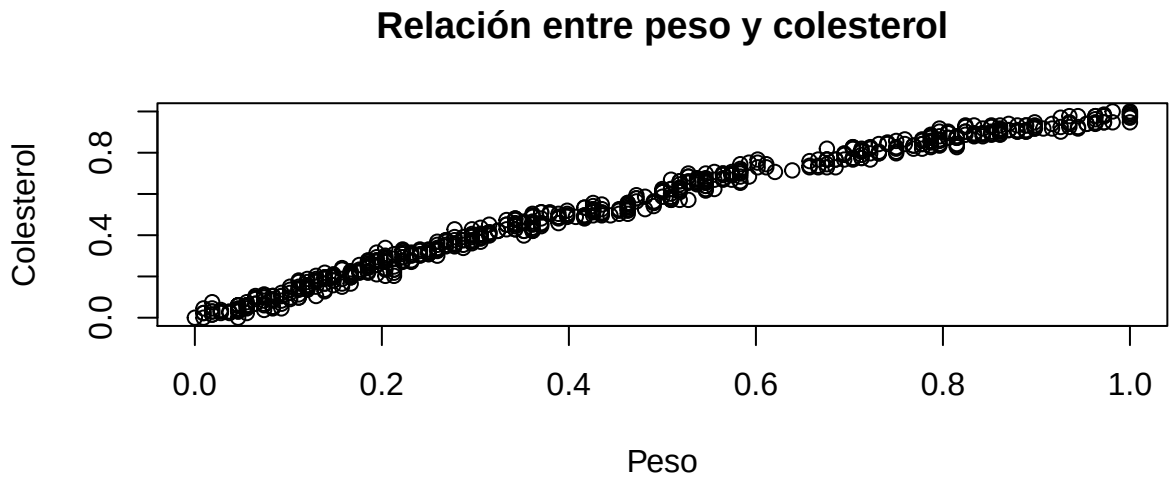
```r
barplot(table(datos$Gender),
        main = 'Cantidad de observaciones por género',
        xlab = 'Género (1 = hombres)',
        ylab = 'Observaciones')
```



En el gráfico anterior, se puede apreciar que hay una leve diferencia entre la cantidad de hombres y mujeres a favor de los hombres.

Se puede analizar también la relación que hay entre las variables colesterol y peso:

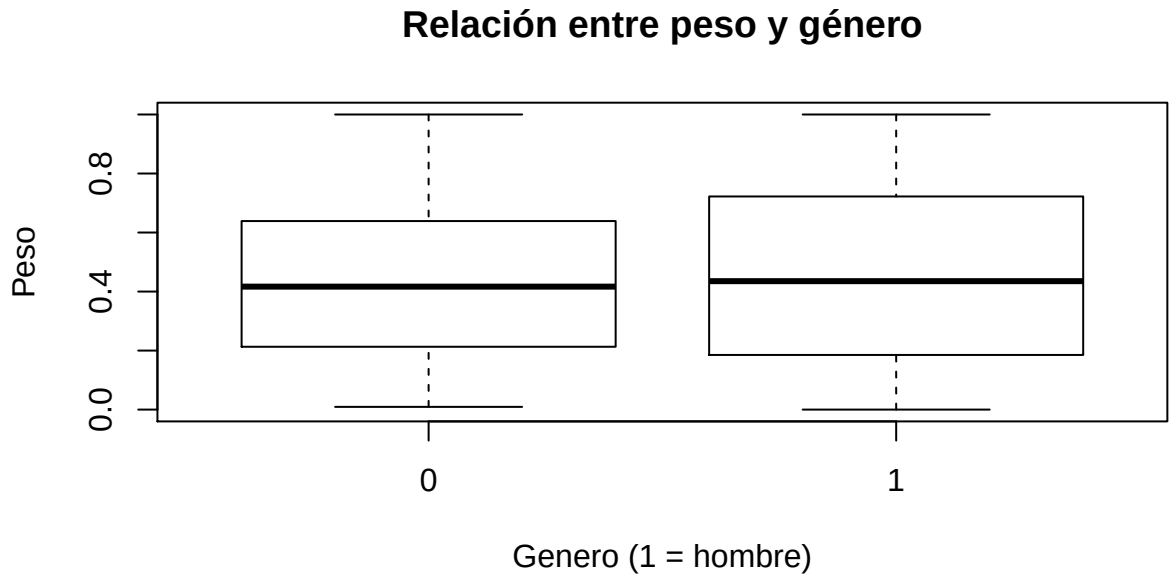
```r
plot(datos$Weight,
     datos$Cholesterol,
     main = 'Relación entre peso y colesterol',
     xlab = 'Peso',
     ylab = 'Colesterol')
```



Como se puede apreciar, hay una relación casi lineal (hay una tendencia a formar un arco) entre ambas variables, y se puede sacar la conclusión que a mayor peso, mayor nivel de colesterol (y viceversa).

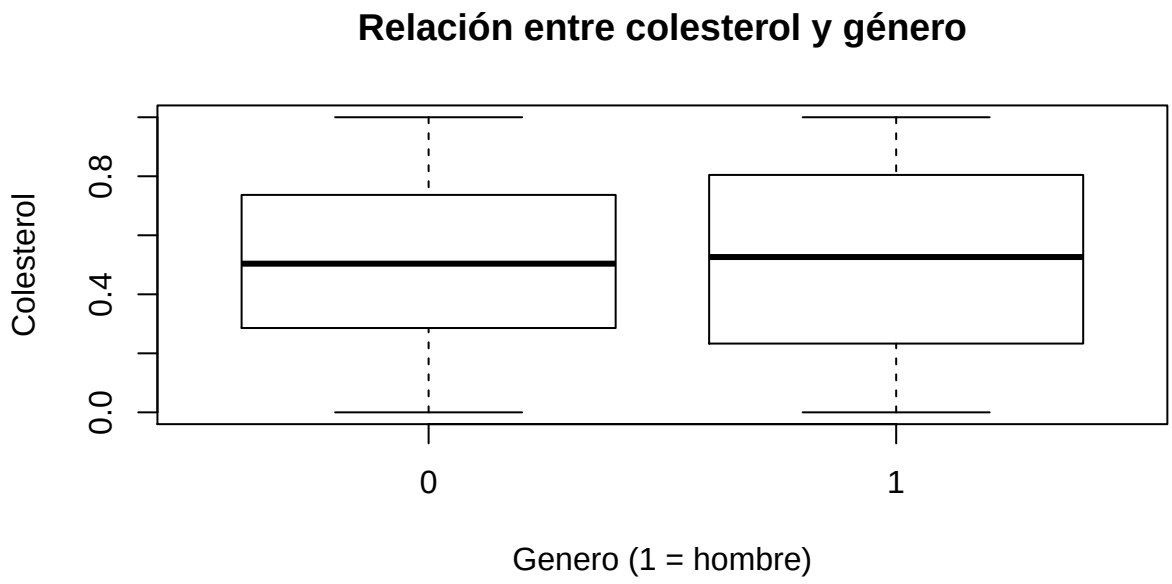
Finalmente, podemos observar la distribución de peso por género:

```
boxplot(datos$Weight ~ factor(datos$Gender),
     main = 'Relación entre peso y género',
     xlab = 'Genero (1 = hombre)',
     ylab = 'Peso')
```



y la distribución de colesterol por género:

```
boxplot(datos$Cholesterol ~ factor(datos$Gender),
     main = 'Relación entre colesterol y género',
     xlab = 'Genero (1 = hombre)',
     ylab = 'Colesterol')
```

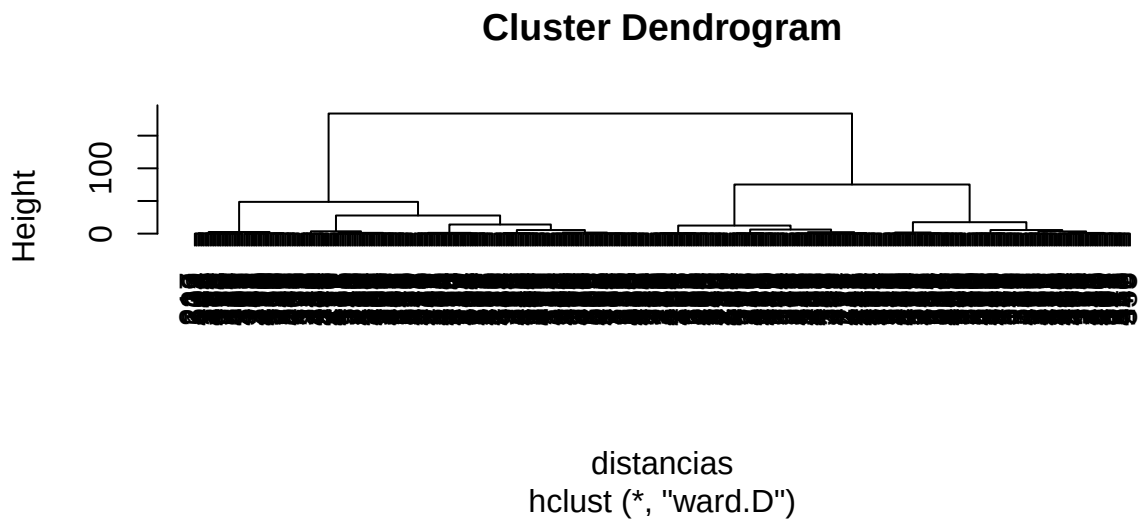


En estos dos gráficos se puede apreciar que en promedio ambos géneros están muy parecidos, pero los hombres tienden a tener mayor variabilidad en ambas variables (la caja es más ancha).

Creación del Modelo

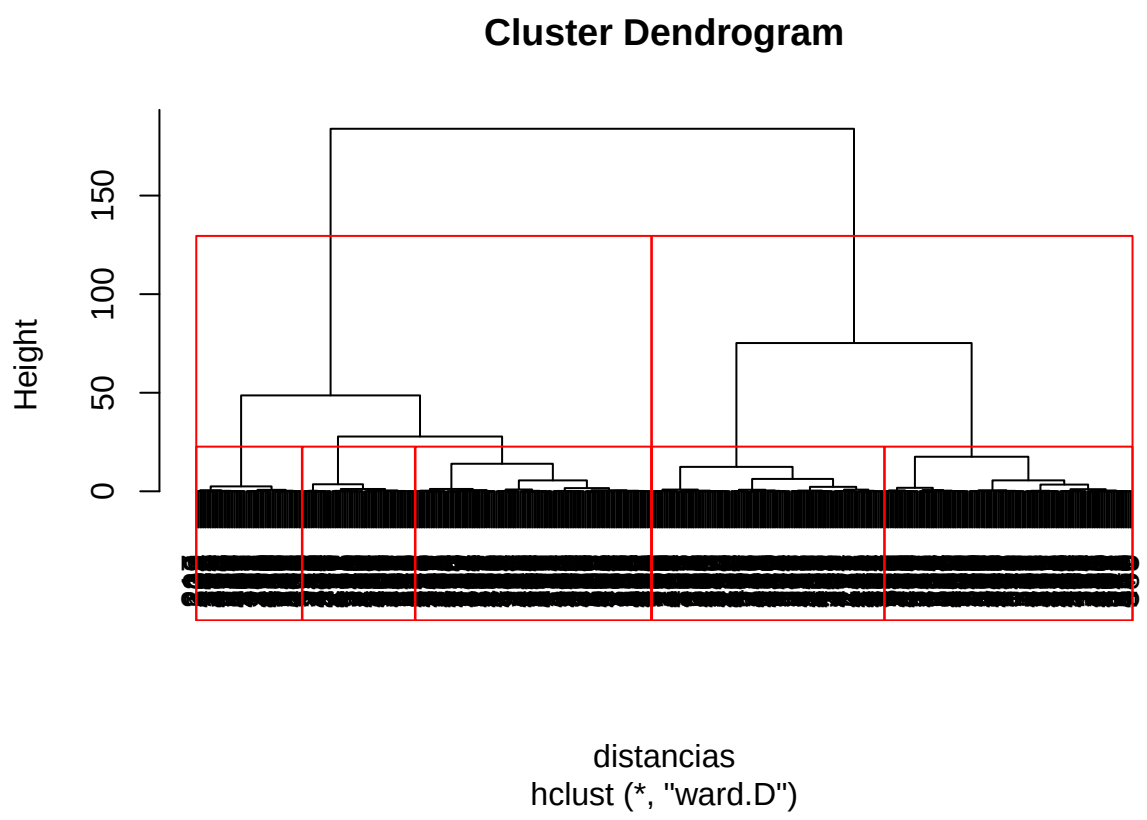
Para determinar la cantidad de clústeres que se pueden crear, se procede a hacer un agrupamiento jerárquico:

```
distancias <- dist(datos, method="euclidean")

datos.jerarquico <- hclust(distancias, method="ward.D")

plot(datos.jerarquico)
```



Determinando las diferentes alturas, se pueden crear desde 2 grupos hasta 6. En realidad, ya la división en 5 ó 6 grupos es un poco "estrecha", pero definitiamente para más de 6 grupos la división es sumamente difícil de hacer. Para este análisis, vamos a utilizar 5

```
plot(datos.jerarquico)
rect.hclust(datos.jerarquico, k = 2, border = "red")
rect.hclust(datos.jerarquico, k = 5, border = "red")
```




```
cluster.jerarquico <- factor(cutree(datos.jerarquico, k=5))
```

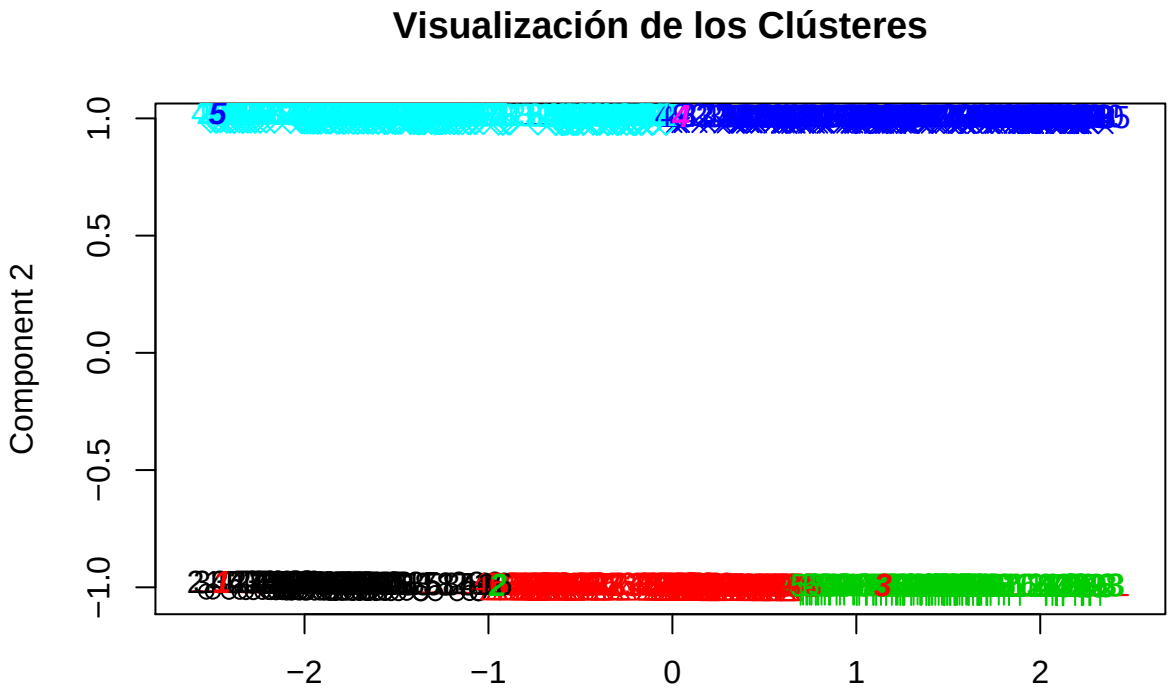
Luego de hacer el análisis jerárquico, se procede a hacer el análisis utilizando el algoritmo KMeans con 5 centros:

```
set.seed(352345) #necesario para replicabilidad
km <- kmeans(datos, centers = 5)

cluster.kmeans <- factor(km$cluster)
```

El resultado se puede visualizar así: (Los colores de los puntos representan el grupo al cual pertenecen)

```
#Visualizar los Clústeres
clusplot(datos,
         km$cluster,
         col.p = km$cluster,
```

```
        color=TRUE,
        shade=TRUE,
        labels=2,
        lines=0,
        main = 'Visualización de los Clústeres')
```



Evaluación

Con el fin de comparar ambos agrupamientos, podemos generar tablas resumen para comparar los valores promedios de cada variable en cada grupo:

```
resultado.jerarquico <- rbind(tapply(datos$Weight, cluster.jerarquico, mean),
                              tapply(datos$Cholesterol, cluster.jerarquico, mean),
                              tapply(datos$Gender, cluster.jerarquico, mean))

rownames(resultado.jerarquico) <- c('Weight', 'Cholesterol', 'Gender')

resultado.jerarquico
```

```
##                    1         2         3         4         5
## Weight      0.1790577 0.4231079 0.7070881 0.8470729 0.1245791
## Cholesterol 0.2277200 0.5209764 0.7777029 0.8958283 0.1557302
## Gender      1.0000000 0.0000000 1.0000000 0.0000000 0.0000000
```

Los grupos creados se pueden resumir así:

- Grupo 1: Hombres con peso y colesterol bajo.
- Grupo 2: Mujeres con peso y colesterol a nivel medio.
- Grupo 3: Hombres con peso y colesterol alto.
- Grupo 4: Mujeres con peso y colesterol alto.
- Grupo 5: Mujeres con peso y colesterol bajo.

De manera similar, se puede generar un resumen para los grupos creados con el algoritmo KMeans:

```
resultado.kmeans <- rbind(tapply(datos$Weight, cluster.kmeans, mean),
                          tapply(datos$Cholesterol, cluster.kmeans, mean),
                          tapply(datos$Gender, cluster.kmeans, mean))

rownames(resultado.kmeans) <- c('Weight', 'Cholesterol', 'Gender')

resultado.kmeans
```

```
##                    1         2         3         4         5
## Weight      0.8364198 0.4562132 0.1624851 0.1875493 0.7173942
## Cholesterol 0.8864206 0.5593423 0.2081817 0.2377753 0.7872180
## Gender      0.0000000 0.0000000 0.0000000 1.0000000 1.0000000
```

Los grupos creados se pueden resumir así:

- Grupo 1: Mujeres con peso y colesterol alto.
- Grupo 2: Mujeres con peso y colesterol medio.
- Grupo 3: Mujeres con peso y colesterol bajo.
- Grupo 4: Hombres con peso y colesterol bajo.
- Grupo 5: Hombres con peso y colesterol alto.

Resultados

De los resúmenes anteriores, podemos sacar la conclusión que con los datos de peso y colesterol normalizados, y con el género, independientemente del algoritmo de agrupamiento (jerárquico o KMeans), los resultados son basicamente los mismos. Si bien es cierto que el número de grupo puede variar, en general al dividir en 5 grupos vamos a tener 3 grupos para mujeres con diferentes niveles de peso y colesterol (alto - medio - bajo) y 2 grupos para los hombres con diferentes niveles de colesterol (alto - bajo).

Ambos modelos dan resultados sumamente válidos para cualquiera de las dos perspectivas mencionadas en el análisis del problema.