# Mineria de Datos

Luis Enrique Olascoaga Dominguez

22/11/2022

## PEC2

Obtenemos el set de datos llamado “Hawks”

```
#Obtenemos librerias a utilizar
if (!require('cluster')) install.packages('cluster')
```

```
## Loading required package: cluster
```

```
library(cluster)
if (!require('Stat2Data')) install.packages('Stat2Data')
```

```
## Loading required package: Stat2Data
```

```
library(Stat2Data)
if (!require('ggplot2')) install.packages('ggplot2')
```

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

```
library(ggplot2)
if (!require('fpc')) install.packages('fpc')
```

```
## Loading required package: fpc
```

```
library(fpc)

#Obtenemos el set de datos llamado "Hawks"
data("Hawks")
summary(Hawks)
```

```
##      Month             Day             Year       CaptureTime   ReleaseTime
##  Min.   : 8.000   Min.   : 1.00   Min.   :1992   11:35  : 14          :842
##  1st Qu.: 9.000   1st Qu.: 9.00   1st Qu.:1995   13:30  : 14   11:00  :  2
##  Median :10.000   Median :16.00   Median :1999   11:45  : 13   11:35  :  2
##  Mean   : 9.843   Mean   :15.74   Mean   :1998   12:10  : 13   12:05  :  2
##  3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:2001   14:00  : 13   12:50  :  2
##  Max.   :11.000   Max.   :31.00   Max.   :2003   13:05  : 12   13:32  :  2
##                                                  (Other):829   (Other): 56
```

```
##       BandNumber  Species  Age     Sex          Wing           Weight
##            :  2   CH: 70   A:224    :576   Min.   : 37.2   Min.   :  56.0
##  1142-09240:  1   RT:577   I:684   F:174   1st Qu.:202.0   1st Qu.: 185.0
##  1142-09241:  1   SS:261           M:158   Median :370.0   Median : 970.0
##  1142-09242:  1                           Mean   :315.6   Mean   : 772.1
##  1142-18229:  1                           3rd Qu.:390.0   3rd Qu.:1120.0
##  1142-19209:  1                           Max.   :480.0   Max.   :2030.0
##  (Other)   :901                           NA's   :1       NA's   :10
##      Culmen          Hallux            Tail        StandardTail
##  Min.   : 8.6   Min.   :  9.50   Min.   :119.0   Min.   :115.0
##  1st Qu.:12.8   1st Qu.: 15.10   1st Qu.:160.0   1st Qu.:162.0
##  Median :25.5   Median : 29.40   Median :214.0   Median :215.0
##  Mean   :21.8   Mean   : 26.41   Mean   :198.8   Mean   :199.2
##  3rd Qu.:27.3   3rd Qu.: 31.40   3rd Qu.:225.0   3rd Qu.:226.0
##  Max.   :39.2   Max.   :341.40   Max.   :288.0   Max.   :335.0
##  NA's   :7      NA's   :6                        NA's   :337
##      Tarsus        WingPitFat        KeelFat           Crop
##  Min.   :24.70   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:55.60   1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:0.0000
##  Median :79.30   Median :1.0000   Median :2.000   Median :0.0000
##  Mean   :71.95   Mean   :0.7922   Mean   :2.184   Mean   :0.2345
##  3rd Qu.:87.00   3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:0.2500
##  Max.   :94.00   Max.   :3.0000   Max.   :4.000   Max.   :5.0000
##  NA's   :833     NA's   :831      NA's   :341     NA's   :343
```

```
#{r pressure, echo=FALSE}
```

## EDA (exploratory data analysis)

Obtenemos nombre de las columnas y su tipo de dato

```
str(Hawks)
```

```
## 'data.frame':    908 obs. of  19 variables:
##  $ Month       : int  9 9 9 9 9 9 9 9 9 9 ...
##  $ Day         : int  19 22 23 23 27 28 28 29 29 30 ...
##  $ Year        : int  1992 1992 1992 1992 1992 1992 1992 1992 1992 1992 ...
##  $ CaptureTime : Factor w/ 308 levels " ","1:15","1:31",..: 181 25 138 42 62 71 181 88 261 192 ...
##  $ ReleaseTime : Factor w/ 60 levels "","  ","10:20",..: 1 2 2 2 2 2 2 2 2 2 ...
##  $ BandNumber  : Factor w/ 907 levels " ","1142-09240",..: 856 857 858 809 437 280 859 860 861 281 .
##  $ Species     : Factor w/ 3 levels "CH","RT","SS": 2 2 2 1 3 2 2 2 2 2 ...
##  $ Age         : Factor w/ 2 levels "A","I": 2 2 2 2 2 2 2 1 1 2 ...
##  $ Sex         : Factor w/ 3 levels "","F","M": 1 1 1 2 2 1 1 1 1 1 ...
##  $ Wing        : num  385 376 381 265 205 412 370 375 412 405 ...
##  $ Weight      : int  920 930 990 470 170 1090 960 855 1210 1120 ...
##  $ Culmen      : num  25.7 NA 26.7 18.7 12.5 28.5 25.3 27.2 29.3 26 ...
##  $ Hallux      : num  30.1 NA 31.3 23.5 14.3 32.2 30.1 30 31.3 30.2 ...
##  $ Tail        : int  219 221 235 220 157 230 212 243 210 238 ...
##  $ StandardTail: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ Tarsus      : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ WingPitFat  : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ KeelFat     : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ Crop        : num  NA NA NA NA NA NA NA NA NA NA ...
```

Observamos en el preview de los datos que las columnas StandardTail, Tarsus, WingPitFat, KeelFat y Crop, contienen datos NA, vamos a revisarlos y a removerlos ya que no se ocuparan para este análsis. Las columnas numericas a utilizar serán:

Wing: Longitud (en mm) de la pluma principal del ala desde la punta hasta la muñeca a la que se une
Weight: Peso corporal (en gm)
Culmen: Longitud (en mm) del pico superior desde la punta hasta donde choca con la parte carnosa del ave
Hallux: Longitud (en mm) de la garra asesina

Y la calumna que se utilizará para comparar nuestros clusters posteriormente será: Species: CH=Halcón de Cooper, RT=Colirrojo, SS=Gavilán

La descripción del layout fue obtenida desde: https://vincentarelbundock.github.io/Rdatasets/doc/Stat2Data/Hawks.html

Para nuestro análsis vamos a generar 1 dataframe llamado: hawks_k_means: Tendra las columnas Wing, Weigh, Culmen y Hallux

Generamos el dataframe k-means

```
hawks_k_means <- na.omit(Hawks[,10:13])
summary(hawks_k_means)
```

```
##       Wing           Weight           Culmen          Hallux
##  Min.   : 37.2   Min.   :  56.0   Min.   : 8.60   Min.   :  9.50
##  1st Qu.:202.0   1st Qu.: 185.0   1st Qu.:12.80   1st Qu.: 15.10
##  Median :370.0   Median : 970.0   Median :25.50   Median : 29.40
##  Mean   :315.9   Mean   : 771.6   Mean   :21.81   Mean   : 26.41
##  3rd Qu.:390.0   3rd Qu.:1120.0   3rd Qu.:27.35   3rd Qu.: 31.40
##  Max.   :480.0   Max.   :2030.0   Max.   :39.20   Max.   :341.40
```

```
hawks_k_original <- na.omit(Hawks[,7:13])
```

Cuando reducimos la dimensionalidad de nuestro dataframe, solo dejamos las 4 columnas a ocupar para el modelo k-means. En el summary encontramos que no hay valores en NA o nulos y todos los valores son númericos.

Revisamos nuevamente el set de datos original

```
summary(Hawks)
```

```
##      Month             Day             Year       CaptureTime   ReleaseTime
##  Min.   : 8.000   Min.   : 1.00   Min.   :1992   11:35  : 14          :842
##  1st Qu.: 9.000   1st Qu.: 9.00   1st Qu.:1995   13:30  : 14   11:00  :  2
##  Median :10.000   Median :16.00   Median :1999   11:45  : 13   11:35  :  2
##  Mean   : 9.843   Mean   :15.74   Mean   :1998   12:10  : 13   12:05  :  2
##  3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:2001   14:00  : 13   12:50  :  2
##  Max.   :11.000   Max.   :31.00   Max.   :2003   13:05  : 12   13:32  :  2
##                                                  (Other):829   (Other): 56
##       BandNumber  Species  Age     Sex          Wing           Weight
##            :  2   CH: 70   A:224    :576   Min.   : 37.2   Min.   :  56.0
##  1142-09240:  1   RT:577   I:684   F:174   1st Qu.:202.0   1st Qu.: 185.0
##  1142-09241:  1   SS:261           M:158   Median :370.0   Median : 970.0
##  1142-09242:  1                           Mean   :315.6   Mean   : 772.1
##  1142-18229:  1                           3rd Qu.:390.0   3rd Qu.:1120.0
```

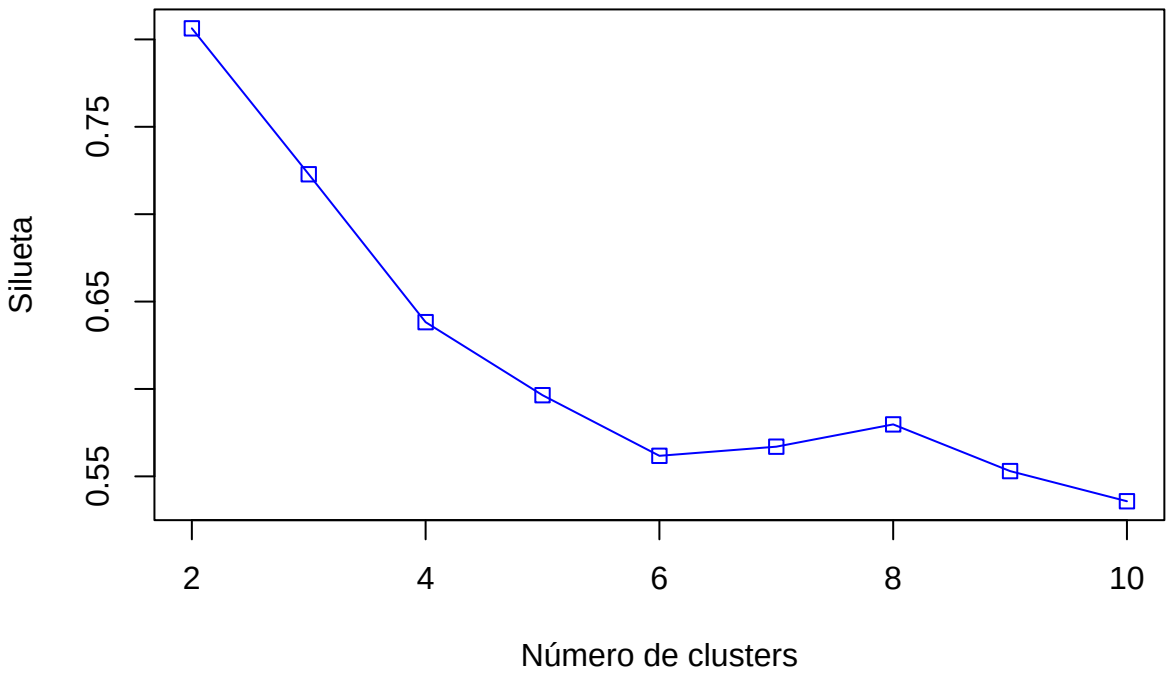
```
##  1142-19209:  1                                Max.   :480.0   Max.   :2030.0
##  (Other)   :901                                NA's   :1       NA's   :10
##      Culmen          Hallux             Tail         StandardTail
##  Min.   : 8.6   Min.   :  9.50   Min.   :119.0   Min.   :115.0
##  1st Qu.:12.8   1st Qu.: 15.10   1st Qu.:160.0   1st Qu.:162.0
##  Median :25.5   Median : 29.40   Median :214.0   Median :215.0
##  Mean   :21.8   Mean   : 26.41   Mean   :198.8   Mean   :199.2
##  3rd Qu.:27.3   3rd Qu.: 31.40   3rd Qu.:225.0   3rd Qu.:226.0
##  Max.   :39.2   Max.   :341.40   Max.   :288.0   Max.   :335.0
##  NA's   :7      NA's   :6                        NA's   :337
##      Tarsus        WingPitFat        KeelFat           Crop
##  Min.   :24.70   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:55.60   1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:0.0000
##  Median :79.30   Median :1.0000   Median :2.000   Median :0.0000
##  Mean   :71.95   Mean   :0.7922   Mean   :2.184   Mean   :0.2345
##  3rd Qu.:87.00   3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:0.2500
##  Max.   :94.00   Max.   :3.0000   Max.   :4.000   Max.   :5.0000
##  NA's   :833     NA's   :831      NA's   :341     NA's   :343
```

Podemos observar que es un set de datos de un modelo supervisado pero para fines de nuestra PEC 2 vamos a ejecutar un modelo no supervisado tomando la columna "Species" como la variable que vamos a predecir. Primeramente observamos 3 valores:
CH=Halcón de Cooper RT=Colirrojo SS=Gavilán Podemos observar que hay más datos para Colirrojo y Gavilán. Vamos a ver como se comporta el algoritmo k-means con esta distribución de datos.

## k-means

Comenzamos a evaluar el número de cluster que necesitamos para nuestra variable "k"

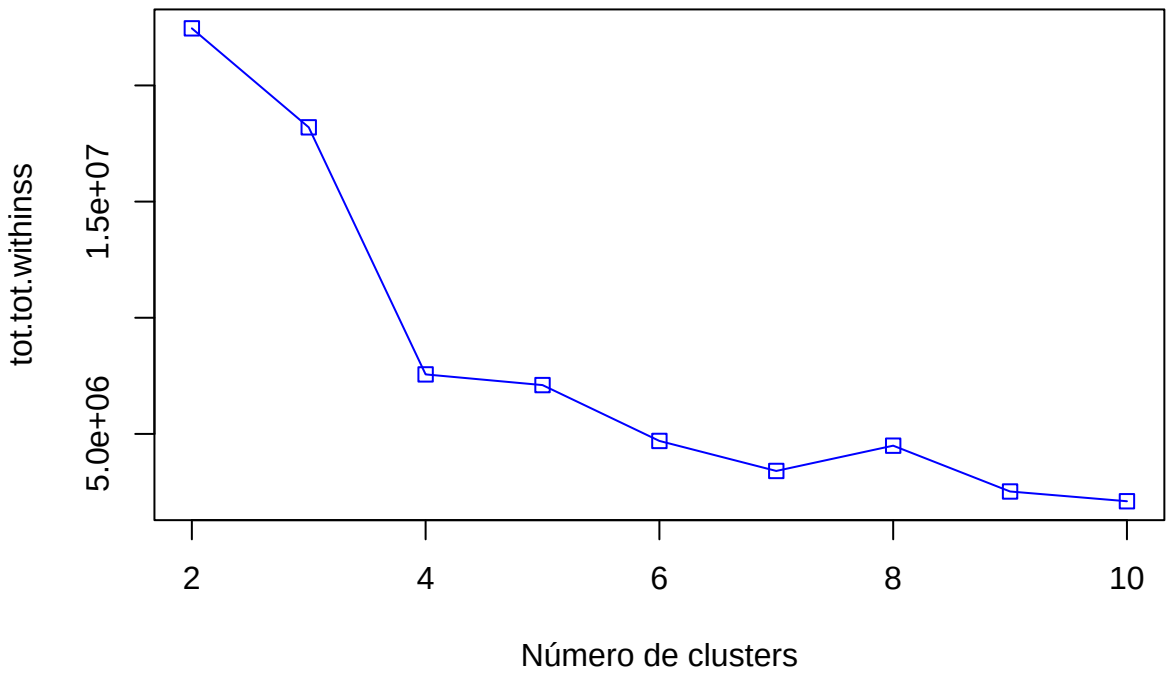
```
distance <- daisy(hawks_k_means)
resultados <- rep(0, 10)
for (i in c(2,3,4,5,6,7,8,9,10))
{
  fit           <- kmeans(hawks_k_means, i)
  y_cluster     <- fit$cluster
  sk            <- silhouette(y_cluster, distance)
  resultados[i] <- mean(sk[,3])
}
plot(2:10,resultados[2:10],type="o",col="blue",pch=0,xlab="Número de clusters",ylab="Silueta")
```




De acuerdo a los valores de las siluetas, el mejor valor para "k" es 2 a pesar que hay 3 tipos de especie.

Vamos a verificar el número de cluster mediante el procedimiento elbow (codo).

```
resultados <- rep(0, 10)
  for (i in c(2,3,4,5,6,7,8,9,10))
  {
    fit           <- kmeans(hawks_k_means, i)
    resultados[i] <- fit$tot.withinss
  }
plot(2:10,resultados[2:10],type="o",col="blue",pch=0,xlab="Número de clusters",ylab="tot.tot.withinss")
```




Como observamos, de acuerdo al método de elbow, el valor mas optimo para "k" podría ser 4 o 6.

Vamos a utilizar los criterios, silueta media ("asw") y Calinski-Harabasz ("ch").

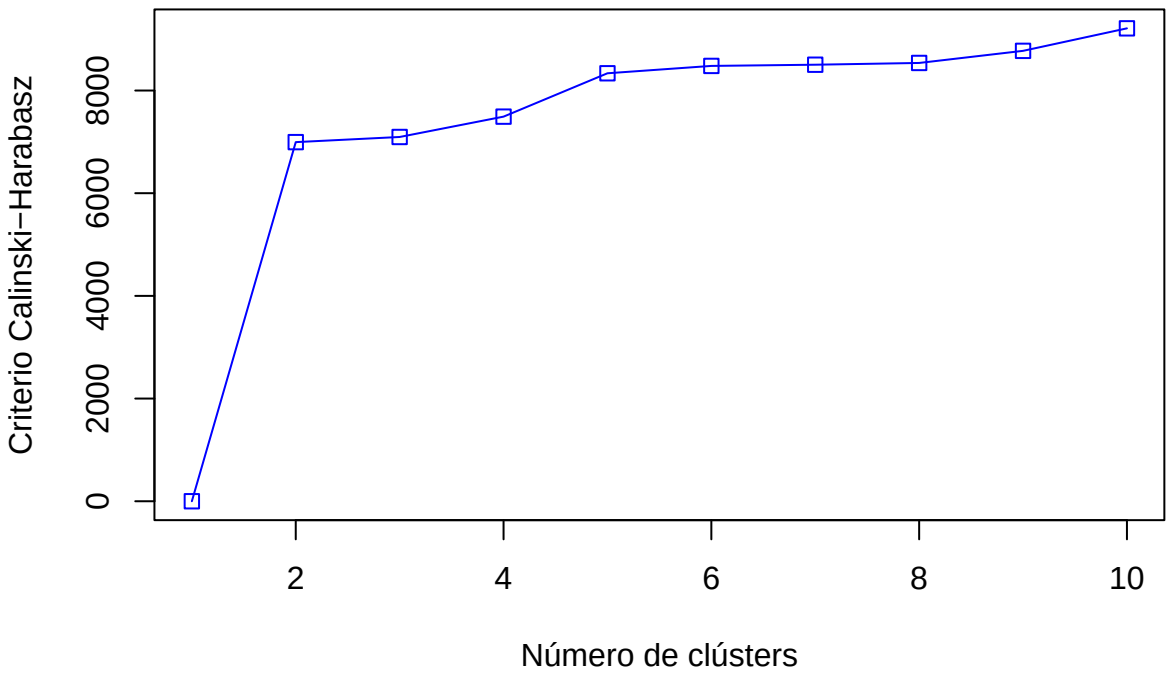
```
if (!require('fpc')) install.packages('fpc')
library(fpc)

fit_ch  <- kmeansruns(hawks_k_means, krange = 1:10, criterion = "ch")
fit_asw <- kmeansruns(hawks_k_means, krange = 1:10, criterion = "asw")

print(fit_ch$bestk)
```

```
## [1] 10
```

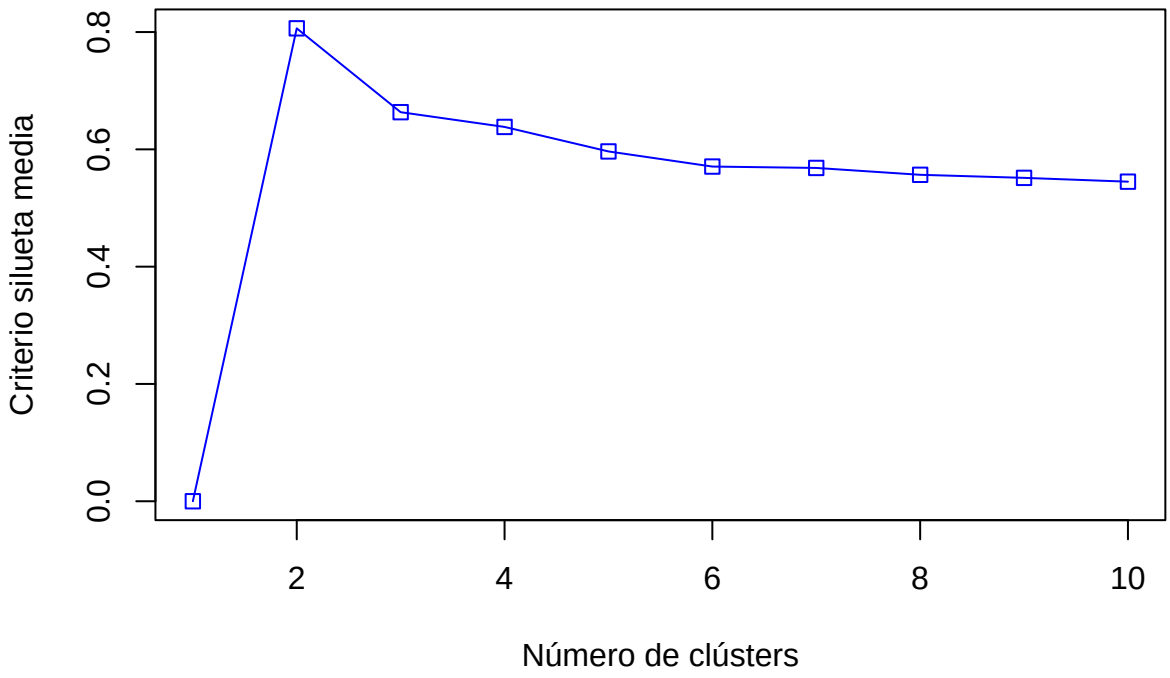
```
print(fit_asw$bestk)
```

```
## [1] 2
```

```
plot(1:10,fit_ch$crit,type="o",col="blue",pch=0,xlab="Número de clústers",ylab="Criterio Calinski-Haraba
```




```
plot(1:10,fit_asw$crit,type="o",col="blue",pch=0,xlab="Número de clústers",ylab="Criterio silueta media
```



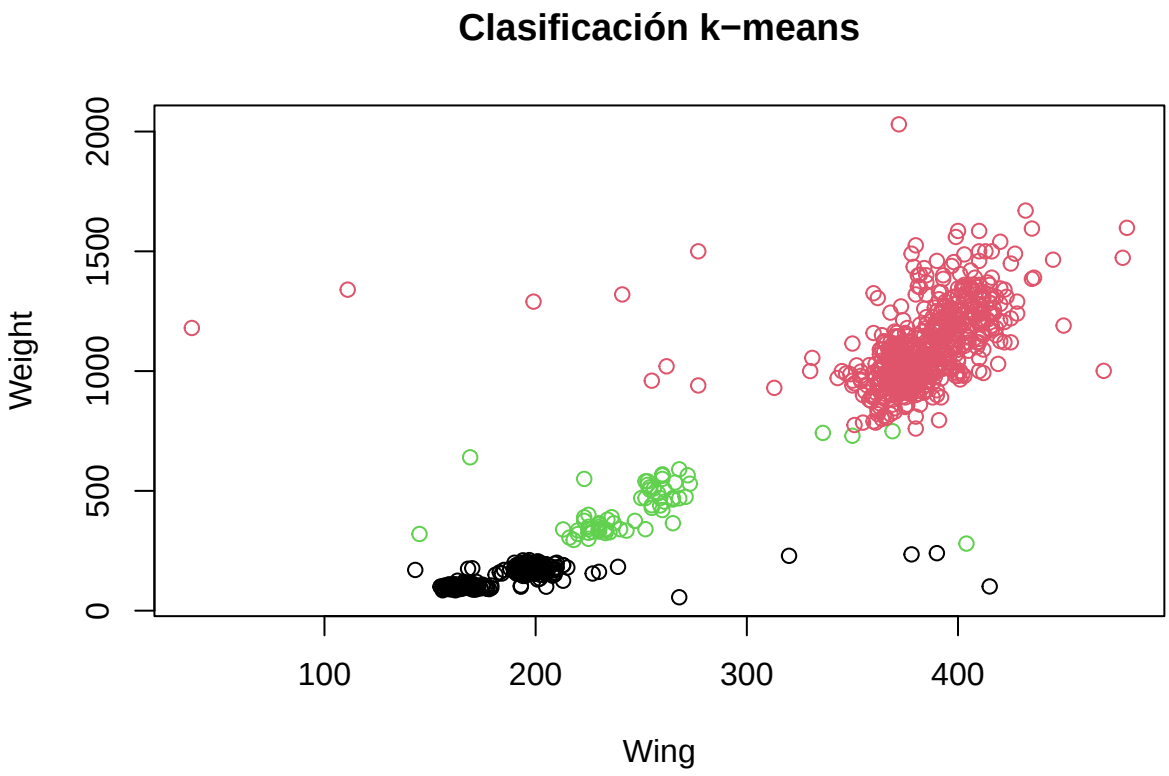

De acuerdo a los criterios ch y asw, el número para "k" podría ser 3, este resultado es el mas cercano al número de especies que ya conocemos. PAra fines de nuestra PEC 2, vamos a continuar con el valor de "k" igual a 3.

### Clasificación k-means

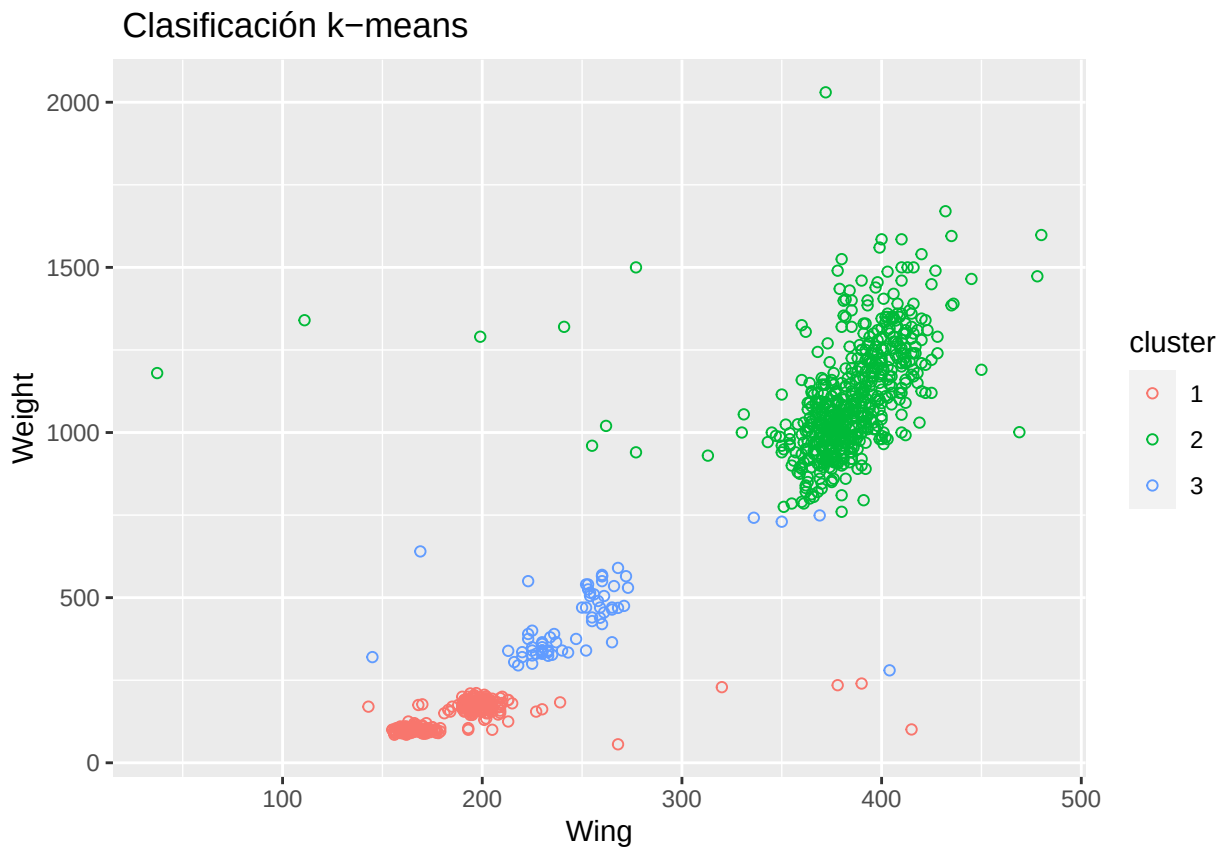
Aplicamos la función de kmeans para 3 clusters

```
hawks3clusters <- kmeans(hawks_k_means, 3)
hawks_k_means$cluster <- as.character(hawks3clusters$cluster)
```

```
#Wing and Weight
plot(hawks_k_means[c(1,2)], col=hawks3clusters$cluster, main="Clasificación k-means")
```

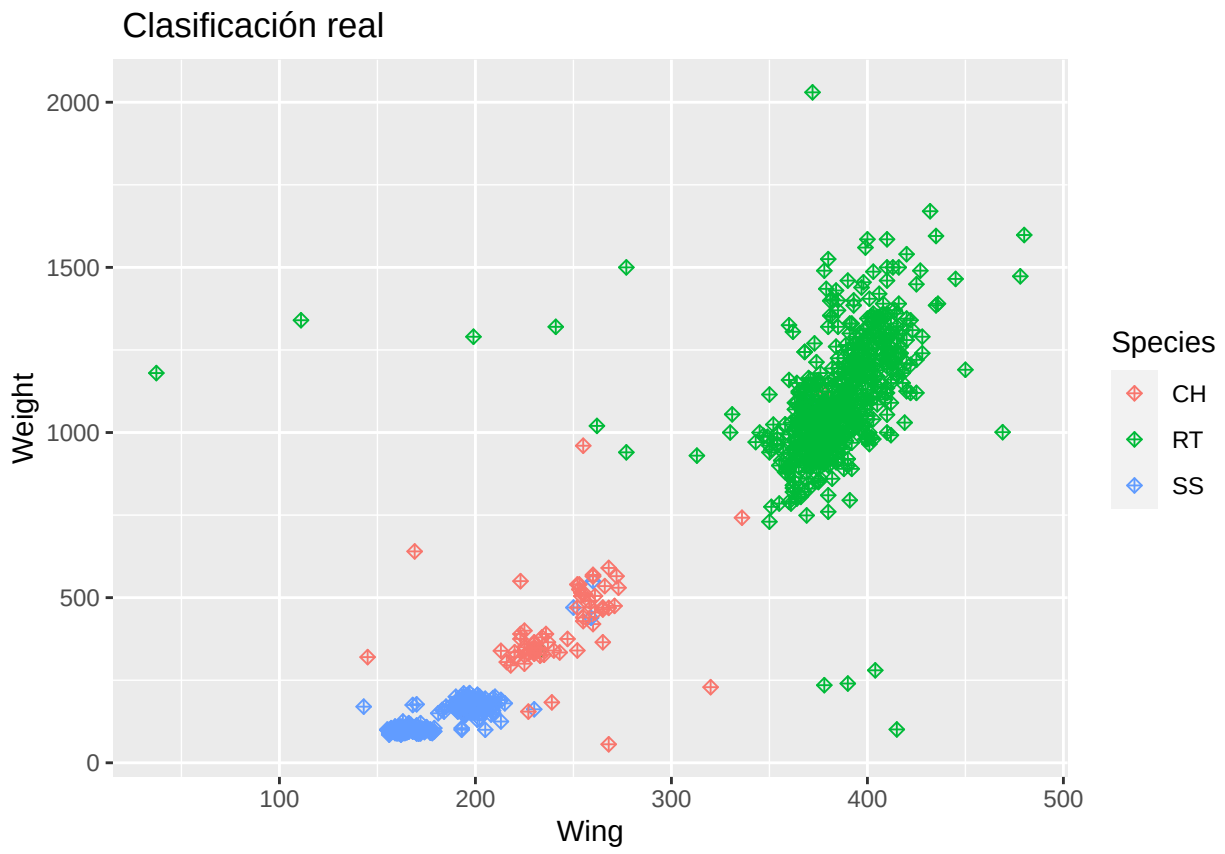



```
ggplot(hawks_k_means) + geom_point(aes(x=Wing, y=Weight, colour=cluster), shape=1) + labs(title= " Clas
```

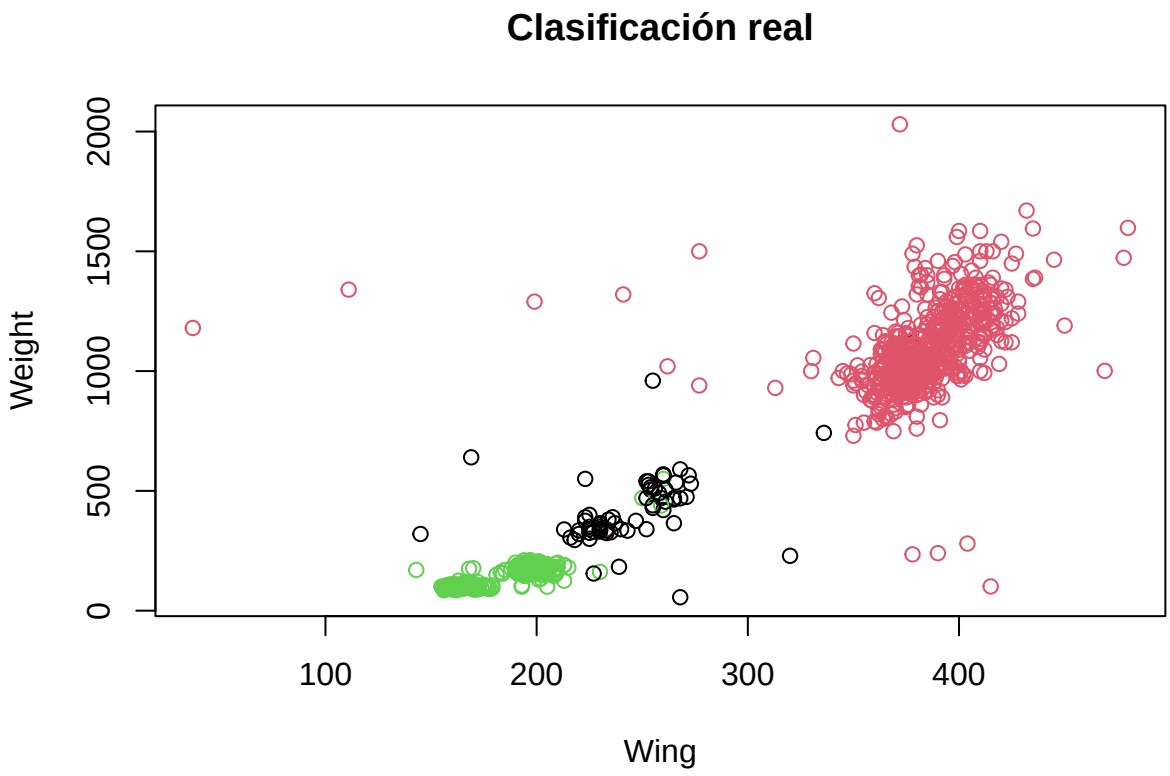



```
#Wing and Weight
ggplot(Hawks) + geom_point(aes(x=Wing, y=Weight, colour=Species), shape=9)+ labs(title= " Clasificación
```

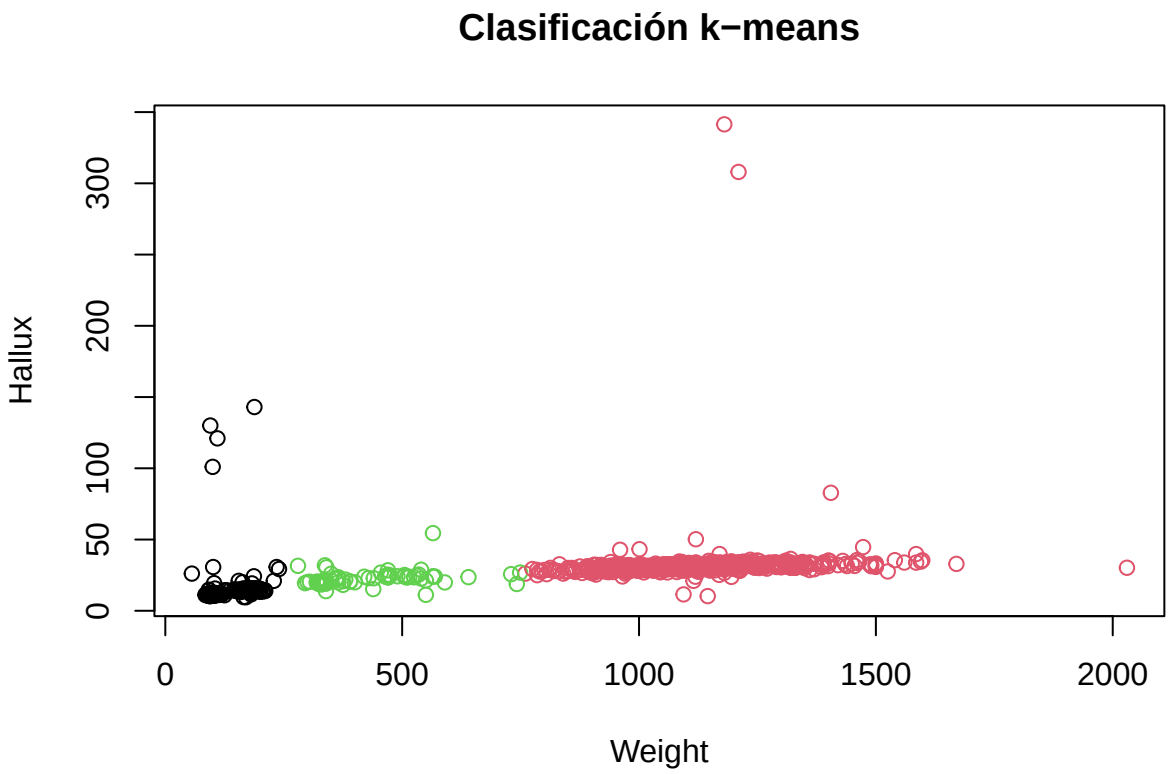
```
## Warning: Removed 11 rows containing missing values (‘geom_point()‘).
```



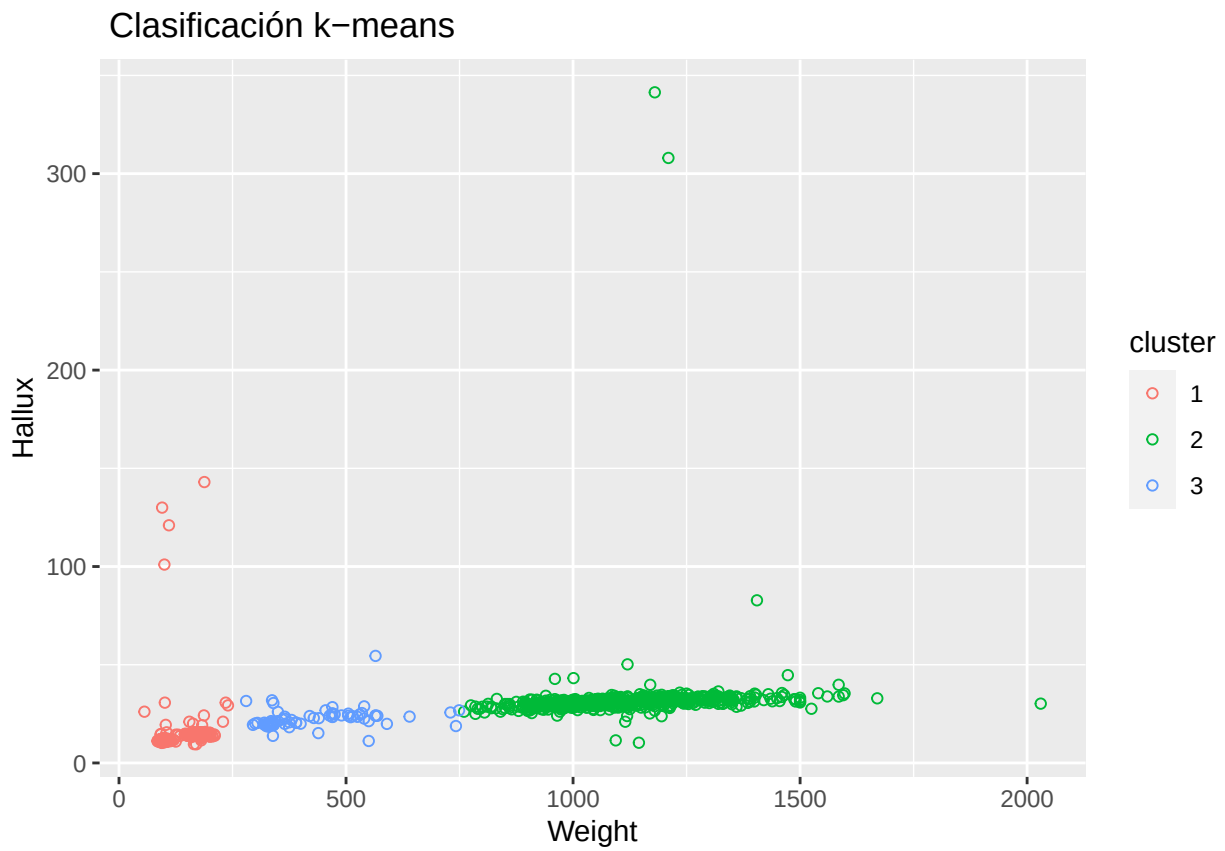

```
plot(hawks_k_original[c(4,5)], col=as.factor(hawks_k_original$Species), main="Clasificación real")
```



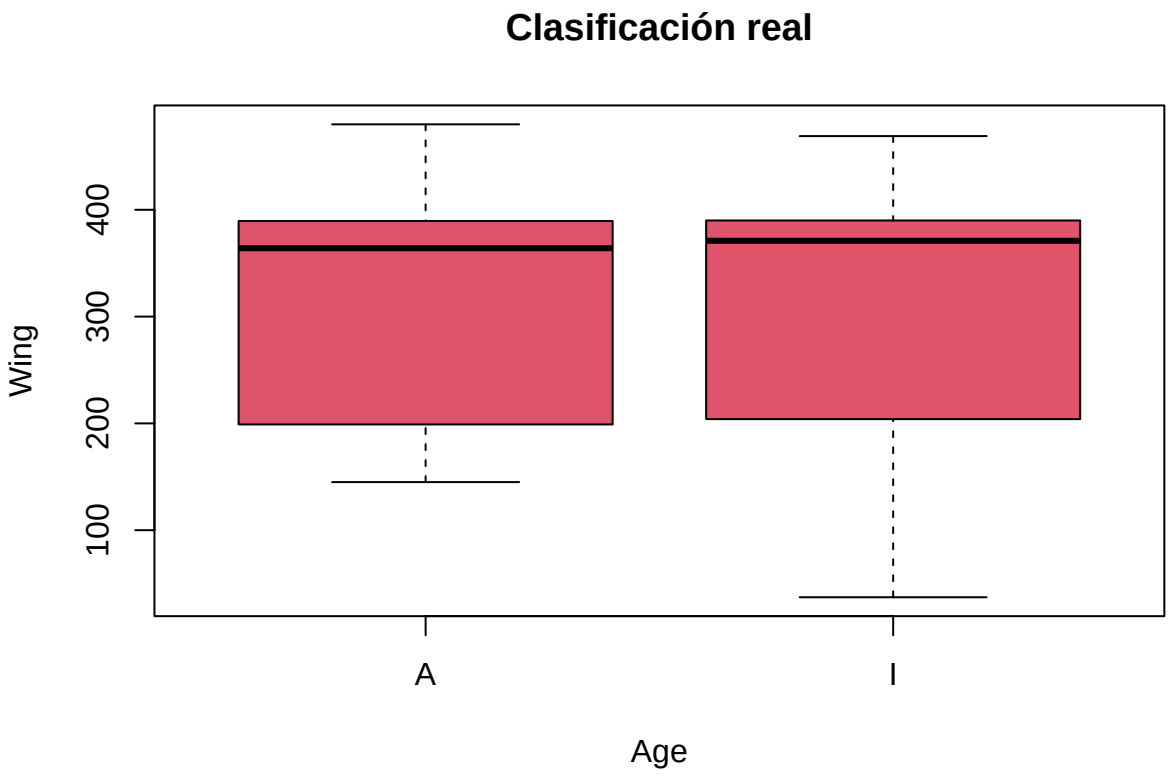

```
#Hallux and Weight
plot(hawks_k_means[c(2,4)], col=hawks3clusters$cluster, main="Clasificación k-means")
```




```
ggplot(hawks_k_means) + geom_point(aes(x=Weight, y=Hallux, colour=cluster), shape=1) + labs(title= " Cla
```

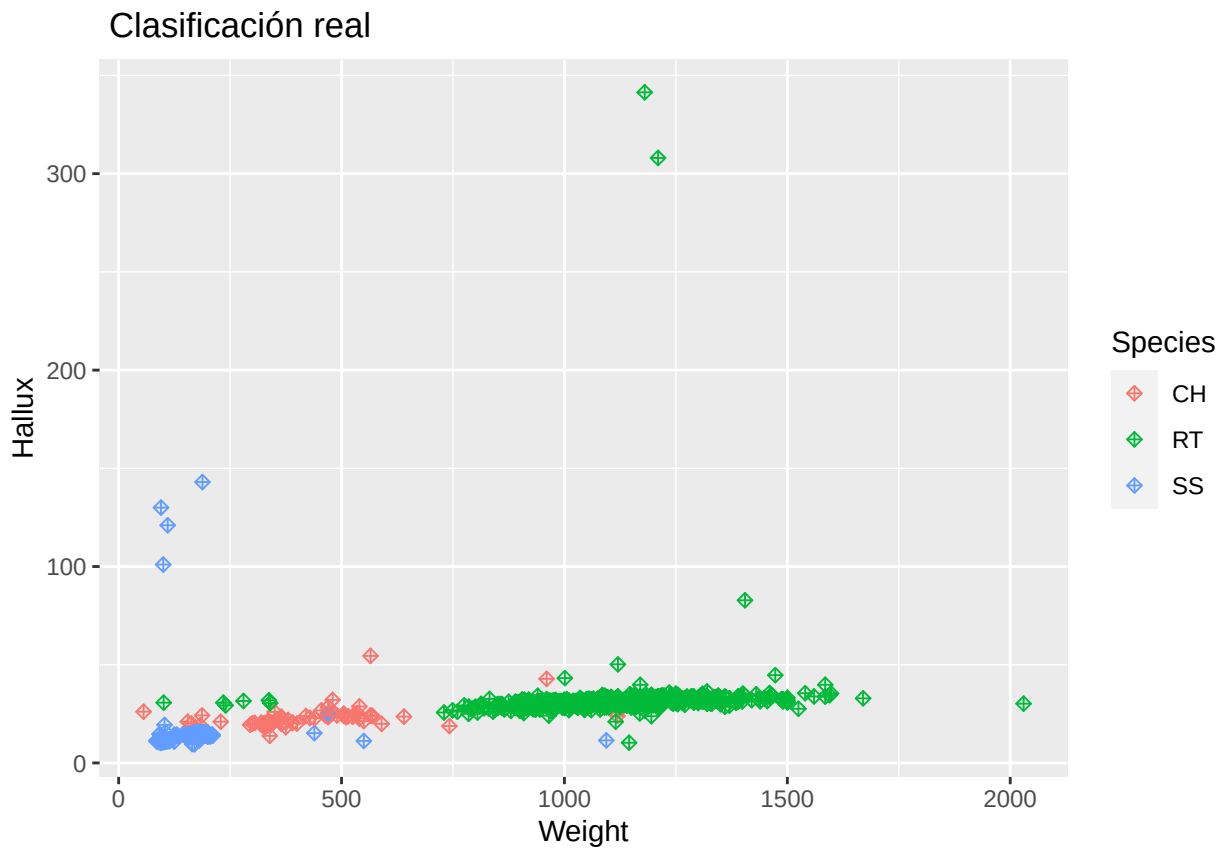



```
#Hallux and Weight
plot(hawks_k_original[c(2,4)], col=as.factor(hawks_k_original$Species), main="Clasificación real")
```

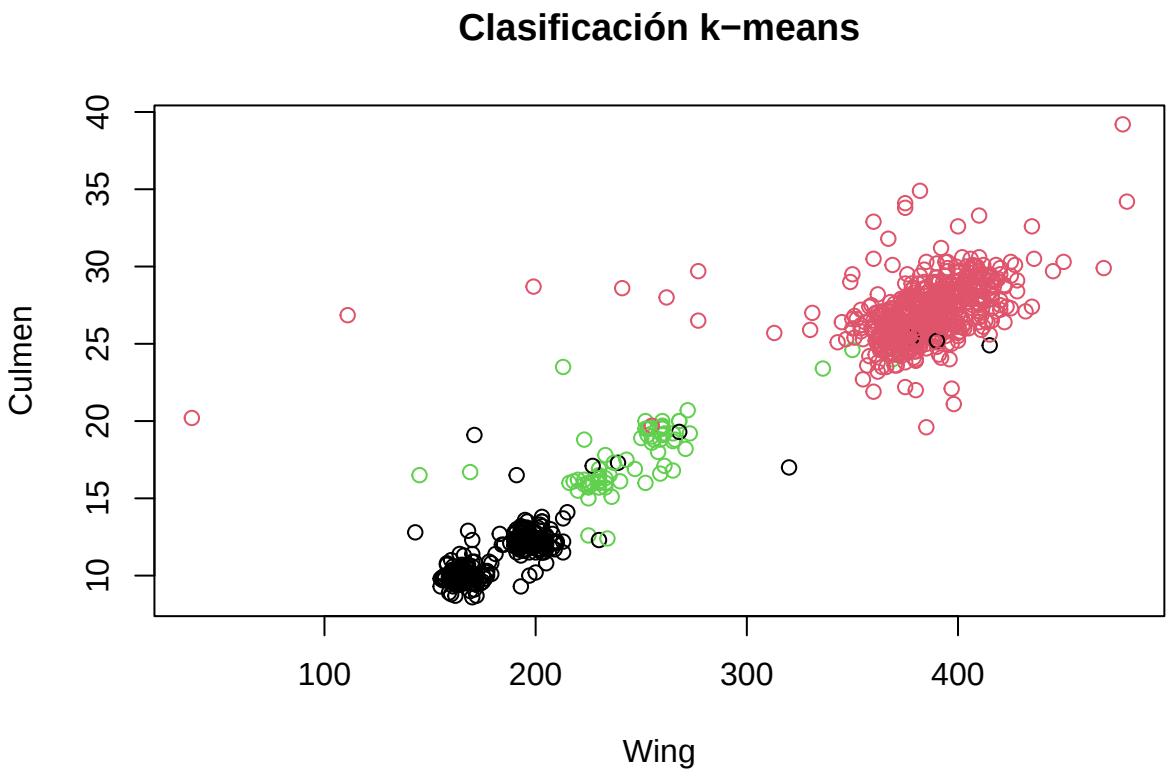



```
ggplot(Hawks) + geom_point(aes(x=Weight, y=Hallux, colour=Species), shape=9)+ labs(title= " Clasificaci
```

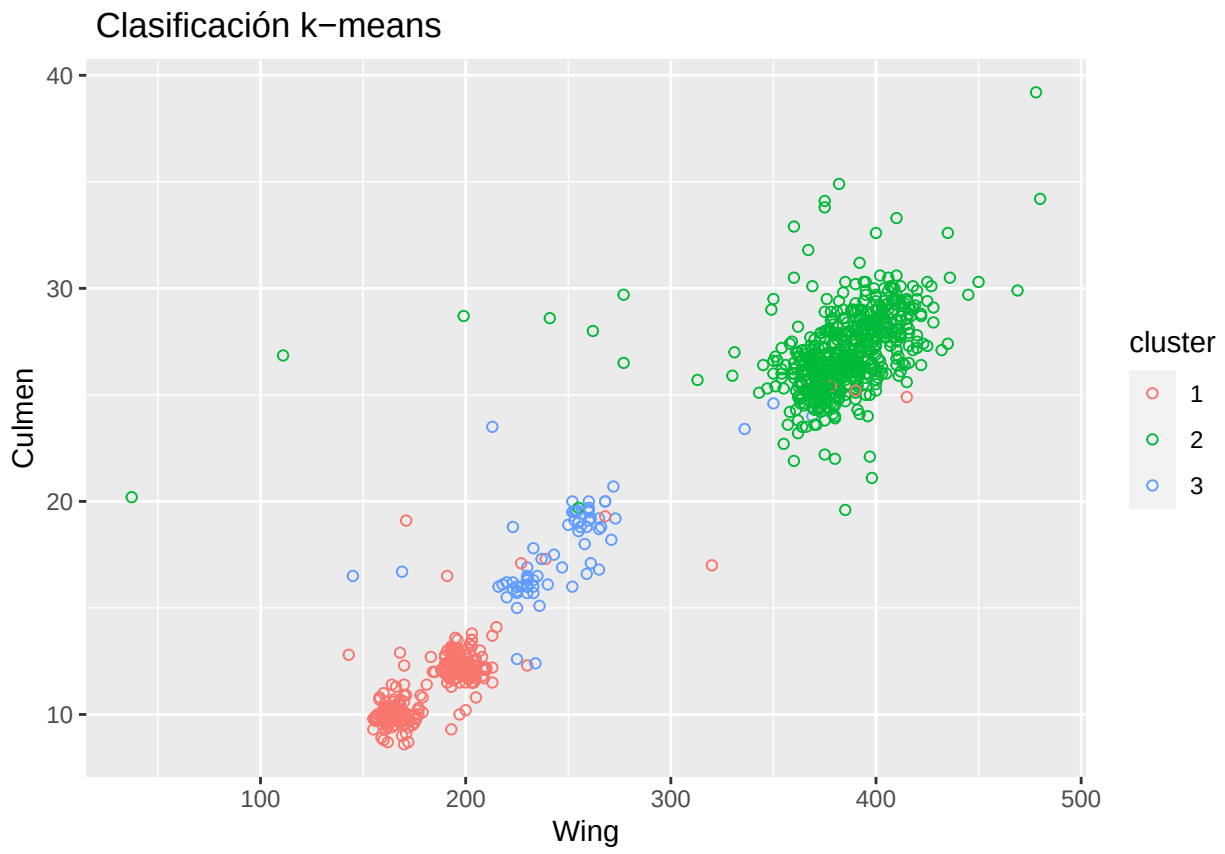
```
## Warning: Removed 14 rows containing missing values (‘geom_point()‘).
```



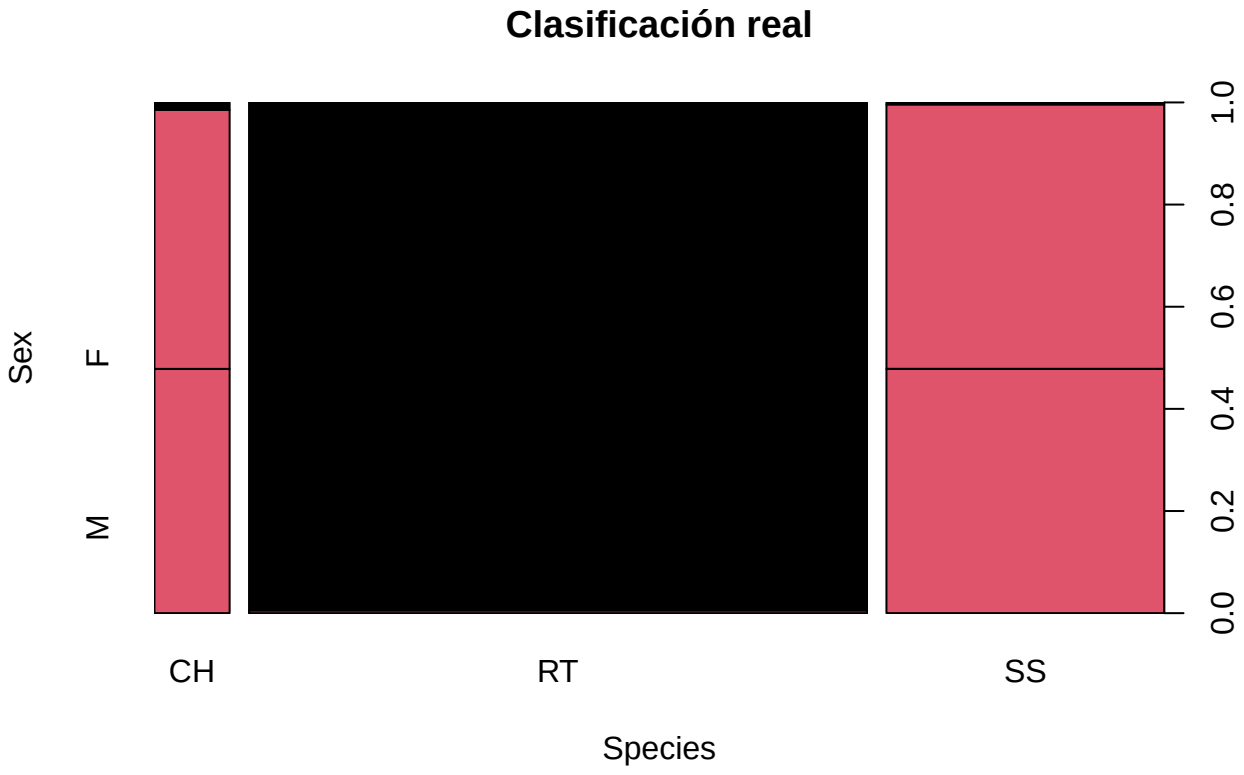

```
#Culnmen and Wing
plot(hawks_k_means[c(1,3)], col=hawks3clusters$cluster, main="Clasificación k-means")
```




```
ggplot(hawks_k_means) + geom_point(aes(x=Wing, y=Culmen, colour=cluster), shape=1) + labs(title= " Clas
```

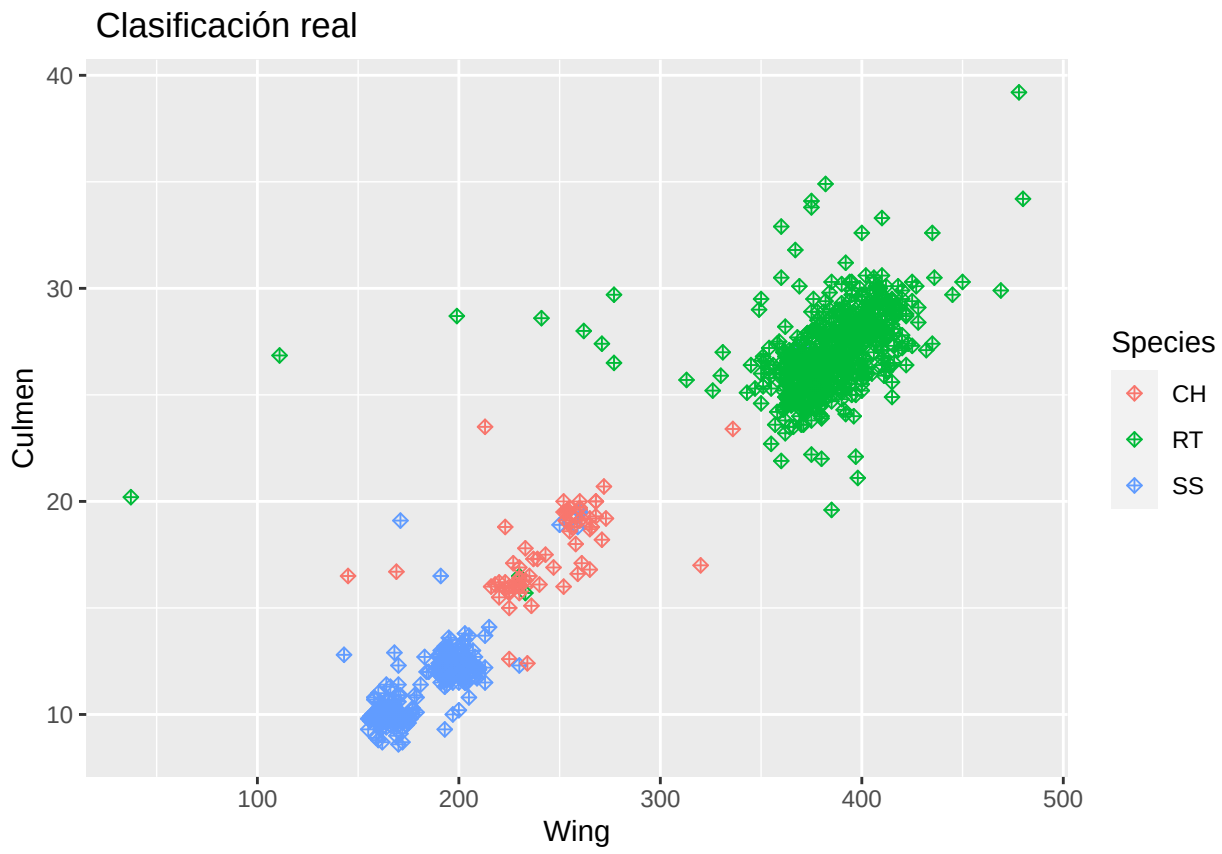



```
#Culnmen and Wing
plot(hawks_k_original[c(1,3)], col=as.factor(hawks_k_original$Species), main="Clasificación real")
```




```
ggplot(Hawks) + geom_point(aes(x=Wing, y=Culmen, colour=Species), shape=9)+ labs(title= " Clasificación
```

```
## Warning: Removed 8 rows containing missing values (‘geom_point()‘).
```


Clasificación real
Culmen
Wing
40
30
20
10
100
200
300
400
500
Species
CH
RT
SS