Práctica Final

Tatiana Dávila Egas

Vamos a utilizar el dataset de semillas que se encuentra aquí: https://archive.ics.uci.edu/ml/datasets/seeds#

Primero vamos a descargarnos el dataset con el siguiente comando:

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
library(magrittr)
```

```
Warning: package 'magrittr' was built under R version 4.0.5
```

```
Attaching package: 'magrittr'

The following object is masked from 'package:purrr':

    set_names

The following object is masked from 'package:tidyr':

    extract
```

```
df_seeds <- read.table('https://archive.ics.uci.edu/ml/machine-learning-databases/00236/seeds
```

PREGUNTA 1

¿Cuantas filas y cuantas columnas tiene el dataframe df_seeds?

Respuesta:

```
dim(df_seeds)
```

```
[1] 210   8
```

PREGUNTA 2

Vamos a convertir en factor la columna tipo. Vamos a reemplazar los números por su correspondiente etiqueta (label). La correspondencia entre el código y el tipo es:

- 1 - Kama
- 2 - Rosa
- 3 - Canadian

Convierte en factor la columna tipo, respetando las etiquetas:

Respuesta:

```
df_seeds$tipo %<>% factor(levels = c(1,2,3), labels = c('Kama','Rosa','Canadian'))
```

PREGUNTA 3

¿Cual es la media del area de cada uno de los tipos?

Respuesta

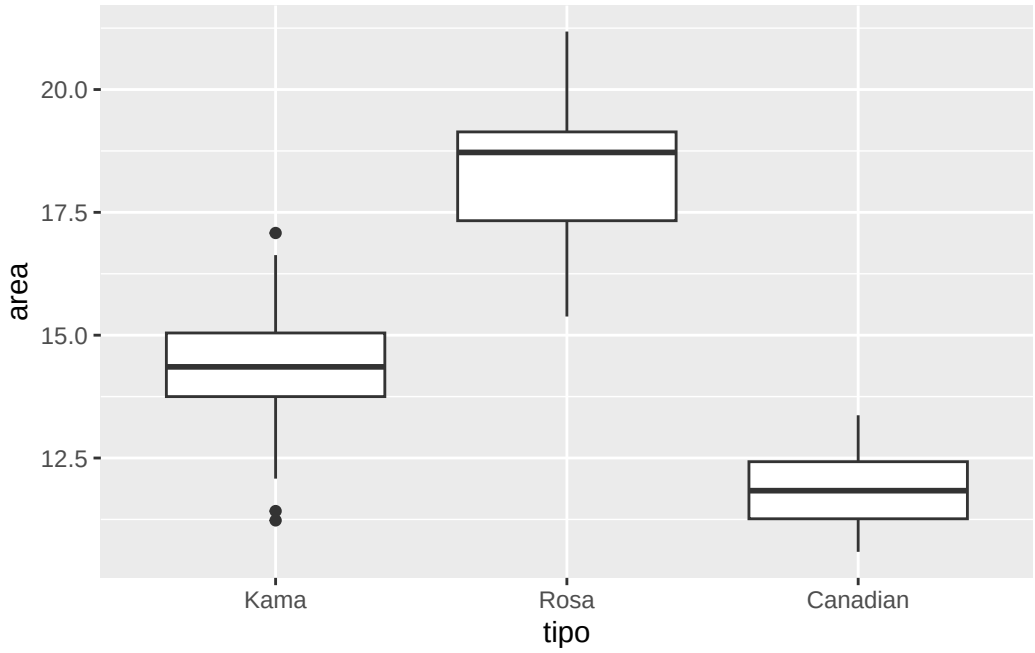
```
df_seeds %>% group_by(tipo) %>% summarise(mean(area))
```

```
# A tibble: 3 x 2
  tipo     `mean(area)`
  <fct>           <dbl>
1 Kama             14.3
2 Rosa             18.3
3 Canadian         11.9
```

PREGUNTA 4

¿Como se llama el siguiente tipo de gráfico?. ¿Qué representa la línea del centro de la caja?

```
ggplot(df_seeds, aes(x=tipo, y=area)) + geom_boxplot()
```

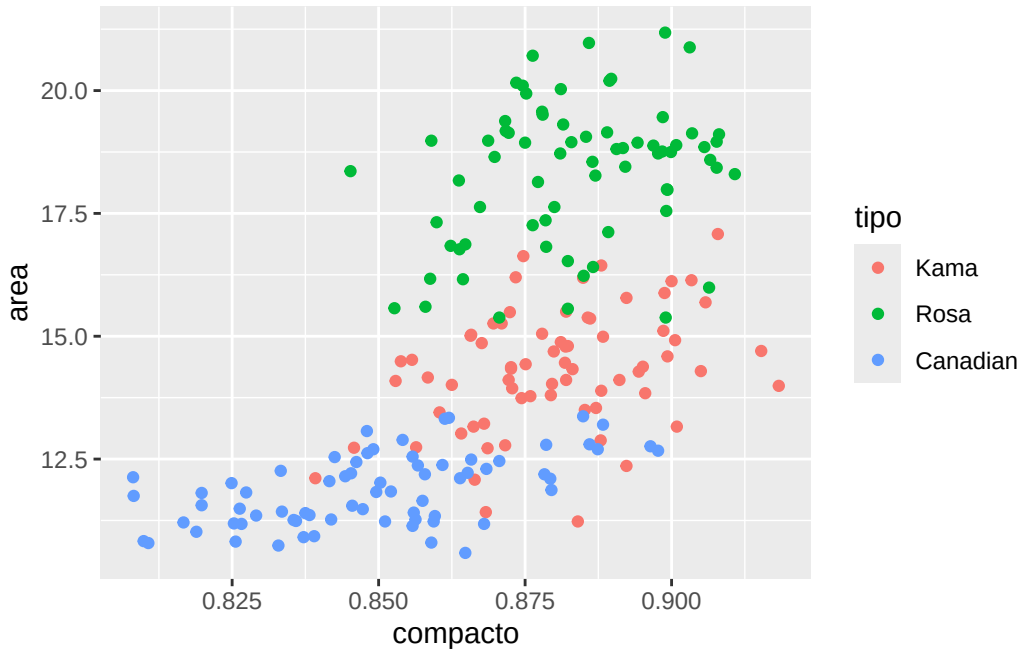


Respuesta: 'Este tipo de gráfico se llama boxplot o diagrama de cajas y bigotes, su objetivo es representar la distribución de los datos dando como referencia los cuartiles 1, 2 y 3 y los outliers, la línea del centro de la caja representa el cuartil o mediana.' #### PREGUNTA 5

¿Como pintarías un diagrama de puntos (o scatterplot) con ggplot con las siguientes características? - En el eje X la variable compacto - En el eje Y la variable area - Cada tipo de semilla debería tener un color diferente

Respuesta:

```
ggplot( aes(x=compacto, y=area, color=tipo), data= df_seeds ) + geom_point()
```



PREGUNTA 6

¿Qué hace la siguiente línea?:

```
df_seeds %>% mutate(is_kama = tipo=='Kama') -> df_seeds
```

Respuesta: 'Esta línea crea una variable llamada is_kama de tipo booleano que será TRUE si el tipo es kama y FALSE si no lo es.'

PREGUNTA 7

Vamos a dividir el conjunto de datos en test y training porque vamos a entrenar un modelo que me permita diferenciar si una semilla es de tipo Kama o no. ¿Por qué es aconsejable dividir el dataset en los grupos de train y test?

```
set.seed(123) # Este set.seed hace que a todos nos generen los mismos número aleatorios
idx <- sample(1:nrow(df_seeds), 0.7*nrow(df_seeds))
df_seeds_train <- df_seeds[idx,]
df_seeds_test <- df_seeds[-idx,]
```

Respuesta: 'Es aconsejable dividir el dataset en dos grupos para medir la calidad de nuestro modelo y evitar overfitting o underfitting, es decir puede darse el caso de que nuestro modelo

"aprenda" muy bien con nuestros datos y al realizar estimaciones con nuevos datos el error que cometa sea muy alto. Al tener dos grupos uno de train y uno de test podemos comparar qué tan bueno es el modelo con cada dataset y podemos estar seguros de que la precisión que tiene se va a mantener cuando evaluemos nuevos datos.'

PREGUNTA 8

Vamos a crear un modelo para realizar una clasificación binaria, donde le pasaremos como entrada las columnas: area, perimetro, compacto, longitud, coeficient.asimetria y longitud.ranura

¿Qué tipo de algoritmo o modelo debería usar?

Respuesta: 'En este caso dado que la variable a predecir es binaria la mejor opción es un modelo de regresión logística.'

PREGUNTA 9

Crea un modelo que me permita clasificar si una semilla es de tipo Kama o no con las siguientes columnas: area, perimetro, compacto, longitud, coeficient.asimetria, longitud.ranura

Respuesta:

```
model<-glm(data=df_seeds_train, formula=is_kama~area + perimetro + compacto + longitud + coe
```

PREGUNTA 10

Si usamos un umbral de 0 en la salida del modelo (lo que equivale a probabilidad de 0.5 cuando usamos el predict con type='response') ¿Cuales son los valores de precisión y exhaustividad?

Respuesta:

```
umbral <- 0.5
pred_test=predict(model, df_seeds_test, type="response")
M=table(real=df_seeds_test$is_kama, pred=pred_test>umbral)

paste("La precisión es:",M[2,2]/(M[1,2]+M[2,2]))
```

```
[1] "La precisión es: 0.958333333333333"
```

```
paste("La exhaustividad es:",M[2,2]/(M[2,1]+M[2,2]))
```

```
[1] "La exhaustividad es: 1"
```

PREGUNTA 11

¿Qué están haciendo las siguientes líneas?

```
set.seed(123)
cl<-df_seeds %>%  select(area,perimetro,compacto,longitud,anchura,coeficient.asimetria,longi
table(real=df_seeds$tipo,cluster=cl$cluster)
```

```
          cluster
real        1  2  3
  Kama      1 60  9
  Rosa     60 10  0
  Canadian  0  2 68
```

Respuesta: 'Estas líneas están clusterizando los datos en 3 grupos, esta clusterización se basa en las variables: área, perímetro, compacto, longitud,anchura, coeficient.asimetria y longitud.ranura. Si observamos el boxplot de la pregunta 4 ya se puede deducir que la variable área es determinante para describir los diferentes tipos, además en base al área también se define los 3 tipos en 3 grupos diferentes. Al observar el gráfico de la pregunta 5 se refuerza la idea de que el area es diferente para cada tipo, pero también observamos que en algunos casos un tipo Kama se mezcla en el grupo del tipo Rosa o del tipo Canadian, por ello también podemos entender que los 3 clústers que se crean contengan más de un tipo, como es el caso del clúster dos que tiene sujetos de los tres tipos.'

OBSERVACIÓN

A lo largo del ejercicio se intuye la alta correlación que existe entre las variables que entran en el modelo y que se usan para clusterizar. Una buena opción sería observar esta correlación y descartar las variables que más correladas están, en este caso la variable área parece recoger la información de las variables perímetro, longitud y anchura, por lo que podríamos descartar estar tres y quedanos solo con área.

```
df_seeds %>%  select(area,perimetro,compacto,longitud,anchura,coeficient.asimetria,longitud.r
```

```
                     area perimetro compacto longitud anchura
area                 1.00      0.99     0.61     0.95    0.97
perimetro            0.99      1.00     0.53     0.97    0.94
compacto             0.61      0.53     1.00     0.37    0.76
longitud             0.95      0.97     0.37     1.00    0.86
anchura              0.97      0.94     0.76     0.86    1.00
coeficient.asimetria -0.23     -0.22    -0.33    -0.17   -0.26
```

```
longitud.ranura       0.86      0.89     0.23     0.93    0.75
                    coeficient.asimetria longitud.ranura
area                               -0.23            0.86
perimetro                          -0.22            0.89
compacto                           -0.33            0.23
longitud                           -0.17            0.93
anchura                            -0.26            0.75
coeficient.asimetria                1.00           -0.01
longitud.ranura                    -0.01            1.00
```

```
df_seeds %>%  select(area,perimetro,compacto,longitud,anchura,coeficient.asimetria,longitud.r
```