Práctica

CUNEF - Máster en Data Science

9/1/23

Ejercicio 1 (0,5 puntos)

Crea un *data frame* o *tibble* a partir de los datos del fichero `penguins`. Contiene información sobre ejemplares de pingüinos:

- La especie a la que pertenece
- La isla en la que se encuentra
- Dimensiones del pico (*bill*)
- Longitud del ala (*flipper*)
- La masa
- El sexo
- El año en que se registró

Ejercicio 2 (1 punto)

Responde a las siguientes preguntas:

- ¿Cuántas especies distintas hay?
- ¿De cuántas islas distintas hay datos?
- ¿Están todas las especies en todas las islas?

Ejercicio 3 (0,5 puntos)

En el ejercicio 2 has visto que en la isla Torgersen solo hay una especie. ¿Cuál es?

Ejercicio 4 (1 punto)

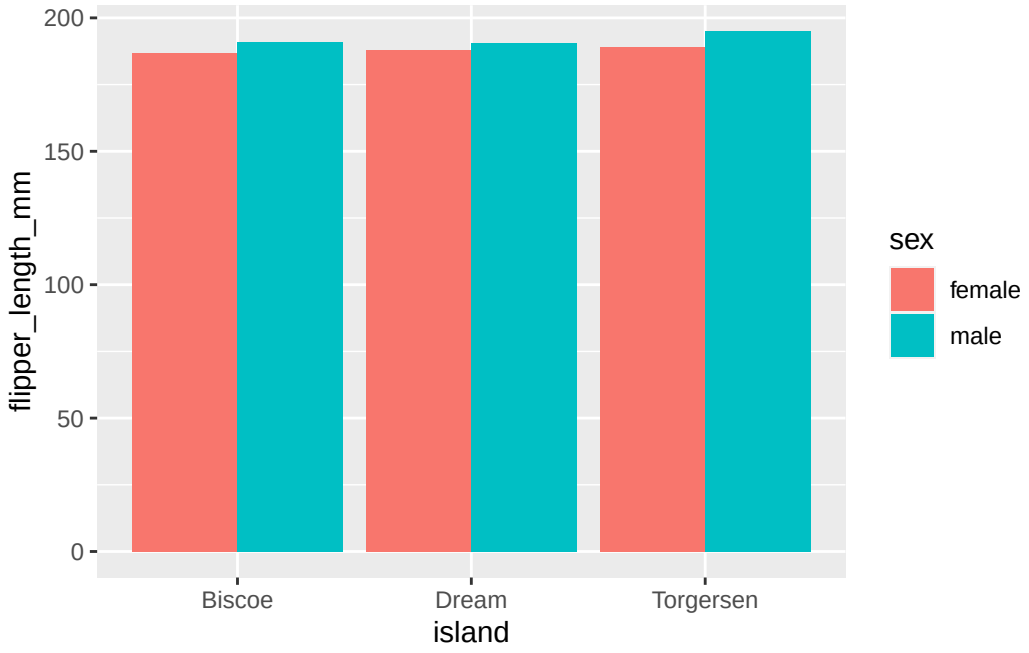
Teniendo en cuenta que cada fila es un pingüino, calcula cuántos pingüinos de cada especie hay.

Ejercicio 5 (1 punto)

Calcula la media de la longitud y la profundidad del pico en función de cada especie. *Pista.* Ninguna de las medias es `NA`.

Ejercicio 6 (1 punto)

Calcula la mediana de la longitud del ala para la especie *Adelie* en cada isla y en función del sexo. Quita los casos en los que `sex` es `NA`. Replica el siguiente gráfico, que tiene todos esos datos coloreando cada columna en función del sexo. Necesitarás jugar con la `position` de las columnas. **Explica qué conclusiones sacas a la vista del gráfico.** *Pista.* De nuevo, ninguno de estos datos es `NA`.



Ejercicio 7 (2 puntos)

Más adelante vas a hacer un modelo con estos datos, concretamente, con las columnas `bill_length_mm`, `bill_depth_mm`, `flipper_length_mm`, `body_mass_g`. Antes de eso, tienes que reemplazar los `NA`. Para ello, calcula las medias de las cuatro columnas y utilízalas para reemplazar los `NA`.

Puedes hacerlo como quieras, pero **si lo haces con un bucle, tendrás la máxima puntuación en este ejercicio. Si no, no :)**

Pista. El bucle yo lo he planteado tratanto a las columnas como vectores (o sea, no lo planteo con dplyr).

Ejercicio 8 (1 punto)

Ahora haz un modelo de segmentación con esas variables. Para ello, usa la función `kmeans()`, disponible en R. La función recibirá dos argumentos:

- un `data.frame` o `tibble` con datos
- el número de grupos que hay que utilizar (ahora te digo cuántos).

El objetivo de esto es que, a partir de las columnas del ejercicio anterior, un algoritmo asigne cada pingüino a un grupo (puede coincidir con la especie o no).

Crea un `data.frame` o `tibble` nuevo con esas columnas mencionadas antes. Y luego llama a la función `kmeans()` con ese data frame y como argumento `centers=` usa `3` (porque hay 3 especies, aunque luego matizamos esto). Guarda el resultado en un objeto.

Ejercicio 9 (2 puntos)

El ejercicio anterior estaba pensado solamente para que te familiarizaras con la sintaxis de un modelo.

Ahora vas a validar si lo de agrupar los pingüinos en 3 grupos es algo razonable o es mejor otro número. Para ello, haz un bucle en el que hagas un modelo para 2 grupos, para 3, para 4 y así hasta 10. Te cuento cómo.

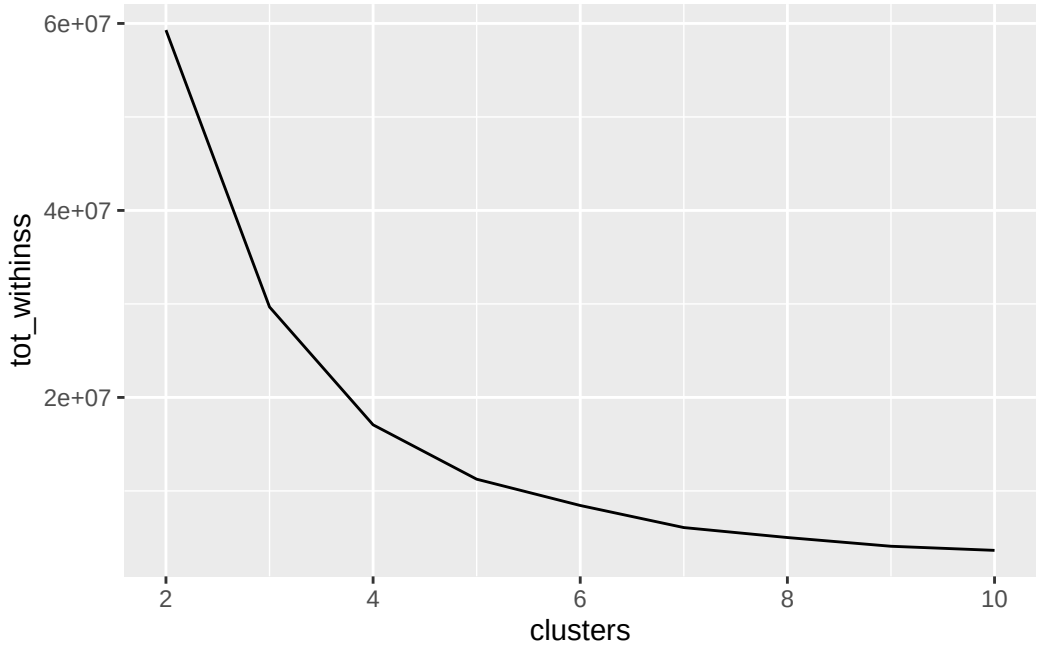
Del modelo, te interesa solo una cosa: el número `tot.withinss`. Para obtenerlo, ten en cuenta que si estás guardando el resultado de `kmeans()` en un objeto llamado `fit` (por ejemplo), puedes acceder a ese número haciendo `fit$tot.withinss`. Es una medida que indica cómo de juntos están los individuos al centro del grupo que se les ha asignado (es decir, si los grupos son homogéneos o no). Luego te explico cómo interpretarlo.

Quiero que construyas un `data.frame` o `tibble` con dos columnas: la primera es el número de grupos (`2:10`) y la segunda es precisamente esa medida que has calculado con el bucle para cada número de grupos.

Si haces el bucle con `sapply()` podrás aspirar a la máxima puntuación en este ejercicio. Además, te devolverá ya el vector que será la segunda columna de este `data.frame` o `tibble`.

Ejercicio 10 (1 punto)

En el ejercicio anterior habrás creado un `data.frame` de 9 filas. Haz un gráfico de líneas parecido al siguiente con esa información (no te saldrán los mismos números, pero sí deberías observar un comportamiento parecido y una magnitud del eje y similar).



Ejercicio 11 (1 punto)

En el gráfico anterior hay que fijarse dónde hay una diferencia significativa en la variación de la métrica.

El ejercicio anterior sugiere que tomar entre 3 ó 4 grupos parece razonable (porque pasar de 2 a 3 clusters mejora mucho, pasar de 3 a 4 mejora algo, pero pasar de 4 a 5 ya no mejora mucho).

Como en el ejercicio 8 ya has ajustado un modelo con 3 clusters, vas a reutilizarlo. De nuevo, si llamaste a ese modelo `fit`, puedes acceder al vector `fit$cluster` para obtener las etiquetas que el modelo ha dado a cada pingüino.

Añade ese vector al `data.frame` original, al que tiene las etiquetas reales e intenta replicar el siguiente gráfico. Es un gráfico de dispersión entre `bill_length_mm` y `flipper_length_mm` y los puntos están coloreados en función de la especie y, además, tienen una forma distinta en función de la etiqueta (para esto, necesitarás convertir esta etiqueta a factor con la función `as.factor()`). Aparte, está separado por isla.

Comenta qué puedes observar a simple vista.