Laboratorio de Datos

Primer Cuatrimestre 2024

Práctica N° 5: Modelo lineal multivariado. Entrenamiento y testeo.

Para estos ejercicios, usar el dataset `penguins`.

1. Crear un subconjunto de datos que contenga sólo pingüinos de la isla Biscoe y que tengan un pico de 48 mm de largo o más.
2. Crear otro dataset con la información de pingüinos Adelie machos que no hayan sido vistos en el año 2008.
3. Del dataset `penguins` quedarse con todas las variables excepto `year`, `sex` y `body_mass_g`.
4. Crear un subconjunto de los datos de penguins sólo con las obsevaciones de pingüinos machos con aletas (flipper) de más de 200 mm de largo y quedarse con todas las columnas que terminan con "mm". (Ayuda: pueden utilizar el método `endswith()` aplicado al string.)
5. Empezando con `penguins`, hacer un pipe (mediante el método `pipe` de DataFrames) que:
   - **(a)** se quede sólo con las observaciones de la isla Dream.
   - **(b)** se quede con las variables species y todas las que empiecen con bill.
6. Convertir todas las variables que empiezan con bill a mayúsculas. (Ayuda: `rename()` y `upper()`.)
7. Empezando con `penguins` hacer las siguientes operaciones utilizando `transform()`:
   - **(a)** Crear una nueva variable que tenga el peso en Kg.
   - **(b)** Convertir la variable island a minúscula.
8. Empezando con `penguins` crear una tabla resumen que contenga para el largo mínimo y máximo de las aletas de los pingüinos Adelie, agrupados por isla.
9. Empezando con `penguins`, agrupar los datos por especie y año, luego crear una tabla de resumen que contenga el ancho del pico (llamarla `bill_depth_mean`) y el largo del pico (llamarla `bill_length_mean`) para cada grupo
10. Empezando con `penguins`, hacer una secuencia de operaciones `pipe` que:
    - **(a)** Agregue una nueva columna llamada `bill_ratio` que sea el cociente entre el largo y el ancho del pico.
    - **(b)** Quedarse sólo con las columnas species y `bill_ratio`.
    - **(c)** Agrupar los datos por especie.
    - **(d)** Crear una tabla de resumen que contenga el promedio de la variable `bill_ratio` por especie y que el nombre de la columna en la tabla sea `bill_ratio_mean`).
11. Usar el método `rename()` de DataFrames para cambiarle el nombre a la variable `body_mass_g` y llamarla `masa_corporal_g`.

12. Calcular la mediana de la masa corporal de los pingüinos de cada especie usando `group_by()` y `summarise()`.

13. Empezando con `penguins`, escribir una secuencia de operaciones `pipe` que:

(a) Excluya a los pingüinos observados en la isla Biscoe.

(b) Sólo se quede con las variables que están entre species y `body_mass_g` inclusive.

(c) Renombrar la variable species a `especie_pinguino`.

(d) Agrupar los datos por la variable `especie_pinguino`.

(e) Encontrar el valor medio de las variables que contienen el string "length", separando por la especie del pingüino, y llamando a las columnas como las originales pero agregando "_mean" al final.

14. Empezando con `penguins`, contar cuántas observaciones hay por especie, isla y año.

15. Empezando con `penguins`, quedarse sólo con los pingüinos de las especies Adelie y Gentoo. Luego contar cuántos hay por cada especie y sexo.

16. Agregar una nueva columna a la base de datos llamada `peso_bin` que contenga:

- "chico" si la masa corporal es menos que 4000 gramos.
- "grande" si la masa corporal es mayor que 4000 gramos.

17. Empezando con `penguins`,

(a) Quedarse sólo con las observaciones correspondientes a pingüinos Chinstrap.

(b) Luego, quedarse sólo con las variables `flipper_length_mm` y `body_mass_g`.

(c) Agregar una nueva columna llamada `fm_ratio` que contenga el cociente entre el largo de la aleta y el peso del pingüino.

(d) Luego quedarse solo con las observaciones que no tienen NaN en ninguna columna (ayuda: `drop_na()`)

(e) Agregar otra columna llamada `ratio_bin` que contenga la palabra "alto" si `fm_ratio` es mayor o igual que 0.05 y "bajo" si el cociente es menor que 0.05.