Guía 4: Transformación de datos

Laboratorio de datos 2023 (comisión: G. Solovey)

1. Transformación de datos

Para estos ejercicios, usar el dataset `penguins` del paquete `palmerpenguins`.

1.1 Usar `filter()` para crear un subconjunto de datos que contenga sólo pinguinos de la isla Biscoe y que tengan un pico de 48 mm de largo o más.

1.2 Crear otro dataset con la información de pinguinos Adelie machos que no hayan sido vistos en el año 2008.

1.3 Del dataset `penguins` quedarse con todas las variables excepto `year`, `sex` y `body_mass_g`.

1.4 Crear un subconjunto de los datos de `penguins` sólo con las obsevaciones de pinguinos machos con aletas (*flipper*) de más de 200 mm de largo y quedarse con todas las columnas que terminan con "mm". (*Ayuda: explorar cómo se usa la función `ends_with()` y sus parientes acá*).

1.5 Empezando con `penguins`, hacer un `pipe` (`%>%`) que:

- se quede sólo con las observaciones de la isla Dream.
- se quede con las variables `species` y todas las que empiece con `bill`.

1.6 Convertir todas las variables que empiezan con `bill` a mayúsculas. (*Ayuda: `?rename_with` y `toupper()`*)

1.7 Empezando con `penguins` hacer lo siguiente con un único llamado a la función `mutate()`:

- Convertir la variable `species` a `character`.
- Crear una nueva variable que tenga el peso en Kg.
- Convertir la variable `island` a minúscula.

1.8 Empezando con `penguins` crear una tabla resumen que contenga para el largo mínimo y máximo de las aletas de los pinguinos Adelie, agrupados por isla.

1.9 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

1.10 Empezando con `penguins`, hacer una secuencia de operaciones `%>%` que:

- Agregue una nueva columna llamada `bill_ratio` que sea el cociente entre el largo y el ancho del pico.
- Quedarse sólo con las columnas `species` y `bill_ratio`.
- Agrupar los datos por especie.
- 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`).

1.11 Usar `rename()` para cambiarle el nombre a la variable `body_mass_g` y llamarla `masa_corporal_g`.

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

1.13 Empezando con `penguins`, escribir una secuencia de operaciones `%>%` que:

- Excluya a los pinguinos observados en la isla Biscoe.
- Sólo se quede con las variables que están entre `species` y `body_mass_g` inclusive.
- Renombrar la variable `species` a `especie_pinguino`.
- Agrupar los datos por la variable `especie_pinguino`.
- Encontrar el valor medio de las variables que contienen el string "length", separando por la especie del pinguino, y llamando a las columnas como las originales pero agregando "_mean" al final.

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

1.15 Empezando con `penguins`, quedarse sólo con los pinguinos Adelie y gentoo penguins. Luego contar cuántos hay por cada especie y sexo.

1.16 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.

1.17 Empezando con `penguins` quedarse sólo con las observaciones correspondientes a pinguinos chinstrap. Luego, quedarse sólo con las variables `flipper_length_mm` y `body_mass_g`. Agregar una nueva columna llamada `fm_ratio` que contenga el cociente entre el largo de la aleta y el peso del pinguino. Luego quedarse solo con las observaciones que no tienen `NA` en ninguna columna (*ayuda:* `?drop_na()`) y 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).

2. Exploración de datos

Para resolver estos ejercicios usando lo que han aprendido de transformación, exploración y visualización de datos usando los paquetes de `tidyverse`.

2.1 ¿Te parece que los pinguinos macho tienen más masa corporal que las hembras? Poner a prueba tu intiuición con visualizaciones y estadística descriptiva.

2.2 ¿Te parece que pinguinos con pico más largo (`bill_length_mm`) tienen a su vez el pico más ancho (`bill_depth_mm`)? Poner a prueba esta intuición con visualizaciones y estadística descriptiva.

2.3 En el 2.2 ¿da igual si consideran cada especie del pinguino por separado? ¿Qué tiene esto que ver con la paradoja de Simpson?

2.4 Repetir 2.1 pero tomando un subconjunto aleatorio de N pinguinos (explorar diferentes números para N) (*Ayuda: buscar* `?sample_n()`).

Referencias

- Recomendamos esta clase de la Dra. Lucía Babino del Instituto de Cálculo.
- tutorial de dplyr.
- Cap. 3 de Data Science. A First Introduction. Tiffany Timbers, Trevor Campbell, and Melissa Lee.
- Cap. 4 de R for Data Science (2e). Hadley Wickham, Mine Çetinkaya-Rundel, and Garrett Grolemund.