Code

Penguins

Castro, A.

2023-10-13

Inicio

Píldoras_R. Material de formación

En esta práctica analizamos el dataset de PalmerPenguins

[mi_blog] https://agustincastro.es

Cargamos la librería de **palmerpenguins** y vemos que contiene un dataset que se llama **penguins**. Este es el que vamos a utilizar en esta práctica.

Hide

```
# usar para ver los datasets que contiene la librería
data(package = "palmerpenguins")
```

Hide

```
library(tidyverse) # librería para manipulación de datos
library(DT) # tablas paginadas
library(palmerpenguins)
library(ggplot2)
library(hrbrthemes)

data("penguins")
```

Exploración general del dataset

Este es el dataset de **penguins**. Aquí puedes ver qué información contiene. He utilizado la librería **DT** y la función **datatable** para que la tabla sea paginada. Esto quiere decir que si hay muchas observaciones, no se mostrarán todas a la vez, sino que se mostrarán por páginas. En este caso, se muestran 10 observaciones por página.

Hide

```
datatable(penguins, options = list(pageLength = 10))
```

Show 10 entries

Search:

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
1	Adelie	Torgersen	39.1	18.7	181	3750	male	2007
2	Adelie	Torgersen	39.5	17.4	186	3800	female	2007
3	Adelie	Torgersen	40.3	18	195	3250	female	2007
4	Adelie	Torgersen						2007
5	Adelie	Torgersen	36.7	19.3	193	3450	female	2007
6	Adelie	Torgersen	39.3	20.6	190	3650	male	2007
7	Adelie	Torgersen	38.9	17.8	181	3625	female	2007
8	Adelie	Torgersen	39.2	19.6	195	4675	male	2007
9	Adelie	Torgersen	34.1	18.1	193	3475		2007
10	Adelie	Torgersen	42	20.2	190	4250		2007

Showing 1 to 10 of 344 entries

Para empezar nos puede interesar saber qué variables contiene el dataset. Con **colnames** podemos ver el nombre de estas. Con **length** podemos contar cuántas son.

Hide

```
length(colnames(penguins))
```

```
## [1] 8
```

Hide

```
colnames(penguins)
```

```
## [1] "species"           "island"            "bill_length_mm"
## [4] "bill_depth_mm"     "flipper_length_mm" "body_mass_g"
## [7] "sex"               "year"
```

Si queremos información sobre las variables y las observaciones podemos utilizar **str**. Interpretamos en resultado. Como vemos, hay 344 observaciones y los datos están en formato de "tibble". Hay tres variables cualitativas: La variable **[species]** es un factor que cuenta con 3 niveles (hay datos para tres especies de pinguinos). Exactamente ocurre lo mismo para **[island]** (hay tres islas). Por último está **[sex]**, que indica el sexo de los pinguinos y es un factor con dos niveles (hembra o macho). El resto de variables son cuantitativas, siendo enteras (sin decimales = int) **[flipper_length_mm]**, **[body_mass_g]** y **[year]** y, numéricas con decimales **[bill_length_mm]** y **[bill_depth_mm]**.

Hide

```
str(penguins)
```

```
## tibble [344 × 8] (S3: tbl_df/tbl/data.frame)
##  $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ bill_length_mm   : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
##  $ bill_depth_mm    : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
##  $ flipper_length_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
##  $ body_mass_g      : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...
##  $ sex              : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ...
##  $ year             : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
```

Como muchas funciones trabajan sobre data frames, creamos uno **df**.

Hide

```
df <- data.frame(penguins)
```

Podemos acceder rápidamente a un resumen de estadísticos descriptivos utilizando **sumary**.

Hide

```
summary(df)
```

```
##       species          island    bill_length_mm  bill_depth_mm
##  Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10
##  Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60
##  Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30
##                                  Mean   :43.92   Mean   :17.15
##                                  3rd Qu.:48.50   3rd Qu.:18.70
##                                  Max.   :59.60   Max.   :21.50
##                                  NA's   :2       NA's   :2
##  flipper_length_mm  body_mass_g       sex           year
##  Min.   :172.0     Min.   :2700   female:165   Min.   :2007
##  1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007
##  Median :197.0     Median :4050   NA's  : 11   Median :2008
##  Mean   :200.9     Mean   :4202                Mean   :2008
##  3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009
```

```
##  Max.   :231.0     Max.   :6300              Max.   :2009
##  NA's   :2         NA's   :2
```

Hacemos algunos cambios

Vamos a realizar una traducción y cambios en los nombres de las variables del dataset utilizando la función **rename**. Además, creamos una nueva variable **ratio** con la función **mutate**, que va a ser la relación entre la masa corporal y la longitud de la aleta. Con **head(penguins, 5)** podemos ver las 5 primeras observaciones del nuevo dataset, ahora llamado **df_col_rename**.

Hide

```
df_col_rename <- df %>%
  rename(especies = species,
         isla = island,
         pico_longitud_mm = bill_length_mm,
         pico_grosor_mm = bill_depth_mm,
         aleta_longitud_mm = flipper_length_mm,
         masa_g = body_mass_g,
         sexo = sex,
         año = year) %>%
  mutate(ratio = masa_g / aleta_longitud_mm)
penguins <- df_col_rename
head(penguins, 5)
```

	especies <fct>	isla <fct>	pico_longitud_mm <dbl>	pico_grosor_mm <dbl>
1	Adelie	Torgersen	39.1	18.7
2	Adelie	Torgersen	39.5	17.4
3	Adelie	Torgersen	40.3	18.0
4	Adelie	Torgersen	*NA*	*NA*
5	Adelie	Torgersen	36.7	19.3

5 rows | 1-5 of 10 columns

Preguntas sobre los pinguinos

Tamaño

¿Qué especie de pinguino tiene un mayor tamaño?

Para empezar queremos saber **qué especie de pinguino tiene un mayor tamaño**. Para responder a esto nos fijaremos en sus **masas corporales**, **longitud de aletas** y **ratio** (que no deja de ser la relación entre las dos últimas). Utilizamos **group_by** para agrupar las observaciones por especie. Posteriormente, con **summarise** y **mean** calculamos la media de las tres variables. Ahora tenemos un nuevo data frame **penguins_size** con estos valores. El resultado lo ordenamos de manera descendiente con **arrange**. Como vemos, la especie que tiene un mayor tamaño es la **Gentoo**. De la misma forma podríamos calcular los valores máximos de estas variables **max**, los mínimos **min**, la desviación típica **sd** o la mediana **median**, por ejemplo.

Hide

```
penguins_means <- penguins %>%
  group_by(especies) %>%
  drop_na() %>%
  summarise(media_masa_g = mean(masa_g),
            media_aleta_longitud_mm = mean(aleta_longitud_mm),
            media_ratio = mean(ratio)) %>%
  arrange(desc(media_masa_g))
penguins_means
```

especies <fct>	media_masa_g <dbl>	media_aleta_longitud_mm <dbl>	media_ratio <dbl>
Gentoo	5092.437	217.2353	23.41415
Chinstrap	3733.088	195.8235	19.04376
Adelie	3706.164	190.1027	19.48037

3 rows

¿Dónde viven los más grandes?

Sabemos que hay datos biométricos para tres islas en el dataset. Nos preguntamos ¿Qué isla tiene los pinguinos más grandes? Para responder a esto, agrupamos por isla y calculamos ahora la media de **masa_g**. El resultado lo ordenamos de manera descendiente con **arrange**. Como vemos, la isla que tiene los pinguinos más grandes es **Biscoe**.

Hide

```
penguins_island <- penguins %>%
  group_by(isla) %>%
  drop_na() %>%
  summarise(media_masa_g = mean(masa_g)) %>%
  arrange(desc(media_masa_g))
penguins_island
```

isla <fct>	media_masa_g <dbl>
Biscoe	4719.172
Dream	3718.902
Torgersen	3708.511

3 rows

Especies en Biscoe

Pero, ¿Qué especies viven en la **isla Biscoe**? ¿Las tres, o solo alguna de ellas? Con **filter** podemos filtrar el dataset para ver solo los datos de los pinguinos de la isla **Biscoe**. Con **unique** vemos que solo Adelie y Gentoo están. De hecho, si contamos con **sum** el número de observaciones de cada especie vemos que hay 44 de Adelie y 124 de Gentoo, y ninguna de Chinstrap. Con **select** nos quedamos con esas columnas para pasarlas al nuevo data frame.

Hide

```
# data frame con los pinguinos de la isla Biscoe
penguins_biscoe <- penguins %>%
  filter(isla == "Biscoe") %>%
  drop_na() %>%
  select(especies, isla, pico_longitud_mm, pico_grosor_mm, aleta_longitud_mm, masa_g, sexo, año, ratio) %>%
  arrange(desc(especies))

# especies que hay en la isla Biscoe
unique(penguins_biscoe$especies)
```

```
## [1] Gentoo Adelie
## Levels: Adelie Chinstrap Gentoo
```

Hide

```
# número de observaciones de cada especie
sum(penguins_biscoe$especies == "Adelie")
```

```
## [1] 44
```

Hide

```
sum(penguins_biscoe$especies == "Gentoo")
```

```
## [1] 119
```

Hide

```
sum(penguins_biscoe$especies == "Chinstrap")
```

```
## [1] 0
```

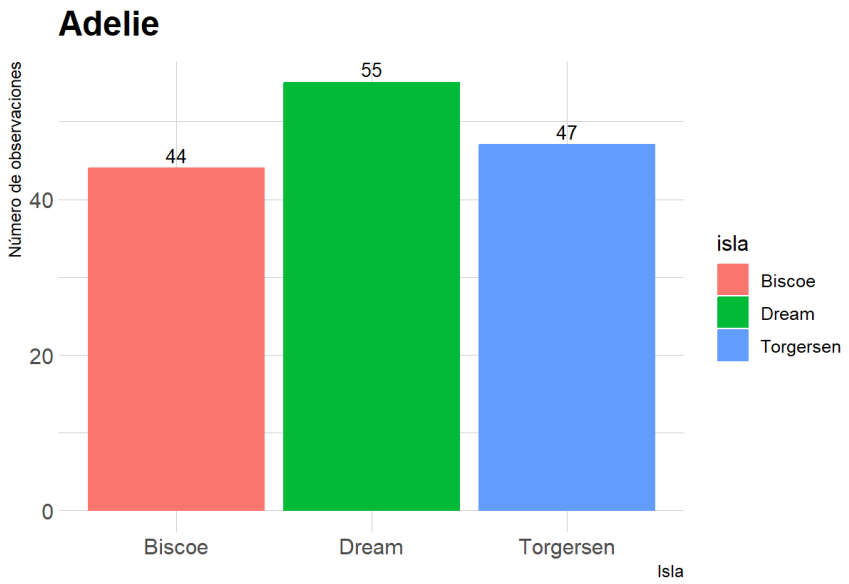
Observaciones de sp en cada isla

Vamos a representar el número de observaciones de cada especie, en cada una de las islas.

Hide

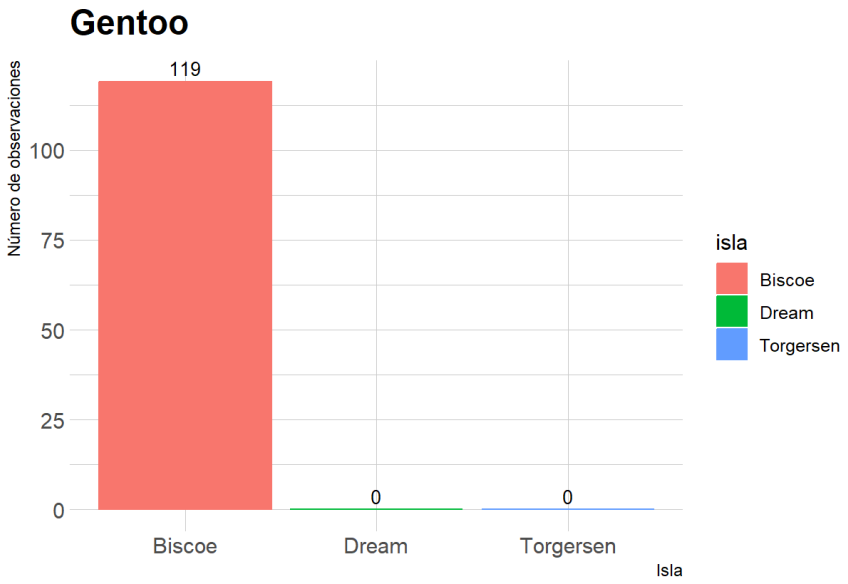
```
penguins_island_especies <- penguins %>%
  group_by(isla) %>%
  drop_na() %>%
  summarise(Adelie = sum(especies == "Adelie"),
            Chinstrap = sum(especies == "Chinstrap"),
            Gentoo = sum(especies == "Gentoo"))

penguins_island_especies %>%
  ggplot(aes(x = isla, y = Adelie, color = isla)) +
  geom_col(aes(fill = isla), position = "dodge") +
  labs(title = "Adelie",
       x = "Isla",
       y = "Número de observaciones") +
  geom_text(aes(label = Adelie), position = position_dodge(width = 0.9),
            color = "black", size = 3.5, vjust = -0.4) +
  theme_ipsum()
```



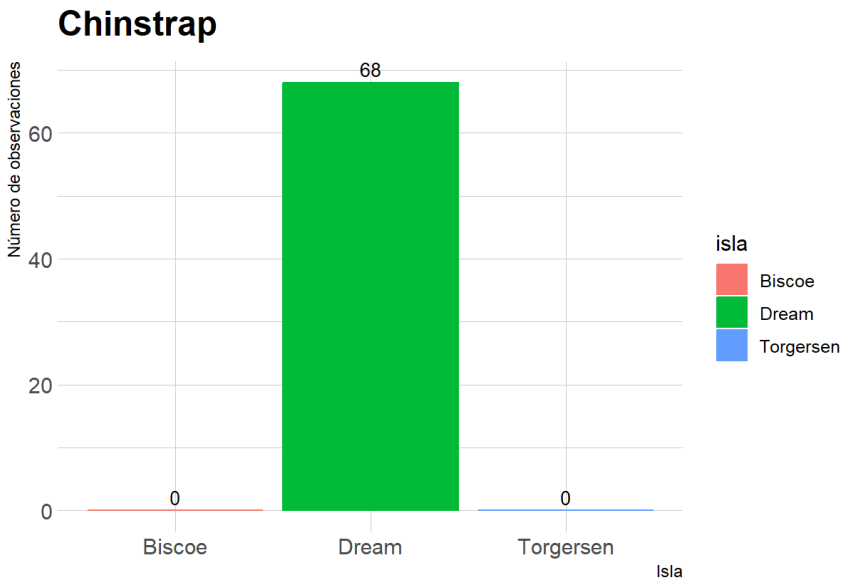
Hide

```
penguins_island_especies %>%
  ggplot(aes(x = isla, y = Gentoo, color = isla)) +
  geom_col(aes(fill = isla), position = "dodge") +
  labs(title = "Gentoo",
       x = "Isla",
       y = "Número de observaciones") +
  geom_text(aes(label = Gentoo), position = position_dodge(width = 0.9),
            color = "black", size = 3.5, vjust = -0.4) +
  theme_ipsum()
```



Hide

```
penguins_island_especies %>%
  ggplot(aes(x = isla, y = Chinstrap, color = isla)) +
  geom_col(aes(fill = isla), position = "dodge") +
  labs(title = "Chinstrap",
       x = "Isla",
       y = "Número de observaciones") +
  geom_text(aes(label = Chinstrap), position = position_dodge(width = 0.9),
            color = "black", size = 3.5, vjust = -0.4) +
  theme_ipsum()
```



Dimofismo sexual

¿Hay dimorfismo sexual en relación al tamaño de machos y hembras en las especies estudiadas (masa corporal y longitud de las aletas)? Para contestar a esta pregunta necesitamos crear un nuevo dataframe con los valores medios de las variables para cada una de las especies. Para ello, agrupamos por especie y sexo y calculamos la media de las variables. El resultado lo ordenamos de manera descendiente con **arrange**. Como vemos, **en las tres especies los machos son más grandes que las hembras, en ambas variables**.

Hide

```r
penguins_dimorfismo_sexual_masa <- penguins %>%
  group_by(especies, sexo) %>%
  drop_na() %>%
  summarise(media_masa_g = mean(masa_g)) %>%
  arrange(especies, desc(sexo))
penguins_dimorfismo_sexual_masa
```

especies <fct>	sexo <fct>	media_masa_g <dbl>
Adelie	male	4043.493
Adelie	female	3368.836
Chinstrap	male	3938.971
Chinstrap	female	3527.206
Gentoo	male	5484.836
Gentoo	female	4679.741

6 rows

Hide

```r
penguins_dimorfismo_sexual_aleta <- penguins %>%
  group_by(especies, sexo) %>%
  drop_na() %>%
  summarise(media_aleta_longitud_mm = mean(aleta_longitud_mm)) %>%
  arrange(especies, desc(sexo))
  penguins_dimorfismo_sexual_aleta
```

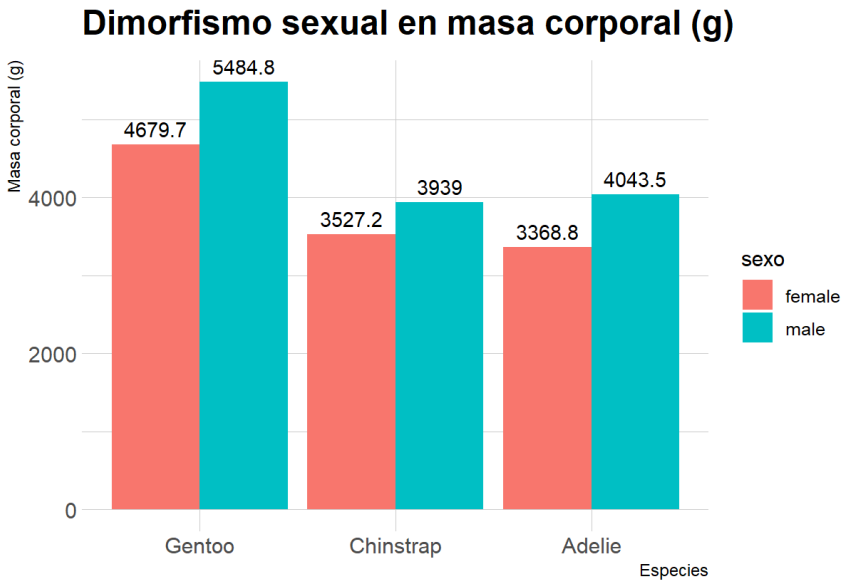
especies <fct>	sexo <fct>	media_aleta_longitud_mm <dbl>
Adelie	male	192.4110
Adelie	female	187.7945
Chinstrap	male	199.9118
Chinstrap	female	191.7353
Gentoo	male	221.5410
Gentoo	female	212.7069

6 rows

Utilizando el data frame de penguins_dimorfismo_sexual_masa, vamos a representar gráficamente la masa de machos y hembras en las tres especies de pinguinos. Para ello, utilizamos la función **ggplot**. Con **aes** indicamos que en el eje x queremos la especie y en el eje y la masa corporal. Con **geom_col** indicamos que queremos representar los datos en forma de columnas. Con **fill** indicamos que queremos que las columnas se rellenen con el color de la variable sexo. Con **labs** añadimos el título y las etiquetas de los ejes. Con **theme** cambiamos el color de fondo del gráfico.

Hide

```r
ggplot(penguins_dimorfismo_sexual_masa, aes(x = reorder(especies, -media_masa_g),
                                            y = media_masa_g)) +
  geom_col(aes(fill = sexo), position = "dodge") +
  labs(title = "Dimorfismo sexual en masa corporal (g)",
       x = "Especies",
       y = "Masa corporal (g)") +
  geom_text(aes(label = round(media_masa_g, 1), group = sexo),
            position = position_dodge(width = 0.9), vjust = -0.5) +
  theme_ipsum()
```



De la misma forma podemos hacerlo para la longitud de las aletas.

Hide

```
ggplot(penguins_dimorfismo_sexual_aleta, aes(x = reorder(especies,
                                                         -media_aleta_longitud_mm),
                                             y = media_aleta_longitud_mm)) +
  geom_col(aes(fill = sexo), position = "dodge") +
  labs(title = "Dimorfismo sexual en longitud de las aletas (mm)",
       x = "Especies",
       y = "Longitud de las aletas (mm)") +
  geom_text(aes(label = round(media_aleta_longitud_mm, 1), group = sexo),
            position = position_dodge(width = 0.9), vjust = -0.5) +
  theme_ipsum()
```