Análisis de Librería “palmerpenguins” - Victor Sojo

1. Objetivo del análisis:

Se trabajó con la librería llamada “*palmerpenguins*”, en donde se realizaron diferentes tipos de análisis. Inicialmente se debe de verificar la información que se localiza en la librería, su estructura, contenido, posibles variables de analizar, entre otras. Dentro de las variables se buscó conocer cuál especie posee mayor masa corporal, mayor longitud de aleta, media de la longitud del pico, estadísticas descriptivas, análisis de ANOVA para comparar la masa corporal de los individuos entre las islas, correlaciones entre longitud de la aleta y la masa corporal, modelo de regresión lineal para predecir la masa corporal en función de la longitud de la aleta, entre otros.

2. Resultados y Discusión:

Dentro de los datos de la librería, tenemos longitud y profundidad del pico, largo de la aleta y masa corporal por especie de pingüino. Entre los datos contamos con tres especies de pingüinos, distribuidos de la siguiente forma: Adelie 152 individuos, Chinstrap 68 individuos y Gentoo con 124 individuos, de los cuales Adelie posee la mayor variedad de pesos, seguida de Gentoo y en última posición Chinstrap. Con respecto al promedio de la longitud de las aletas y su masa corporal, la especie Gentoo es que posee una medida de 217 mm y 5.08 kg, seguida de Chinstrap con 196 mm y 3,73 Kg y por último la especie Adelie con 190 mm y 3,70 Kg. Así mismo, se logró identificar cual es el pingüino por especie que posee mayor tamaño en su aleta, como Adelie que presentó un individuo con 210 mm, Chinstrap un individuo con 212 mm y Gentoo un individuo con 231 mm. Se estimó, además cuales eran los individuos con mayor masa corporal, presentándose para la especie Adelie un individuo con 4,78 Kg, Chinstrap un individuo con 4,8 Kg y Gentoo un individuo con 6,3 Kg.

La especie Adelie existe mayor variedad de pesos, seguido de la especie Gentoo y por último la especie Chinstrap.

La mayor cantidad de pingüinos se localizan entre los 3 y 4 kilos, y un número muy reducido en los valores mayores a 6 kilogramos.

Al realizar la comparación de la masa corporal de los pingüinos por isla, se logra extraer que en la isla Biscoe, es donde se localizan los pingüinos de mayor peso, seguida de la isla Dream y por último la isla Torgersen.

Del análisis de ANOVA a la masa corporal, se logra extraer existe una diferencia significativa en la masa corporal entre los pingüinos de las diferentes islas

Se tiene que existe una correlación entre la longitud de la aleta y la masa corporal de 0,87 lo cual podría sugerir que existe una alta correlación entre las variables.

Finalmente al pretender generar una relación entre la longitud y la profundidad del piso por especie, se logró obtener que en muy pocos casos existe una relación entre dichas variables, concluyendo que no existe un patrón entre estas variables.

3. Resultados:

Gráfico 1. Variedad de pesos por especie de pingüinos

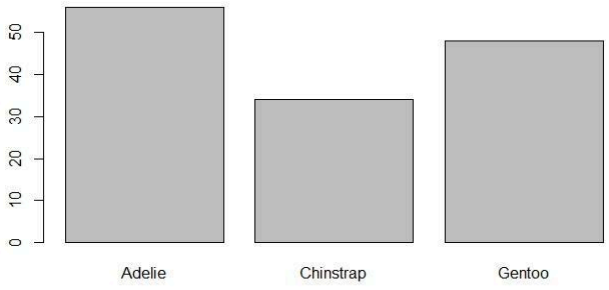


Gráfico 2. Histograma de la distribución de la masa por corporal en gramos de los pingüinos

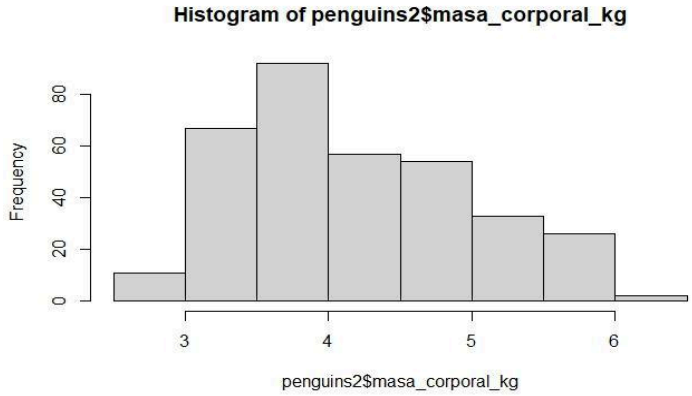


Gráfico 3. Distribución de la longitud de la aleta según la masa corporal de los pingüinos.

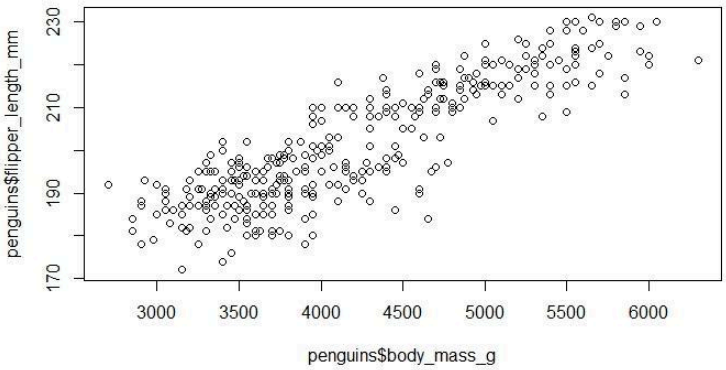


Gráfico 4. Comparación de la masa corporal de los pingüinos por isla.

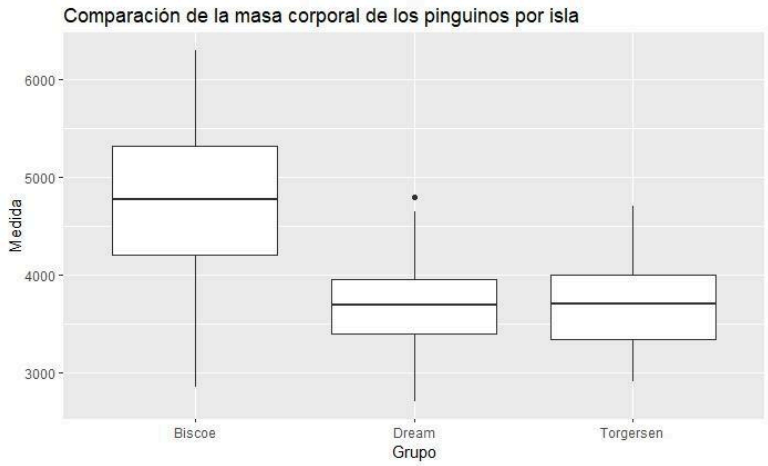
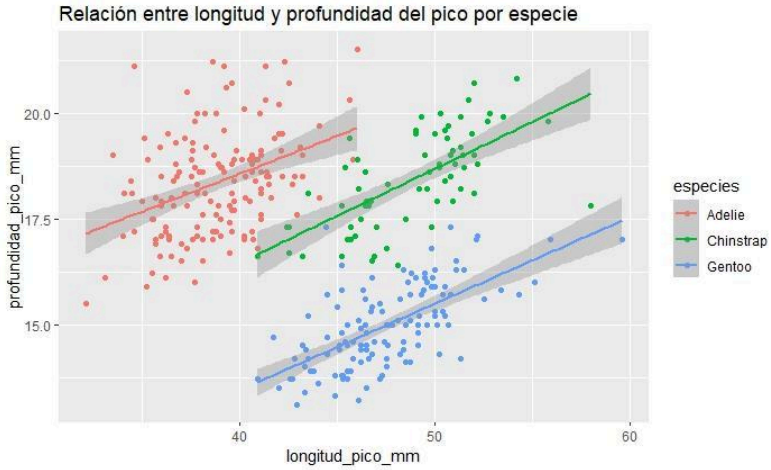


Gráfico 5. Relación entre la longitud y la profundidad del pico por especie.



4. **Anexo:**

```
#Se instala el paquete de datos "palmerpenguins"
installed.packages("palmerpenguins")
#Se carga la libreria necesarias
library(palmerpenguins)
library(tidyverse)
```

```
#cargar los valores de "penguins"
data(penguins)

#Ver la tabla de "penguins"
head(penguins)
head(penguins_raw)

#Conocer la estructura de los datos "penguins" y penguis_raw"
str(penguins)
str(penguins_raw)

#Verificar la clases de los datos "penguins" y "penguins_raw"
class(penguins)
class(penguins_raw)

#conocer la cantidad de filas en los datos "penguins" y "penguins_raw"
nrow(penguins)
nrow (penguins_raw)

#conocer la cantidad de columnas en los datos "penguins" y "penguins_raw"
ncol(penguins)
ncol(penguins_raw)

#conocer la dimensión de columnas en los datos "penguins" y "penguins_raw"
dim(penguins)
dim(penguins_raw)
```

```
#conocer la posición exacta de los datos nulos en "penguins" y "penguins_raw"
posiciones_na_penguins<- which(is.na(penguins))
(posiciones_na_penguins)
posiciones_na_penguins_raw<- which(is.na(penguins_raw))
(posiciones_na_penguins_raw)

#conocer la cantidad de valores nulos en el set de datos "penguins" y "penguins_raw"
length(which(is.na(penguins)))
length(which(is.na(penguins_raw)))

#conocer la cantidad de valores nulos por columna en el set de datos "penguins" y "penguins_raw"
cantidad_nulos_por_columna_penguins <- colSums(is.na(penguins))
cantidad_nulos_por_columna_penguins

cantidad_nulos_por_columna_penguins_raw <- colSums(is.na(penguins_raw))
cantidad_nulos_por_columna_penguins_raw

#conocer la cantidad de especies en el set de datos de "penguins" y "penguins_raw"
length(unique(penguins$species))
length(unique(penguins_raw$Species))

#Exploración de los datos
library(skimr)
```

```
skim(penguins)

skim(penguins_raw)


#Conocer los valores unicos del set de datos de "penguins" y "penguins_raw", en la columna "species"

unique(penguins$species)

unique(penguins_raw$Species)


#Realizar una gráfica sencilla del set de datos de "penguins", tomando como variable independiente masa corporal en gramos y la variable dependiente la longitud de la aleta en milímetros

plot(penguins$body_mass_g,penguins$flipper_length_mm)


#Generar un gráfico que nos muestre cual es la especie de pinguinos que posee mayor variedad de pesos

tipos <- by(penguins$body_mass_g,
            penguins$species,
            FUN = function(x) length(unique(x)))

tipos <- as.data.frame(tipos)

barplot(tipos$x,
        names.arg=rownames(tipos))


#Agregar una columna que transforme el peso de gramos a kilos, agregandome un tibble

penguins |>
  select(body_mass_g) |>
  mutate(body_mass_Kg = body_mass_g/1000)
```

```
#Agregar una columna al set de datos (Original) de "penguins" que muestre el peso en kilogramos

penguins <- penguins |>

  mutate(body_mass_kg = body_mass_g / 1000)

penguins


#Ordenar los datos en orden ascendente de peso en kilogramos del set de datos "penguins"

penguins |>

  select(body_mass_kg) |>

  arrange(body_mass_kg)


#Ordenar los datos en orden descendente de peso en kilogramos del set de datos "penguins"

penguins |>

  select(body_mass_kg) |>

  arrange(desc(body_mass_kg))


#Conocer la cantidad de pinguinos por especie

penguins |>

  group_by(species) |>

  summarise(cantidad = n())


#Obtener por tipo de especie, la media y desviación estandar de la longitud de la aletas,

#Ambas excluyendole los valores de "NA".

penguins |>

  group_by(species) |>
```

```
  summarise(Promedio = mean(flipper_length_mm, na.rm = TRUE),
            Desviación_Estandar = sd(flipper_length_mm, na.rm = TRUE))

#Obtener por tipo de especie, la media y desviación estandar de la masa corporal,
#Ambas excluyendole los valores de "NA".
penguins |>
  group_by(species) |>
  summarise(Promedio = mean(body_mass_kg, na.rm = TRUE),
            Desviación_Estandar = sd(flipper_length_mm, na.rm = TRUE))

#Resume de cual es la Longitud Total de las aletas por especie
penguins |>
  group_by(species) |>
  summarise(Longitud_de_la_Aleta = sum(flipper_length_mm,na.rm = TRUE)) |>
  ungroup()

#Se quiere conocer por especie cual es el pinguino con mayor longitud en su aleta
penguins |>
  group_by(species) |>
  arrange(desc(flipper_length_mm)) |>
  slice(1)

#Se quiere conocer por especie cual es el pinguino con mayor masa corporal
penguins |>
  group_by(species) |>
  arrange(desc(body_mass_kg)) |>
```

```
  slice(1)

# Calcular la media de la longitud del pico para cada especie, eliminando los NA
penguins |>
  group_by(species) |>
  summarize(media_pico = mean(bill_length_mm, na.rm = TRUE))

#Pruebas estadisticas para los datos penguins
library(rstatix)

# Preprocesamiento limpieza de datos, modificación de nombre de columnas

penguins2 <- penguins |>
  drop_na(bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g) |>
  select(species, island, bill_length_mm, bill_depth_mm, flipper_length_mm,
body_mass_g, sex) |>
  rename("especies" = "species",
         "isla" = "island",
         "longitud_pico_mm" = "bill_length_mm",
         "profundidad_pico_mm" = "bill_depth_mm",
         "longitud_aleta_mm" = "flipper_length_mm",
         "masa_corporal_gr" = "body_mass_g",
         "sexo" = "sex")

# Generar un estadisticas descriptivas
penguins2 |>
  get_summary_stats(type = "mean_sd")
```

```
# Generar una tablka de frecuencias por especie de pinguino
table(penguins2$especies)

#Histograma de masa por corporal de los pinguinos
hist(penguins2$masa_corporal_gr)

# Análisis de varianza (ANOVA) para comparar la masa corporal entre islas
anova_masa <- aov(masa_corporal_gr ~ isla, data = penguins2)
summary(anova_masa)

penguins2 |>
  anova_test(masa_corporal_gr ~ especies)


#Generar una grafico que compare la masas corporales de los pinguinos por isla.
library(ggplot2)
ggplot(penguins2, aes(x = isla, y = masa_corporal_gr)) +
  geom_boxplot() +
  labs(x = "Grupo", y = "Medida") +
  ggtitle("Comparación de la masa corporal de los pinguinos por isla")

# Calcular la correlación entre la longitud de la aleta y la masa corporal,
eliminando los NA
cor(penguins2$longitud_aleta_mm, penguins2$masa_corporal_gr, use =
"complete.obs")
```

```
# Modelo de regresión lineal para predecir la masa corporal en función de la longitud de la aleta

modelo_regresion <- lm(masa_corporal_gr ~ longitud_aleta_mm, data = penguins2, na.action = na.exclude)

summary(modelo_regresion)

plot(modelo_regresion)



#Crear un gráfico que me permita visualizar la relación entre la longitud del pico y la profundidad del pico por especie

library(ggplot2)

ggplot(penguins2, aes(x = longitud_pico_mm, y = profundidad_pico_mm, color = especies)) +

  geom_point() +

  geom_smooth(method = "lm") +

  ggtitle("Relación entre longitud y profundidad del pico por especie")


#Correlación parametrica pearson

penguins2 |>

  select(longitud_pico_mm, longitud_aleta_mm, profundidad_pico_mm, masa_corporal_gr) |>

  cor_test(vars = c(longitud_pico_mm, longitud_aleta_mm, profundidad_pico_mm, masa_corporal_gr),

           vars2 = c(longitud_pico_mm, longitud_aleta_mm, profundidad_pico_mm, masa_corporal_gr),

           method = "pearson")
```