Ciência de Dados com R

Avaliação

Antes de iniciar os exercícios, exceute os códigos a seguir:

```
# Instalando os pacotes necessários
install.packages("dplyr")
install.packages("ggplot2")
install.packages("palmerpenguins")

# Lendo os pacotes necessários
library(dplyr)
library(ggplot2)
library(palmerpenguins)

# Documentação da base de dados
?penguins

# Organizando a base de dados
dados = na.omit(penguins)         # Removendo os NAs (valores faltantes)
dados$year = factor(dados$year)   # Transformando a variável year em fator
```

1. Quantas linhas e colunas existem na base de dados?
2. Calcule o mínimo, a média e o máximo do peso dos pinguins.
3. Com o auxílio das funções do pacote `dplyr`, faça o que se pede:
   (a) Renomeie as variáveis do dataframe `dados` como quiser e salve no objeto `dados1`.
   (b) Ordene o dataframe `dados1` de acordo com uma variável de sua escolha salve no objeto `dados2`.
   (c) Filtre o dataframe `dados2` para remover a espécie `Gentoo` e salve no objeto `dados3`.
   (d) Adicione uma nova coluna ao dataframe `dados3` com a razão do comprimento pela profundidade (comprimento/profundidade) do bico dos pinguins e salve no objeto `dados_final`.
4. Com o auxílio da função `str`, identifique quais variáveis são fatores e quais são números/inteiros. A partir disso, faça os gráficos a seguir com o auxílio do pacote `ggplot2`.
   (a) Para as variáveis do tipo fator, faça o gráfico de barras de cada uma delas. (Nesses gráficos, espera-se que no eixo x tenhamos as categorias do fator e no eixo y a quantidade de observações para cada uma das categorias.)
   (b) Para as variáveis numéricas (números e inteiros), faça o histograma de cada uma delas. (Nesses gráficos, espera-se que no eixo x tenhamos os valores das variável em questão e no eixo y a quantidade de observações nos intervalos criados pelo gráfico.)

5. Explore graficamente a associação entre cada uma das variáveis do dataframe `dados_final` com o peso dos pinguins.

   (a) Faça boxplots para os fatores.

   (b) Faça gráficos de dispersão para variáveis numéricas.

   (c) Que variáveis parecem ter relação com o peso dos pinguins?

   Observação: nos gráficos acima, espera-se o peso no eixo y e as demais variáveis no eixo x.

6. A análise exploratória pode nos ajudar a escolher variáveis para compor um bom modelo de regressão linear. Dado isso, faça o que se pede a seguir.

   (a) Ajuste pelo menos 3 modelos de regressão para explicar o peso dos pinguins com o auxílio da função `lm`.

   (b) Compare os modelos através da função `AIC` e escolha o melhor entre eles.

   (c) Quais variáveis são boas para explicar o peso dos pinguins? (Dica: veja se o p-valor da variável é menor que 0.05 no melhor modelo da letra (c).)

   (d) Qual a porcentagem da variável peso é explicada pelas demais no melhor modelo? (Dica: olhe para o `Adjusted R-squared`.)