Exercícios de Fixação*

Augusto Peroni Baldino e Bruno Grigoletti Laitano

Coleta, Preparação e Análise de Dados

1 Etapa Teórica

f1	f2	f3	f4
0.02	True	red	8.25
−2.00	False	red	1.35
0.57	True	yellow	21.86
1.00	False	black	3.06
−2.00	True	black	1.58
0.84	True	red	1.25
0.66	False	red	1.68
−1.00	False	yellow	1.01
0.44	True	red	1.19
0.32	True	red	1.47

Para responder as questões listadas abaixo, transformamos a tabela acima em um *DataFrame*, utilizando operações nativas da biblioteca *Pandas*:

```
    data = {
        "f1": [0.02, -2, 0.57, 1, -2, 0.84, 0.66, -1, 0.44,
           0.32],
        "f2": [True, False, True, False, True, True, False,
           False, True, True],
        "f3": ["red", "red", "yellow", "black", "black", "red"
           , "red", "yellow", "red", "red"],
        "f4": [8.25, 1.35, 21.86, 3.06, 1.58, 1.25, 1.68,
           1.01, 1.19, 1.47]
    }

    df = pd.DataFrame(data)
```

*O nosso repositório está disponível aqui.

a) Quais atributos são categóricos e quais atributos são numéricos? Considerando as informações dadas, é possível identificar dentre os atributos numéricos quais são os seus tipos?

R.: Os atributos categóricos estão listados nas colunas `f2` e `f3`. Os atributos numéricos, por sua vez, estão listados nas colunas `f1` e `f4`. Dada a formatação dos valores, podemos concluir que os atributos numéricos são do tipo `float64`.

```
print(df.dtypes)
```

b) Qual é a moda de `f3`?

R.: A moda de `f3` é `red`.

```
moda_f3 = df['f3'].mode()[0]
print(f"Moda da coluna 'f3': {moda_f3}")
```

c) Compute a média, a mediana, o desvio padrão e a variância dos atributos `f1` e `f4`.

R.: Para a coluna `f1`, temos os seguintes valores:

- Média: `-0.11500000000000002`;
- Mediana: `0.38`;
- Desvio padrão: `1.1367717448986847`;
- Variância: `1.2922500000000001`.

Para a coluna `f4`, temos os seguintes valores:

- Média: `4.269999999999999`;
- Mediana: `1.525`;
- Desvio padrão: `6.551145786128645`;
- Variância: `42.917511111111104`.

```
mean_f1 = df['f1'].mean()
print(f"Média da coluna 'f1': {mean_f1}")

median_f1 = df['f1'].median()
print(f"Mediana da coluna 'f1': {median_f1}")

desvio_f1 = df['f1'].std()
print(f"Desvio padrão da coluna 'f1': {desvio_f1}"
    )

variancia_f1 = df['f1'].var()
```

```
print(f"Variância da coluna 'f1': {variancia_f1}")

mean_f4 = df['f4'].mean()
print(f"Média da coluna 'f4': {mean_f4}")

median_f4 = df['f4'].median()
print(f"Mediana da coluna 'f4': {median_f4}")

desvio_f4 = df['f4'].std()
print(f"Desvio padrão da coluna 'f4': {desvio_f4}")

variancia_f4 = df['f4'].var()
print(f"Variância da coluna 'f4': {variancia_f4}")
```

d) Calcule a covariância entre os atributos `f1` e `f4`.

R.: A covariância entre os atributos `f1` e `f4` é de `1.8753888888888883`.

```
covariancia = df['f1'].cov(df['f4'])
print(f"Covariância entre 'f1' e 'f4': {covariancia}")
```

e) Calcule a correlação existente entre os atributos `f1` e `f4`.

R.: A correlação existente entre os atributos `f1` e `f4` é de `0.25182613734536186`.

```
correlacao = df['f1'].corr(df['f4'])
print(f"Correlação entre 'f1' e 'f4': {correlacao}")
```

2 Etapa Prática

1. Considerando que a variável `target` refere-se à espécie do pinguim, quais são as *features* categóricas e numéricas que temos disponíveis?

   R.: Utilizando operações nativas da biblioteca *Pandas*, observamos os tipos de dados de cada uma das colunas do *dataset*:

```
data = pd.read_csv("penguins.csv")

data.dtypes
```

   A seguir, a saída da operação utilizada:

```
studyName            object
Sample Number         int64
```

```
Species                object
Region                 object
Island                 object
Stage                  object
Individual ID          object
Clutch Completion      object
Date Egg               object
Culmen Length (mm)    float64
Culmen Depth (mm)     float64
Flipper Length (mm)   float64
Body Mass (g)         float64
Sex                    object
Delta 15 N (o/oo)     float64
Delta 13 C (o/oo)     float64
Comments               object
dtype: object
```

Com base nos resultados acima, concluímos que as *features* categóricas são: `studyName`, `Species`, `Region`, `Island`, `Stage`, `Individual ID`, `Clutch Completion`, `Date Egg`, `Sex` e `Comments`.

Por outro lado, as *features* numéricas são: `Sample Number`, `Culmen Length (mm)`, `Culmen Depth (mm)`, `Flipper Length (mm)`, `Body Mass (g)`, `Delta 15 N (o/oo)` e `Delta 13 C (o/oo)`.

2. Existem dados faltantes no conjunto de dados? Quantos e em quais *features*?

   R.: Sim, há dados faltantes em algumas das colunas do arquivo `penguins.csv`. Encontramos a quantidade de valores em cada uma das *features* utilizando a seguinte operação, executada sobre o mesmo *DataFrame* `data`:

```
1          data.isnull().sum()
```

A seguir, a saída da operação utilizada:

```
studyName              0
Sample Number          0
Species                0
Region                 0
Island                 0
Stage                  0
Individual ID          0
Clutch Completion      0
Date Egg               0
Culmen Length (mm)     2
Culmen Depth (mm)      2
Flipper Length (mm)    2
Body Mass (g)          2
Sex                   10
Delta 15 N (o/oo)     14
Delta 13 C (o/oo)     13
Comments             290
dtype: int64
```

3. Quais espécies de pinguim existem no conjunto de dados? As classes estão balanceadas?

   R.: Utilizamos as seguintes operações para descobrir a quantidade de espécies existentes no *dataset*, além de contabilizar a quantidade de registros relacionados a cada uma das espécies:

```
data['Species'].unique()
data['Species'].value_counts()
```

   A seguir, a saída das operações utilizadas:

```
Espécies presentes no dataset:
['Adelie Penguin (Pygoscelis adeliae)' 'Gentoo penguin (Pygoscelis
↪  papua)'
 'Chinstrap penguin (Pygoscelis antarctica)']

Quantidade de registros por espécie:
Species
Adelie Penguin (Pygoscelis adeliae)          152
Gentoo penguin (Pygoscelis papua)            124
Chinstrap penguin (Pygoscelis antarctica)     68
Name: count, dtype: int64
```

   Podemos ver com base nos resultados obtidos que não há um balanceamento de espécies em meio ao *dataset*.

4. Pesquise como plotar um *boxplot* usando a biblioteca `Seaborn`. Escreva o código necessário para criar o gráfico mostrado no enunciado. Escreva uma pequena análise sobre o gráfico gerado.

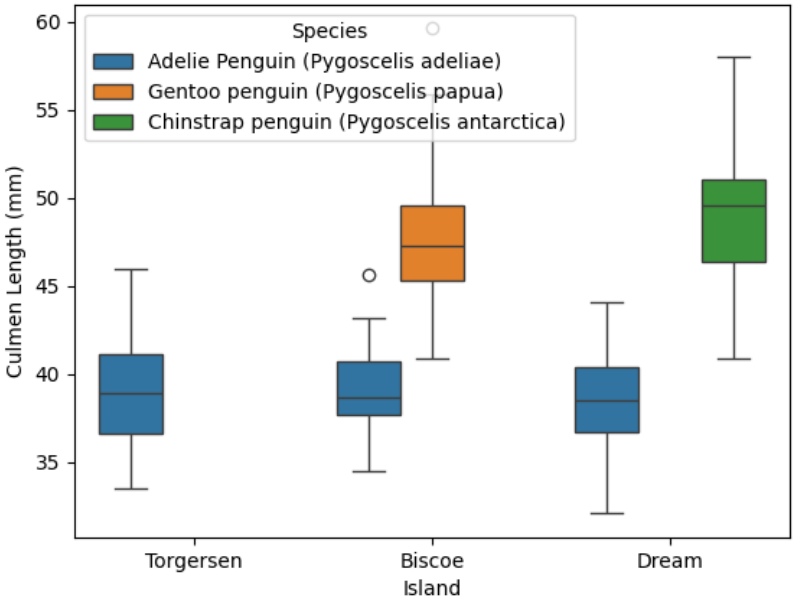

Figura 1: *Boxplot* gerado com a biblioteca `Seaborn`

R.: O gráfico acima foi gerado utilizando o seguinte código:

```
1          sns.boxplot(data=data, x="Island", y="Culmen
               Length (mm)", hue="Species")
2          plt.show()
```

O *boxplot* representa graficamente uma distribuição. No caso do gráfico gerado pelo nosso código, podemos ver os limites superiores e inferiores de cada `Culmen Length (mm)`, considerados com base na ilha de origem das três espécies listadas no conjunto. Além disso, o gráfico exibe valores estatísticos importantes, como o terceiro e o primeiro quartis, bem como a mediana (segundo quartil).

Nesse sentido, podemos concluir que o bico dos pinguins da espécie Adelie Penguin são menores nas ilhas observadas em comparação ao bico dos pinguins das espécies Gentoo Penguin e Chinstrap Penguin.

Outro dado relevante evidenciado pelo nosso *boxplot* é que há *outliers* em dois subconjuntos específicos — pinguins das espécies Adelie Penguin e Gentoo Penguin provenientes da ilha de Biscoe Island.

5. Faça o mesmo para o *scatter plot* mostrado no enunciado.

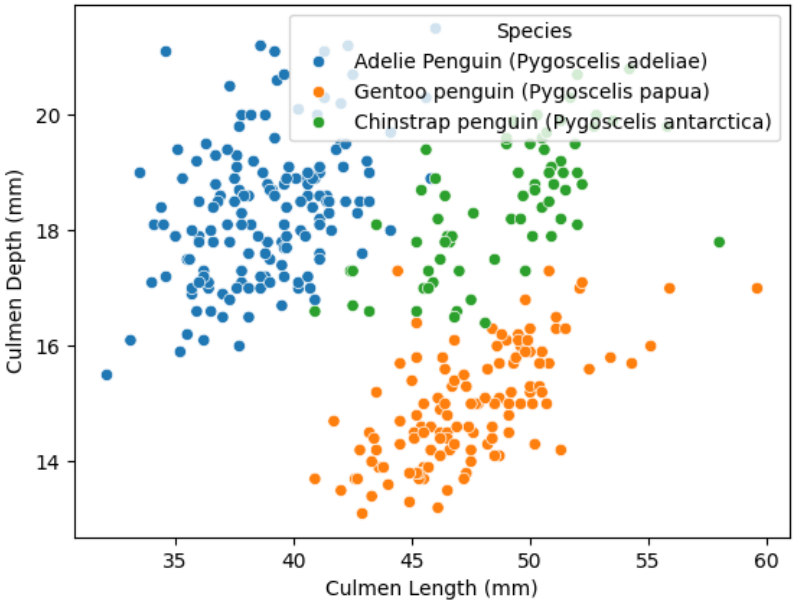

Figura 2: *Scatter plot* gerado com a biblioteca `Seaborn`

R.: O gráfico acima foi gerado utilizando o seguinte código:

```
1          sns.scatterplot(data=data, x="Culmen Length (mm)",
               y="Culmen Depth (mm)", hue="Species")
2          plt.show()
```

Em um *scatter plot*, ou gráfico de dispersão, cada par de valores é tratado como um par de coordenadas e plotado como pontos no plano. Em essência, é uma representação dos dados de duas ou mais variáveis.

No gráfico gerado, observamos o grau de dispersão entre o comprimento e a profundidade do bico dos pinguins das três espécies observadas no *dataset*. No caso da espécie Adelie Penguin, o bico é mais curto e mais profundo em comparação às outras duas espécies. Os pinguins da espécie Gentoo Penguin possuem bicos um pouco mais longos, porém menos profundos. Já os pinguins da espécie Chinstrap Penguin possuem bicos tão profundos quanto os bicos da espécie Adelie Penguin, mas são mais longos, tanto quanto são os bicos da espécie Gentoo Penguin.