Aluno: Pedro Wendling Hernandes

Análise Exploratória: Conjunto de dados sobre pinguins providenciado da Palmer Station LTER, na Antártica

Conjunto retirado de Gorman KB, Williams TD, Fraser WR (2014) Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus *Pygoscelis*). PLoS ONE 9(3): e90081. doi:10.1371/journal.pone.0090081.

Arquivo penguins_lter.csv retirado de https://www.kaggle.com/datasets/parulpandey/palmer-archipelago-antarctica-penguin-data.

Contém estudos sobre dismorfismo sexual ecológico e variabilidade ambiental sobre 3 espécies de pinguins contidas no arquipélago Palmer, na Antártica.

Gráficos gerados com o auxílio do módulo matlibplot.pyplot, em Python.

Varíaveis analisadas:

- studyName: código alfanumérico que diz em qual estudo que a amostra coletada está contida;

- Sample Number: identificador númerico de amostra;
  - Baixa relevância para a análise, alta relevância organizacional;

- Species: indica espécie do pinguim analisado;

- Region: região na qual foram encontrados os pinguins;
  - Irrelevante para esta análise, todas as amostras foram coletadas de apenas da região de Anvers;

- Island: ilha na qual foram encontrados os pinguins;
  - Contida em região;

- Stage: estágio de vida do pinguim estudado;
  - Irrelevante para a análise, visto que neste caso não há atributos discrepantes assim como em Region;

- Individual ID: código alfanumérico identificador de amostra;
    - baixa relevância para a análise, alta relevância organizacional;
    - em conjunto com studyName e Sample Number forma identificador da amostra;

- Clutch Completion: varia entre 'Yes' ou 'No', diz respeito se o pinguim estudado passou do período de aninhamento de ovos;

- Date Egg: data de iniciação de aninhamento de ovos;

- Culmen Length (mm): medidas de comprimento de cúlmen coletadas das amostras;
    - o cúlmen ou cumeeira é a crista dorsal do bico das aves;

- Culmen Depth (mm): medidas de profundidade de cúlmen coletadas das amostras;

- Flipper Length (mm): medidas de comprimento das nadadeiras coletadas das amostras;

- Body Mass (g): medidas da massa corporal coletadas das amostras;

- Sex: indica se o pinguim analisado é macho ou fêmea, apenas valores 'MALE' ou 'FEMALE';

- Delta 15 N (o/oo): razão entre isótopos estáveis de nitrogênio em sangue coletado;

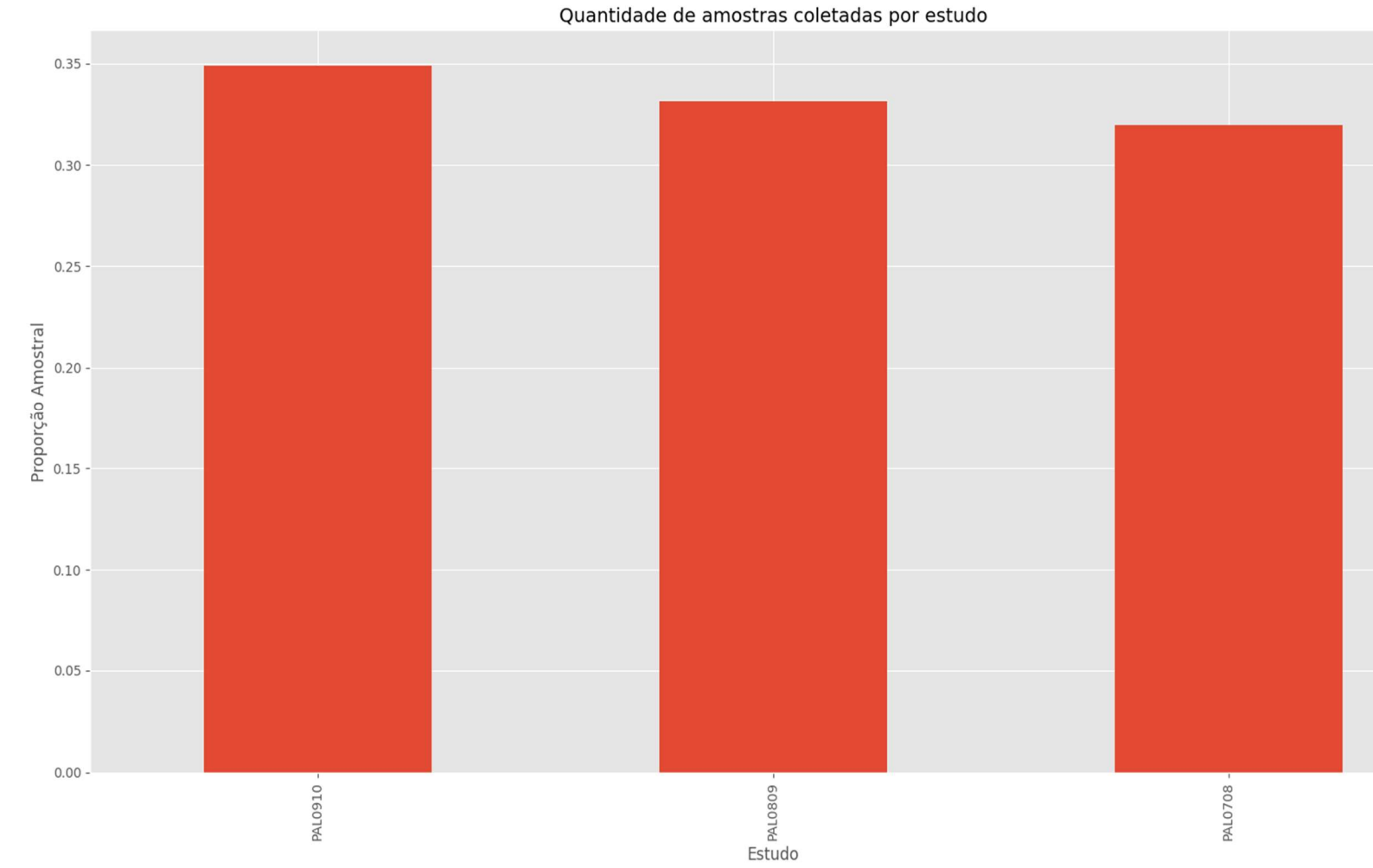
- Delta 13 C (o/oo): razão entre isótopos estáveis de carbono em sangue coletado;

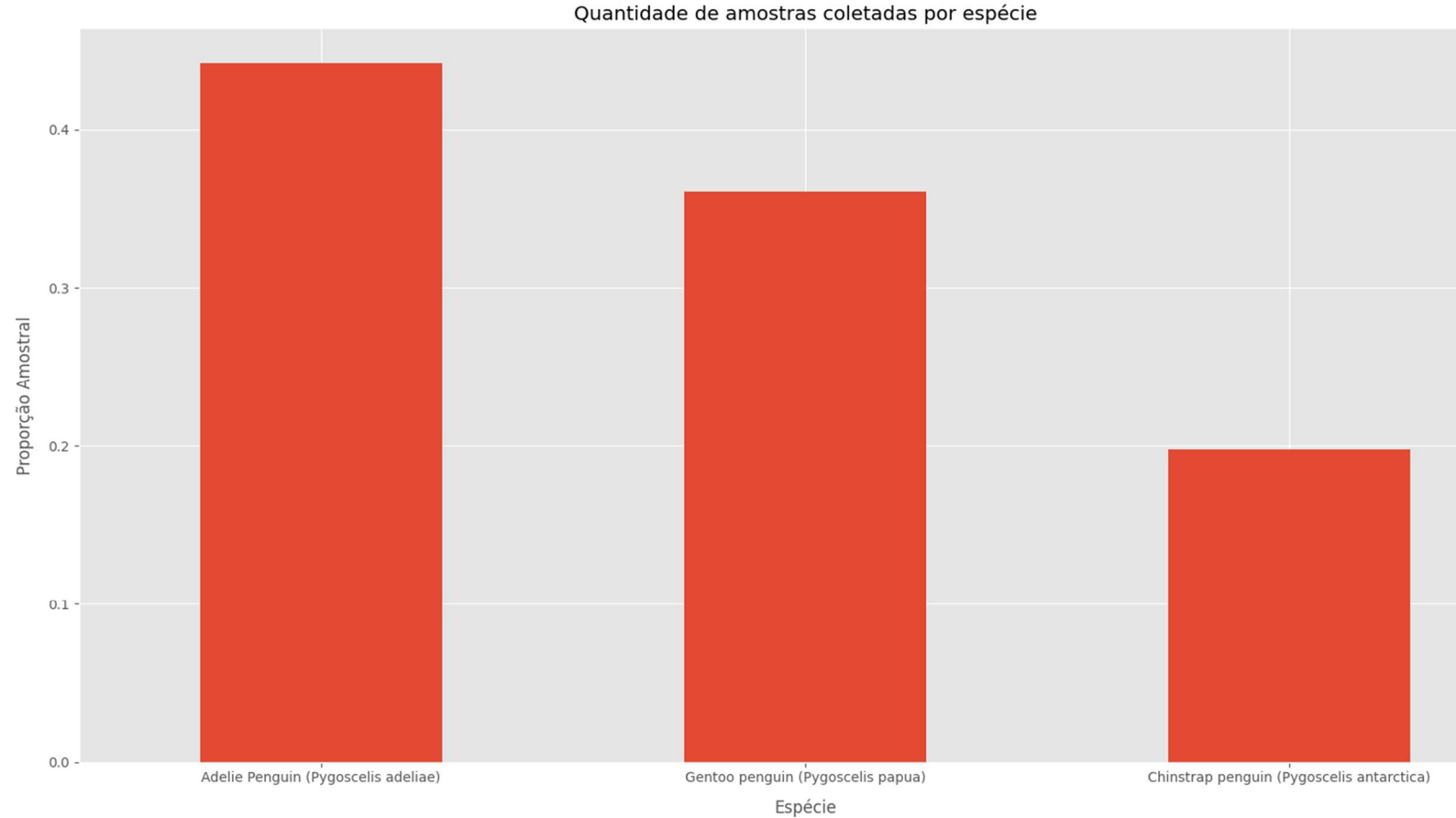
- Comments: comentários adicionais sobre o processo de coleta.

Questões pertinentes:

1. As amostras estão distribuídas balanceadamente?

Observaremos os seguintes gráficos, que dizem respeito à distribuição das amostras por estudo e por espécie, respectivamente:

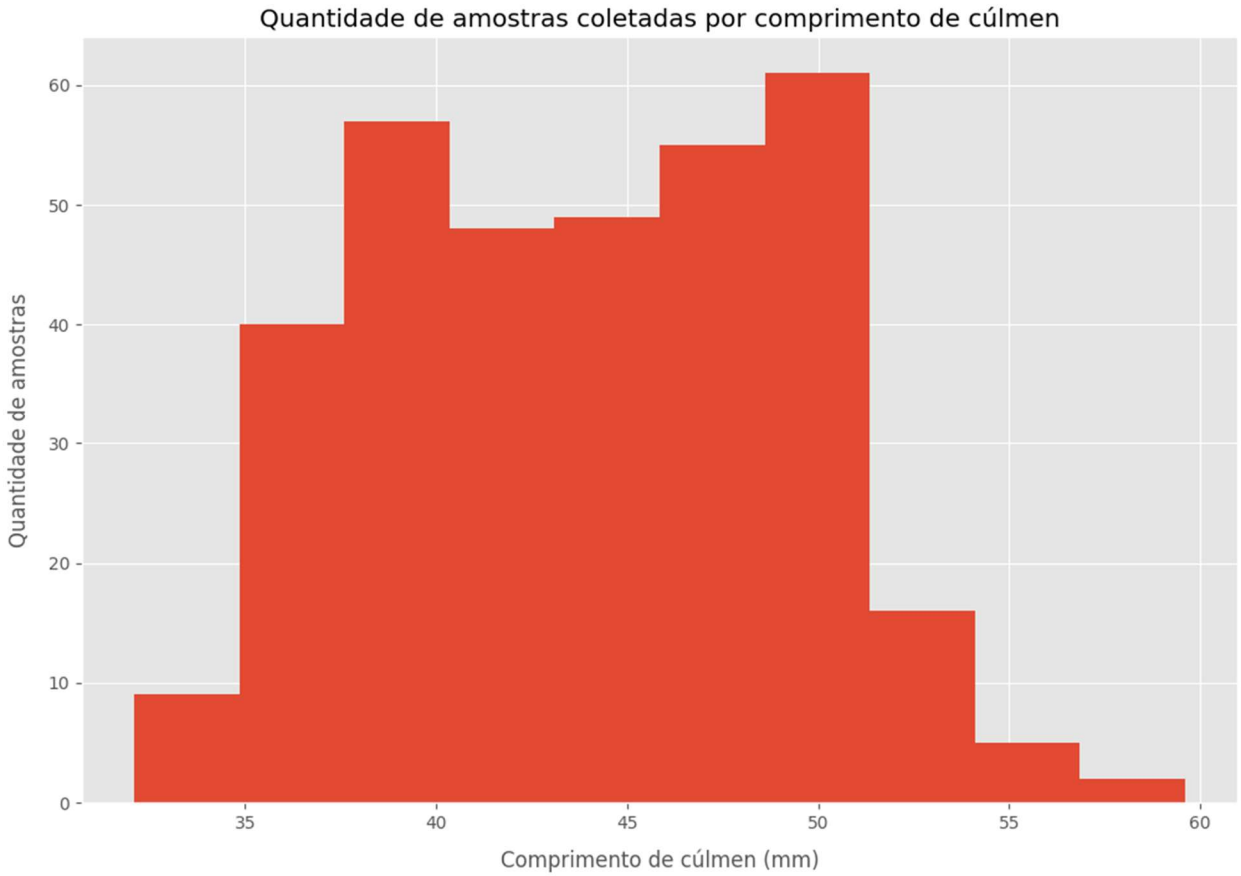


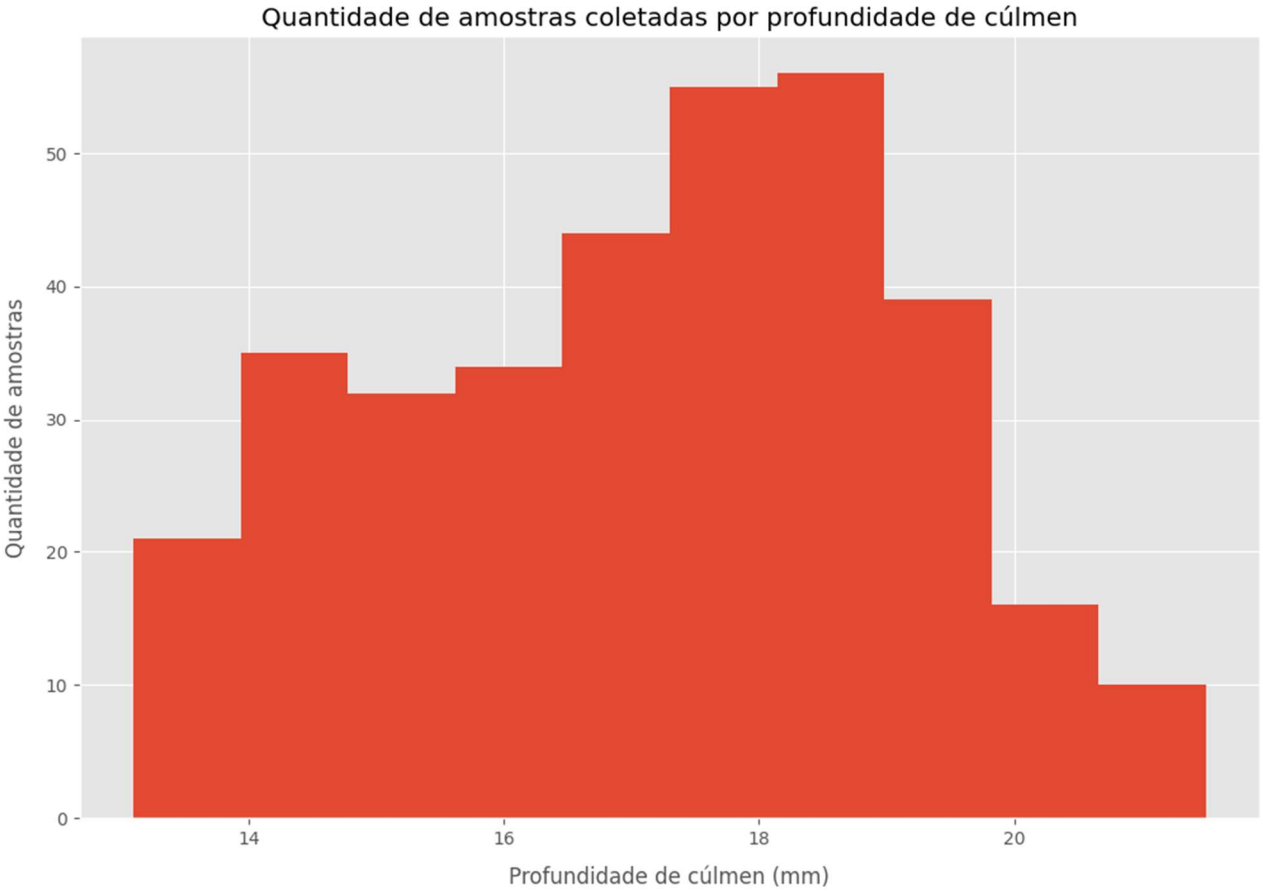


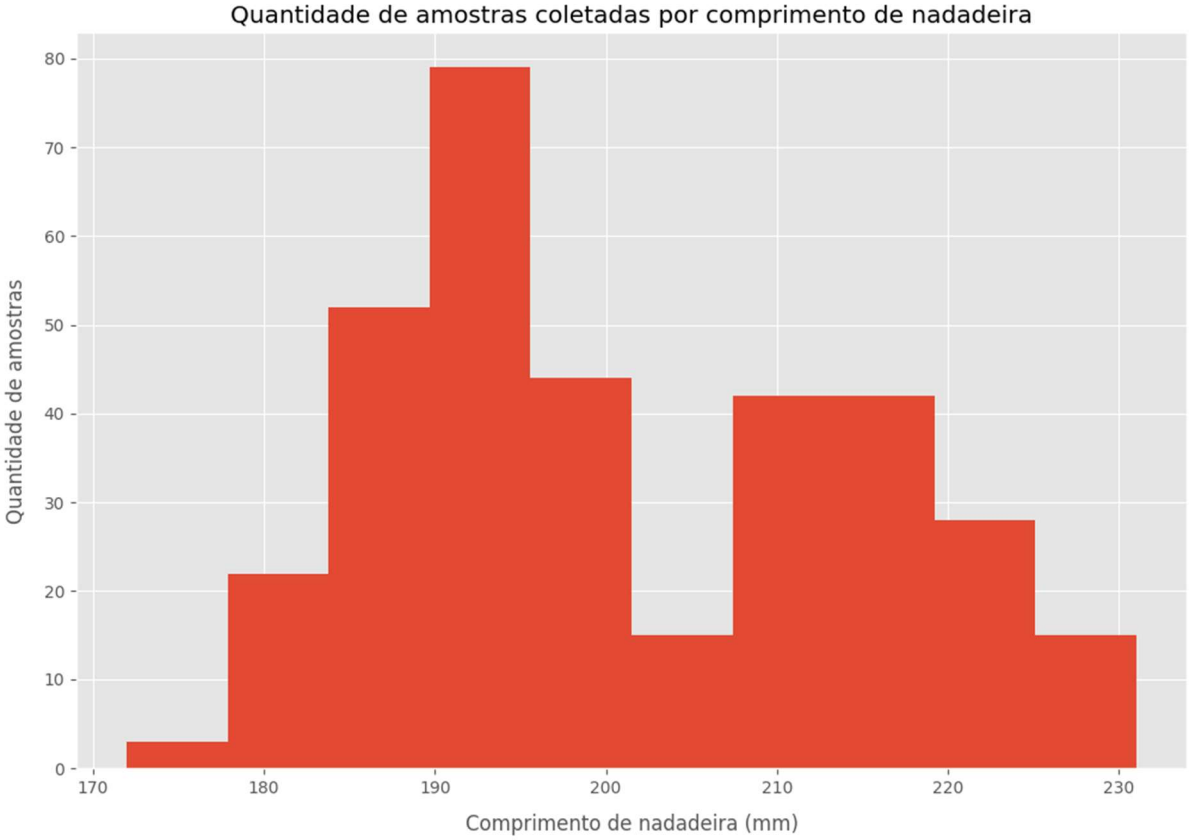
Nota-se que cada estudo possui aproximadamente 1/3 das amostras coletadas, porém existe discrepância significativa na proporção amostral entre as espécies de pinguim, principalmente entre a quantidade de pinguins-de-adélia e pinguins de barbicha. Podemos afirmar que os estudos possuem uma quantidade balanceada de amostras entre si, mas devemos ter em mente que os dados coletados sobre pinguins de barbicha podem ser inconclusivos.

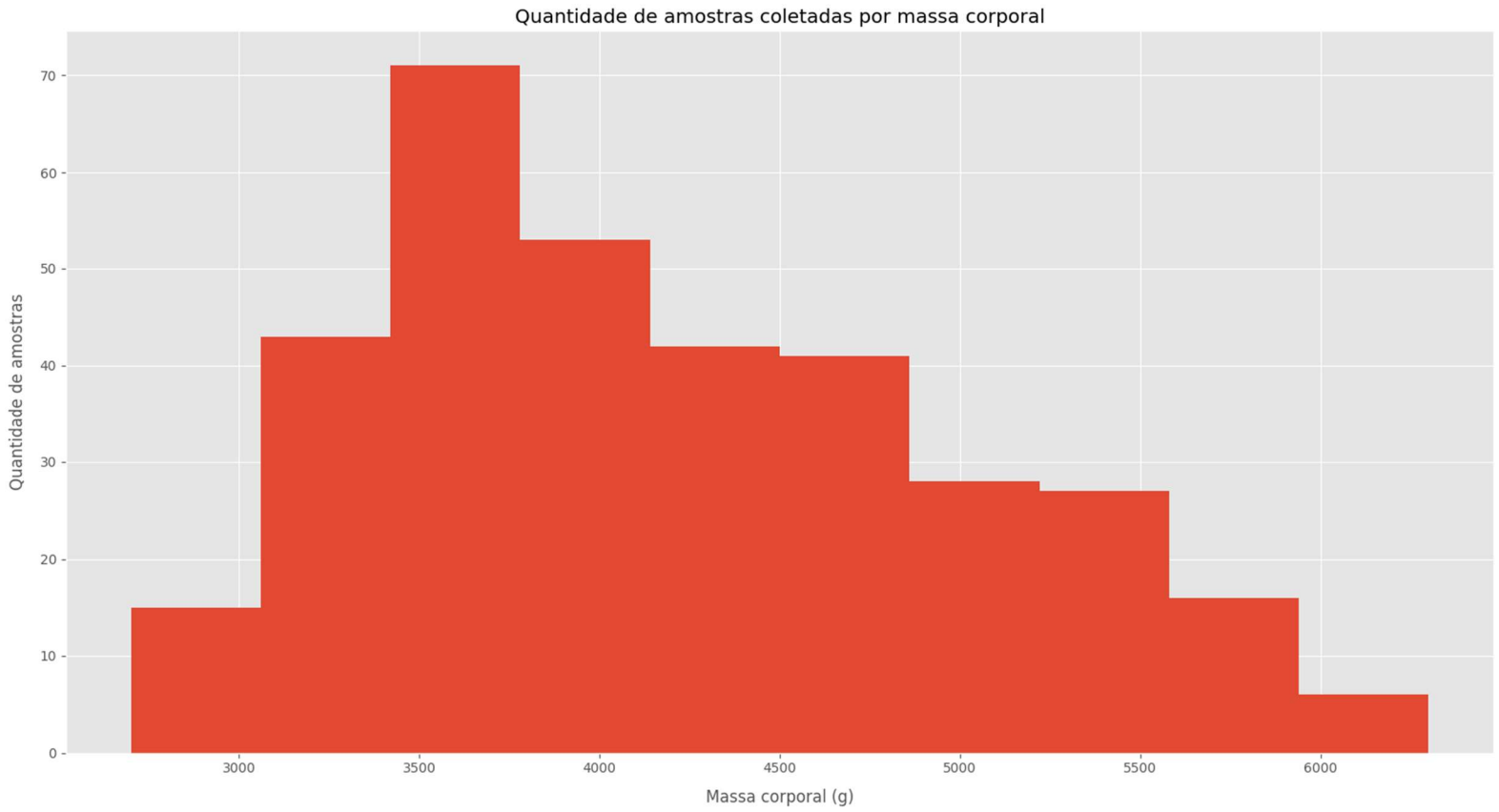
2. A) Quais são as medidas de tamanho e massa mais frequentes, e as médias delas?

   Histogramas de medidas de cúlmen, nadadeira e massa em relação à quantidade de pinguins examinados, respectivamente:




Quantidade de amostras coletadas por profundidade de cúlmen
Quantidade de amostras
0
10
20
30
40
50
14
16
18
20
Profundidade de cúlmen (mm)


Quantidade de amostras coletadas por comprimento de nadadeira
Quantidade de amostras
0
10
20
30
40
50
60
70
80
170
180
190
200
210
220
230
Comprimento de nadadeira (mm)



Observando os histogramas, podemos assumir que:

- Tamanho de cúlmen mais comum: comprimento aproximadamente entre 48,5mm e 51,5mm, profundidade aproximadamente entre 18mm e 19mm;
- Comprimento de nadadeira mais comum: aproximadamente de 190mm a 195mm;
- Quantidade de massa mais comum: aproximadamente de 3450g a 3750g

Aplicando Python com o auxílio do módulo Pandas para armazenar dados do arquivo .csv em uma variável dataframe e manuseá-la para a análise, conseguimos encontrar dados estatísticos assim como a média precisa de valores numéricos. Sendo df o dataframe correspondente a penguins.lter.csv, temos:

- Média de comprimento de cúlmen: `df.CulmenLength_mm.mean()`
  -Aproximadamente 43,92mm;

- Média de profundidade de cúlmen: `df.CulmenDepth_mm.mean()`
  -Aproximadamente 17,15mm;

- Média de comprimento de nadadeira: `df.FlipperLength_mm.mean()`
  -Aproximadamente 200,92mm;

- Massa média: `df.BodyMass_g.mean()` = Aproximadamente 4201,75g.

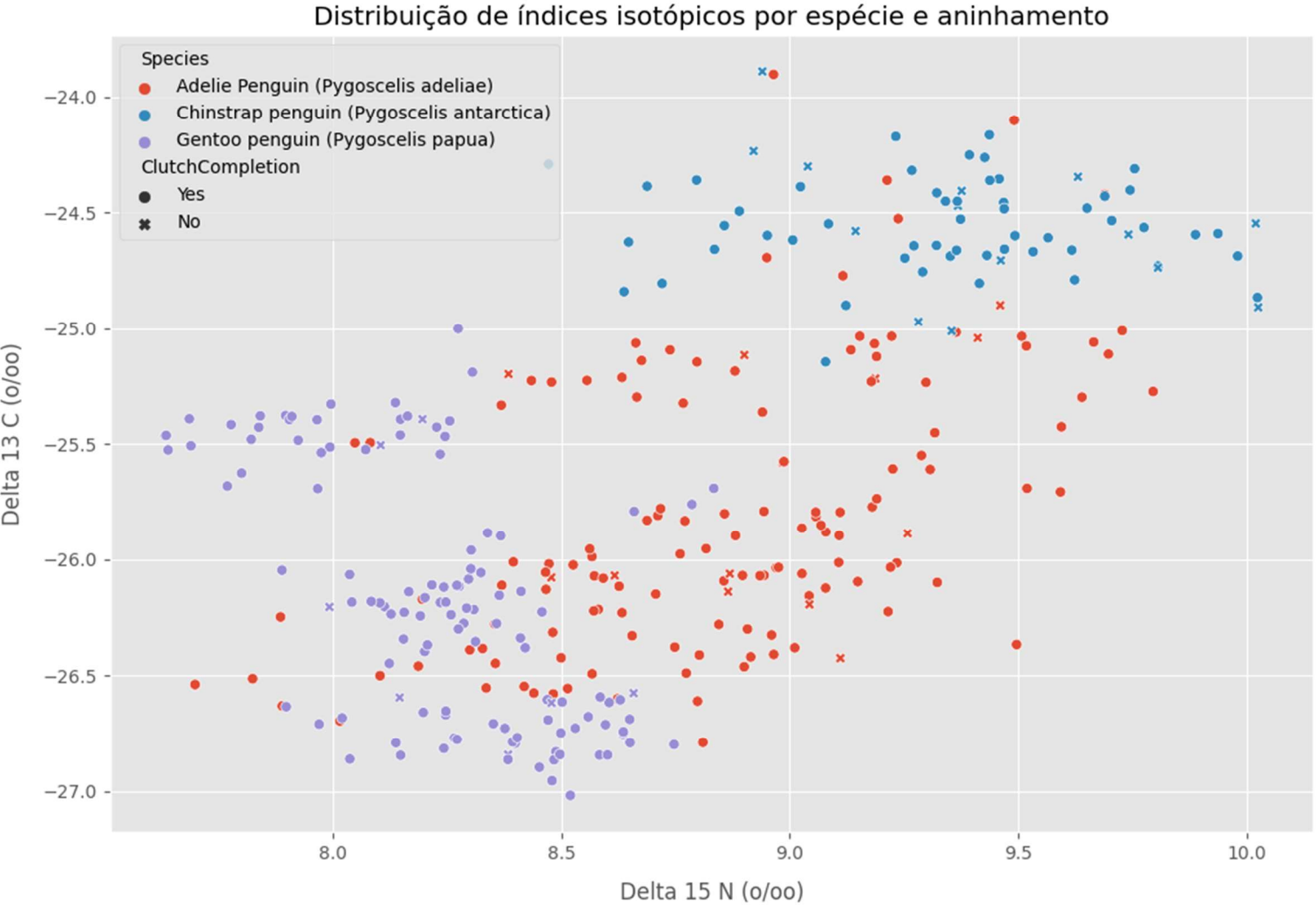
B) Descubra também as médias entre os índices de isótopos:

- Média de nitrogênio: `df.Delta15N_ooo.mean()` = 8,73 o/oo
- Média de carbono: `df.Delta13C_ooo.mean()` = -25,69 o/oo

*o/oo é medida para fração molar

3. Veja se há discrepâncias de índices isotópicos e de completude de aninhamento por espécie.

Conseguimos notar pelo seguinte gráfico de dispersão:



Observando o gráfico, notamos que pinguins Gentoo costumam possuir índices de isótopos mais baixos em relação às demais espécies, enquanto pinguins de barbicha costumam possuir índices isotópicos maiores. Nota-se que o fechamento do período de aninhamento não afeta os índices isotópicos.