

MODELO DE ARTIGO (GRUPO DE TRABALHO)

REGRESSÃO DE GAMBÁ

GT: Vinícius Kawasugui Santiago

Acadêmico do Curso de Engenharia de Software do Centro Universitário Cesumar – UNICESUMAR, Curitiba – PR. Kawav6390@gmail.com

RESUMO

Prever a idade de um gambá, comprimento da cabeça e se é macho ou fêmea usando modelo de regressão utilizando dados do dataset openintro-possum.

PALAVRAS-CHAVE:

Análise de dados; Gambá; Regressão.

INTRODUÇÃO

O gambá tem se mostrado ser uma ferramenta essencial para a compreensão de variações morfológicas dentro da espécie. Diferenças populacionais e sexuais podem indicar processos evolutivos ou adaptações ecológicas. Utilizando o dataset openintro-possum disponível no site kaggle, reúne dados morfométricos de indivíduos capturados em diferentes regiões, contendo uma base relevante para estudos exploratórios. O objetivo é identificar padrões que possam ajudar para estudos futuros sobre diversidade intraespecífica e adaptação ambiental.

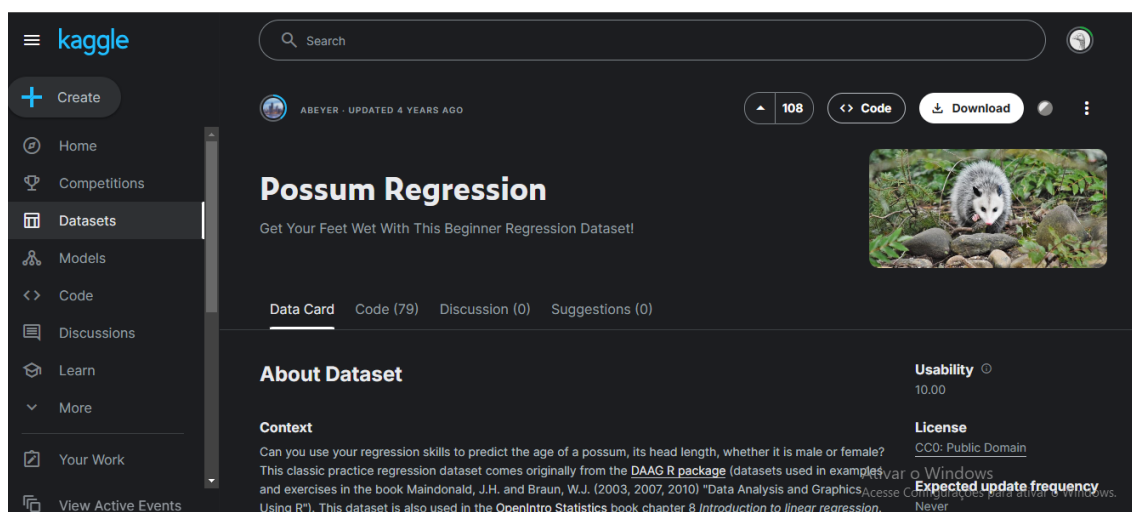


Figura 1 – Dataset Possim regression

Fonte: Kaggle/possum regressio

1 DESENVOLVIMENTO

Análise realizada a partir do dataset openintro-possum, contendo informações dos gambas, com variáveis consideradas foram sexo, população, comprimento da cabeça, largura, comprimento total, cauda, tórax abdômen. Foi feita a coleta dos dados em python, utilizando pandas, e tabelas com matplotlib

Resumo estatístico

Métrica	Valor
Idade Média	4.60
Peso Médio (kg)	2.78
Comprimento Cabeça (cm)	92.48
Comprimento Skul (cm)	57.28
Comprimento Dente (cm)	23.84
Comprimento Pé (cm)	35.79
Comprimento Cauda (cm)	36.76

Figura 2 – resumo estático descritivo

Fonte: Dados de pesquisa

Amostra dos dados

ID	Site	Sexo	Idade	Peso	Compr. Cabeça	Compr. Skul	Dente	Pé	Cauda
1	Victoria	male	6.5	2.55	94.1	57.6	24.7	36.5	36.0
2	Victoria	male	6.5	2.75	94.0	57.1	24.5	36.0	36.5
3	Victoria	male	4.5	2.45	93.0	56.3	24.2	35.5	35.5
4	Victoria	male	4.0	2.55	92.5	56.0	24.0	35.3	36.0
5	Victoria	male	3.5	2.40	91.0	55.5	23.7	35.0	35.0
6	Victoria	male	3.0	2.30	90.5	55.0	23.5	34.8	34.5
7	Victoria	male	5.5	2.60	94.3	57.2	24.6	36.3	36.2
8	Victoria	male	5.0	2.70	95.0	57.8	25.0	36.7	37.0
9	Victoria	male	4.0	2.50	92.0	56.0	24.0	35.2	35.5
10	Victoria	male	6.0	2.80	95.5	58.0	25.2	37.0	37.5

Figura 3 – Dado de gambá

Fonte: Dados de pesquisa

A figura 1 resume as médias de algumas variáveis importantes do conjunto de dados como a idade média, peso, comprimento da cabeça e tudo relacionado aos gambás.

Idade média: Os gambás da amostra têm, em média, 4,6 anos. A idade influencia o desenvolvimento físico, então é útil para comparações.

Peso médio: Os gambás pesam em média 2,78 kg. Essa métrica pode ser usada, por exemplo, para observar se há diferenças de peso entre sexos ou regiões.

Comprimento cabeça: Mede a distância da ponta do focinho até o final do corpo (sem contar a cauda). É uma das medidas mais utilizadas para comparar crescimento.

Comprimento do crânio: importante em análises taxonômicas e para distinguir espécies ou gêneros.

Comprimento Dente: Pode estar relacionado ao comprimento da mandíbula ou de dentes importantes (como caninos), útil em estudos alimentares ou sexuais.

Comprimento do pé: Mede o comprimento do pé traseiro. Essa medida pode variar conforme o habitat (ex: gambás que vivem mais em árvores tendem a ter Pés maiores).

Comprimento da cauda: A cauda é importante para equilíbrio e locomoção, especialmente em ambientes arborizados.

A figura 2 mostra os 10 primeiros registros da base de dados original que contém os dados coletados.

ID: Número único para identificar cada animal no banco de dados.

Site: Região de coleta do animal, neste caso, todos os dados vêm de “Victoria”, mas o dataset completo também tem amostras de “New South Wales”. Isso permite comparar populações de lugares diferentes.

Sexo: Sexo do gambá: “male” (masculino). O dataset tem tanto machos quanto fêmea, esse campo é essencial em análises comparativas.

Idade: Idade estimada do animal. Pode ser contínua (ex: 4.5 anos) e está relacionada com as outras medidas morfológicas.

Peso: Peso em quilogramas.

Comprimento cabeça: Comprimento da cabeça e corpo, em centímetros.

Comprimento do crânio: Tamanho do crânio. Importante para estudar maturidade e diferenciação entre sexos.

Dente: Pode representar o comprimento da mandíbula/dente, refletindo dieta e estágio de desenvolvimento.

Pé: Tamanho do pé traseiro. Ajuda a entender hábitos de locomoção.

Cauda: Comprimento da cauda. Pode variar bastante e ter valor classificatório.

Todos os indivíduos mostrados na tabela de amostra são **machos de Victoria**, com idade entre **3 e 6,5 anos**, e os valores seguem uma tendência esperada: quanto mais velho o animal, geralmente, maiores são suas medidas corporais.

Esses dados servem para visualizar a estrutura bruta da base, identificando padrões iniciais e ajudar em análise comparativas como diferenças entre machos e fêmeas.

Comprimento da cabeça

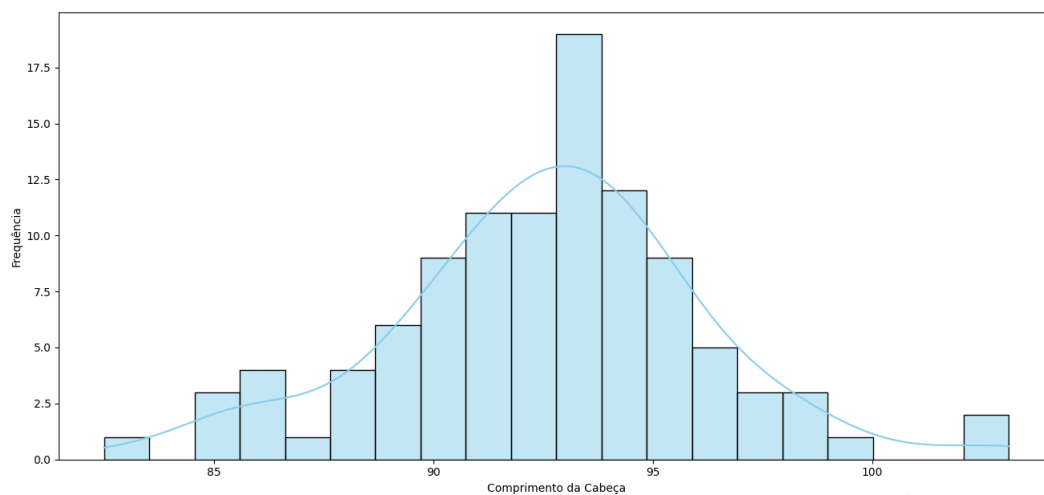


Gráfico 2 – Distribuição de comprimento da cabeça

Fonte: Dados de pesquisa

Na distribuição do comprimento da cabeça o gráfico apresenta um histograma, com a distribuição do comprimento da cabeça dos indivíduos na amostra, indicando o comprimento da cabeça e a frequência. A distribuição tem uma curva para a direita, significando que a maioria dos gambás possuem comprimentos da cabeça próximos a media, mas existem alguns casos com valores mais altos.

Gênero dos gambás

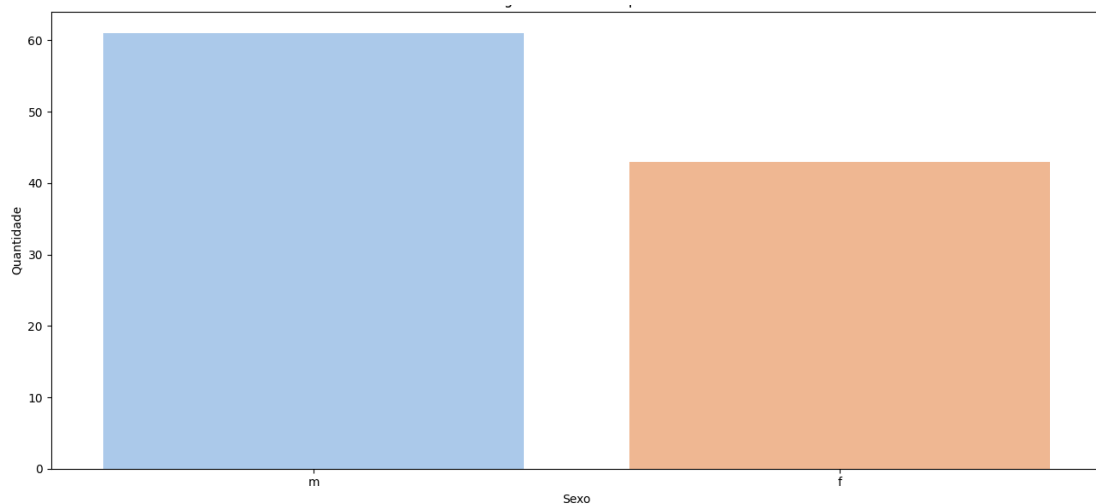


Gráfico 2 – Contagem de gambás por sexo
Fonte: Dados de pesquisa

O gráfico 2 é mostrado em barras demonstrando a distribuição de amostras de acordo com o sexo do gambá. Indicando a barra azul (M) como masculino e a barra laranja (F) como fêmea, também a quantidade de cada sexo.

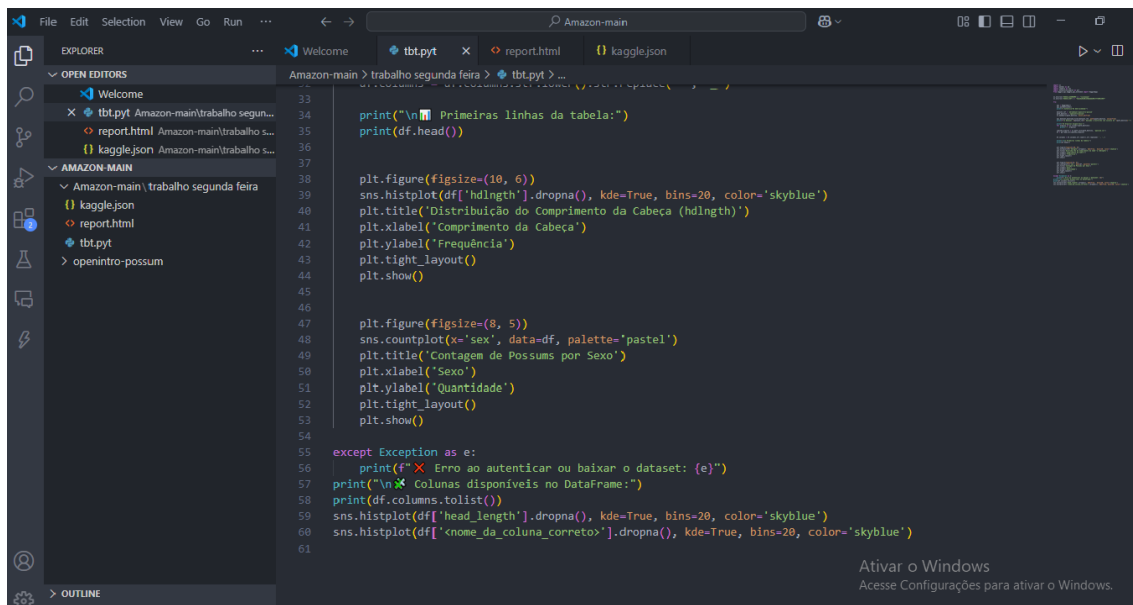
É possível observar que a amostra é composta por uma quantidade maior de indivíduos do sexo masculino, aproximadamente 61 em comparação ao feminino, aproximadamente 43, essa diferença pode influenciar análises posteriores, especialmente se variáveis como o comprimento da cabeça estiverem relacionadas ao sexo.

```

1  import os
2  import pandas as pd
3  import seaborn as sns
4  import matplotlib.pyplot as plt
5  from kaggle.api.kaggle_api_extended import KaggleApi
6
7
8  os.environ['KAGGLE_USERNAME'] = "vinikawa"
9  os.environ['KAGGLE_KEY'] = "4ba26a0180c25ddede09cf7b9013057"
10
11
12  try:
13      api = KaggleApi()
14      api.authenticate()
15      print("✅ Autenticação bem-sucedida!")
16
17      dataset_ref = 'abrambeyer/openintro-possum'
18      path_destino = 'openintro-possum'
19      os.makedirs(path_destino, exist_ok=True)
20
21      api.dataset_download_files(dataset_ref, path= path_destino, unzip=True)
22      print(f"📁 Dataset '{dataset_ref}' baixado e extraído com sucesso em '{path_destino}'!")
23
24      print("📁 Arquivos disponíveis:")
25      for arquivo in os.listdir(path_destino):
26          print("-", arquivo)
27
28      caminho_arquivo = os.path.join(path_destino, 'possum.csv')
29      df = pd.read_csv(caminho_arquivo)
30
31      df.columns = df.columns.str.lower().str.replace(' ', '_')
32
33

```

Figura 4 – Código em python
Fonte: Vscode



```
33
34
35 print("\n Primeiras linhas da tabela:")
36 print(df.head())
37
38
39 plt.figure(figsize=(10, 6))
40 sns.histplot(df['hdlngh'].dropna(), kde=True, bins=20, color='skyblue')
41 plt.title('Distribuição do Comprimento da Cabeça (hdlngh)')
42 plt.xlabel('Comprimento da Cabeça')
43 plt.ylabel('Frequência')
44 plt.tight_layout()
45 plt.show()
46
47
48 plt.figure(figsize=(8, 5))
49 sns.countplot(x='sex', data=df, palette='pastel')
50 plt.title('Contagem de Possums por Sexo')
51 plt.xlabel('Sexo')
52 plt.ylabel('Quantidade')
53 plt.tight_layout()
54 plt.show()
55
56 except Exception as e:
57     print(f"✗ Erro ao autenticar ou baixar o dataset: {e}")
58 print("\n ✗ Colunas disponíveis no DataFrame:")
59 print(df.columns.tolist())
60 sns.histplot(df[['head_length']].dropna(), kde=True, bins=20, color='skyblue')
61 sns.histplot(df[['nome_da_coluna_correto']].dropna(), kde=True, bins=20, color='skyblue')
```

Figura 5 – Código em python
Fonte: Vscode

A figura 4 e 5 são do código do dataset openintro-possum do kaggle, autentica com as credenciais do usuário e salva os arquivos em uma pasta local. Em seguida lê o arquivo csv em um dataframe do pandas, padroniza os nomes das colunas e exibe as primeiras linhas da tabela. Depois, gera os gráficos 1 e 2. Por fim lista todas as colunas disponíveis no dataframe.

```
1 import os
2 import pandas as pd
3 import seaborn as sns
4 import matplotlib.pyplot as plt
5 from kaggle.api.kaggle_api_extended import KaggleApi
```

Figura 6 – Código em python
Fonte: Vscode

- Linha 1 (IMPORT OS): importa o módulo “os”, que fornece funções para interagir com o sistema operacional, como manipular diretórios, arquivos, variáveis de ambiente.
- Linha 2 (IMPORT PANDAS AS PD): usado para manipular e analisar dados em tabela.
- Linha 3 (IMPORT SEABORN AS SNS): Biblioteca para visualização de dados baseada no Matplotlib, mas com visual mais bonito e integração com pandas. Ideal para criar gráficos estatísticos como histogramas

Linha 4 (IMPORT MATPLOTLIB.PYPLT AS PLT): Biblioteca para criação de gráficos em geral. seaborn usa matplotlib por baixo dos panos. Usada para ajustar detalhes de gráficos e exibir visualizações.

Linha 5 (FROM KAGGLE.API.KAGGLE_API_EXTEND IMPORT KAGGLEAPI): Importa a classe KaggleApi da biblioteca oficial da Kaggle. Ela permite interagir com o site do Kaggle por meio da API, como baixar datasets, participar de competições, enviar notebooks, etc.

Essa parte prepara o ambiente para trabalhar com dados, criar gráficos, baixar datasets diretamente do kaggle e interagir com o SO.

```
8 os.environ['KAGGLE_USERNAME'] = "vinikawwa"
9 os.environ['KAGGLE_KEY'] = "4ba26a0180c25ddedde09cf7b9013057"
```

Figura 7 – Código em python
Fonte: Vscod

Linha 8 (OS.ENVIRON[KAGGLE_USERNAME] = "VINIKAWWA"): Aqui, você está dizendo ao Python para definir a variável de ambiente KAGGLE_USERNAME com o valor "vinikawwa", que é o meu nome de usuário no Kaggle

Linha 9 (OS.ENVIRON[KAGGLE_KEY] = "4ba26a0180c25ddedde09cf7b9013057"): Essa linha define a variável de ambiente KAGGLE_KEY com a sua chave secreta da API do Kaggle. Essa chave funciona como uma "senha" para que você possa acessar e baixar datasets do Kaggle via código, sem precisar fazer login manualmente.

```

11 try:
12
13     api = KaggleApi()
14     api.authenticate()
15     print("✅ Autenticação bem-sucedida!")
16
17     dataset_ref = 'abrambeyer/openintro-possum'
18     path_destino = 'openintro-possum'
19     os.makedirs(path_destino, exist_ok=True)
20
21     api.dataset_download_files(dataset_ref, path=path_destino, unzip=True)
22     print(f"📁 Dataset '{dataset_ref}' baixado e extraído com sucesso em '{path_destino}'!")
23
24     print("📁 Arquivos disponíveis:")
25     for arquivo in os.listdir(path_destino):
26         print("-", arquivo)
27
28     caminho_arquivo = os.path.join(path_destino, 'possum.csv')
29     df = pd.read_csv(caminho_arquivo)
30
31
32     df.columns = df.columns.str.lower().str.replace(' ', '_')
33
34     print("\n📄 Primeiras linhas da tabela:")
35     print(df.head())
36
37
38     plt.figure(figsize=(10, 6))
39     sns.histplot(df['hdlngh'].dropna(), kde=True, bins=20, color='skyblue')
40     plt.title('Distribuição do Comprimento da Cabeça (hdlngh)')
41     plt.xlabel('Comprimento da Cabeça')

```

Ativar o W
Acesse Config

Figura 8 – Código em python
Fonte: Vscod

```

42     plt.ylabel('Frequência')
43     plt.tight_layout()
44     plt.show()
45
46
47     plt.figure(figsize=(8, 5))
48     sns.countplot(x='sex', data=df, palette='pastel')
49     plt.title('Contagem de Possums por Sexo')
50     plt.xlabel('Sexo')
51     plt.ylabel('Quantidade')
52     plt.tight_layout()
53     plt.show()
54
55 except Exception as e:
56     print(f"❌ Erro ao autenticar ou baixar o dataset: {e}")
57     print("\n❌ Colunas disponíveis no DataFrame:")
58     print(df.columns.tolist())
59     sns.histplot(df['head_length'].dropna(), kde=True, bins=20, color='skyblue')
60     sns.histplot(df['<nome_da_coluna_correto>'].dropna(), kde=True, bins=20, color='skyblue')
61

```

Figura 9 – Código em python
Fonte: Vscod

Linha 11 (TRY): inicia um bloco de código que será testado para erros. Se ocorrer algum erro dentro desse bloco, ele pode ser tratado com um except (que não está presente aqui, mas pode ser adicionado).

Linha 13 (`api = KaggleApi()`): cria uma instância da classe `KaggleApi`, que permite interagir com a API do Kaggle para baixar datasets, competições, kernels, etc.

Linha 14 (`api.authenticate()`): autentica a sessão com o Kaggle usando as credenciais do usuário, geralmente armazenadas no arquivo `~/.kaggle/kaggle.json`.

Linha 15 (`print(" Autenticação bem-sucedida!")`): exibe uma mensagem de sucesso se a autenticação for concluída sem erros.

Linha 17 (`dataset_ref = 'abrambeyer/openintro-possum'`): define a referência (ID) do dataset que será baixado do Kaggle.

Linha 18 (`path_destino = 'openintro-possum'`): define o diretório local onde o dataset será salvo.

Linha 19 (`os.makedirs(path_destino, exist_ok=True)`): cria o diretório de destino, se ele ainda não existir. O parâmetro `exist_ok=True` evita erro caso a pasta já exista.

Linha 21 (`api.dataset_download_files(...)`): baixa o dataset especificado em `dataset_ref` e o extrai automaticamente no caminho `path_destino`.

Linha 22 (`print(f"...")`): imprime uma mensagem confirmando que o download e a extração do dataset foram feitos com sucesso.

Linha 24 (`print(" Arquivos disponíveis:")`): imprime um cabeçalho indicando que a listagem de arquivos começará.

Linha 25 (`for arquivo in os.listdir(path_destino)`): percorre todos os arquivos no diretório onde o dataset foi extraído.

Linha 26 (`print("-", arquivo)`): imprime o nome de cada arquivo encontrado no diretório do dataset.

Linha 28 (`caminho_arquivo = os.path.join(path_destino, 'possum.csv')`): monta o caminho completo do arquivo `possum.csv`, que é o principal dataset que será carregado.

Linha 29 (`df = pd.read_csv(caminho_arquivo)`): lê o arquivo CSV e armazena os dados em um `DataFrame` do pandas para análise.

Linha 32 (`df.columns = ...`): padroniza os nomes das colunas, deixando tudo em letras minúsculas e substituindo espaços por underline (`_`), para facilitar a manipulação posterior.

Linha 34 (`print("Primeiras linhas da tabela:")`): imprime uma mensagem indicando que as primeiras linhas da tabela serão exibidas.

Linha 35 (`print(df.head())`): exibe as cinco primeiras linhas do DataFrame, dando uma visão geral do conteúdo da tabela.

```
38 plt.figure(figsize=(10, 6))
39 sns.histplot(df['hdlngh'].dropna(), kde=True, bins=20, color='skyblue')
40 plt.title('Distribuição do Comprimento da Cabeça (hdlngh)')
41 plt.xlabel('Comprimento da Cabeça')
42 plt.ylabel('Frequência')
43 plt.tight_layout()
44 plt.show()
45
46
47 plt.figure(figsize=(8, 5))
48 sns.countplot(x='sex', data=df, palette='pastel')
49 plt.title('Contagem de Possums por Sexo')
50 plt.xlabel('Sexo')
51 plt.ylabel('Quantidade')
52 plt.tight_layout()
53 plt.show()
54
```

Figura 10 – Código em python
Fonte: Vscode

Linha 38 (`plt.figure(figsize= (10, 6))`): cria uma nova figura para o gráfico, definindo o tamanho como 10 de largura e 6 de altura. Isso ajuda a melhorar a visualização.

Linha 39 (`sns.histplot(...)`): cria um histograma da coluna `hdlngh` (comprimento da cabeça).

Linha 40 (`plt.title(...)`): define o título do gráfico como "Distribuição do Comprimento da Cabeça (`hdlngh`)".

Linha 41 (`plt.xlabel(...)`): define o rótulo do eixo X como "Comprimento da Cabeça".

Linha 42 (`plt.ylabel(...)`): define o rótulo do eixo Y como "Frequência", ou seja, quantas vezes cada valor aparece.

Linha 43 (`plt.tight_layout()`): ajusta automaticamente o layout para evitar sobreposição de textos e eixos.

Linha 44 (`plt.show()`): exibe o gráfico na tela.

Linha 47 (`plt.figure(figsize=(8, 5))`): cria uma nova figura para o próximo gráfico, agora com tamanho 8x5.

Linha 48 (`sns.countplot(...)`): cria um gráfico de barras mostrando a quantidade de registros de possums por sexo.

Linha 49 (plt.title(...)): define o título do gráfico como "Contagem de Possums por Sexo".

Linha 50 (plt.xlabel(...)): define o rótulo do eixo X como "Sexo".

Linha 51 (plt.ylabel(...)): define o rótulo do eixo Y como "Quantidade", ou seja, quantos possums de cada sexo existem no conjunto de dados.

Linha 52 (plt.tight_layout()): ajusta automaticamente o layout do gráfico.

Linha 53 (plt.show()): exibe o segundo gráfico na tela.

```
55 except Exception as e:
56     print(f"❌ Erro ao autenticar ou baixar o dataset: {e}")
57     print("\n❌ Colunas disponíveis no DataFrame:")
58     print(df.columns.tolist())
59     sns.histplot(df['head_length'].dropna(), kde=True, bins=20, color='skyblue')
60     sns.histplot(df['<nome_da_coluna_correto>'].dropna(), kde=True, bins=20, color='skyblue')
```

Figura 11 – Código em python
Fonte: Vscode

Linha 55 (except Exception as e): inicia o bloco que captura erros ocorridos dentro do try.

Linha 56 (print(f" Erro ao autenticar ou baixar o dataset: {e}")): mostra uma mensagem de erro amigável no console, exibindo também o motivo específico do erro capturado na variável e.a o bloco que captura erros ocorridos dentro do try.

Linha 57 (print("\n❌ Colunas disponíveis no DataFrame:")): exibe um título informando que a lista de colunas do DataFrame será apresentada em seguida.

Linha 58 (print(df.columns.tolist())): imprime a lista completa de colunas do DataFrame como uma lista Python, útil para conferir os nomes disponíveis para análise.

Linha 59 (sns.histplot(df['head_length'].dropna(), ...)): tenta criar um histograma da coluna head_length, removendo valores nulos.

Linha 60 (sns.histplot(df['<nome_da_coluna_correto>'].dropna(), ...)): exemplo genérico de outro histograma, mas com o nome da coluna ainda a ser definido (<nome_da_coluna_correto>).

CONCLUSÃO

Com essa data set permite compreender os padrões nas características dos gambás, como comprimento da cabeça, peso, idade e diferenças entre os sexos. Utilizando o python, pandas, seaborn foi possível organizar, visualizar e interpretar essas informações.

Os gráficos revelam uma predominância dos gambás do sexo masculino na amostra e a assimetria no comprimento da cabeça, demonstrando que a maioria dos gambas possui medidas dentro de uma faixa específica. Essas informações são relevantes para estudos sobre dimorfismo sexual e variações regionais na espécie

REFERÊNCIAS

BEYER, Abram. Openintro Possum. Kaggle, 2022. Disponível em: <https://www.kaggle.com/datasets/abrambeyer/openintro-possum>. Acesso em: 20 abr. 2025

TÍTULO DO TRABALHO EM LINGUA EwSTRANGEIRA

POSSUM REGRESSION

RESUMO

Predict an opossum's age, head length and whether it is male or female using regression model using data from the openintro-possum dataset.

PALAVRAS-CHAVE

Data analysis; Possum; Regression