DAYANANDA SAGAR UNIVERSITY

SCHOOL OF ENGINEERING

HAROHALLI, KANAKAPURA ROAD – 562112

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
(DATA SCIENCE)

STATISTICAL FOUNDATIONS OF DATA SCIENCE

PROJECT REPORT

ON

“EDA ON PALMER PENGUINS DATASET”

2024-2025

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE & ENGINEERING (DATA SCIENCE)

Submitted by

Kumari Nainshi-ENG23DS0016

Anuja Suresh Shinde-ENG23DS0056

Under The Supervision of:

Prof. Sindhu A

Assistant Professor

Department of CSE (Data Science), DSU

DAYANANDA SAGAR UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

(DATA SCIENCE)

SCHOOL OF ENGINEERING

HAROHALLI, KANAKAPURA ROAD – 562112

CERTIFICATE

It is certified that the mini project work entitled "EDA ON PALMER PENGUINS DATASET" has been carried out at *Dayananda Sagar University*, Bangalore, by Kumari Nainshi-ENG23DS0016 ; Anuja Suresh Shinde-ENG23DS0056 Bonafide student of fourth Semester, B.Tech in partial fulfilment for the award of degree in *Bachelor of Technology in Computer Science & Engineering (Data Science)* during academic year *2024-25*. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the report deposited in departmental library.

The project report has been approved as it satisfies the academic requirements in respect of project work for the said degree.

Signature of the Guide **Signature of the Chairperson**

ACKNOWLEDGEMENT

A project's successful completion offers a sense of satisfaction, but it is never finished without expressing gratitude to everyone who contributed to its accomplishment. We would like to convey our sincere gratitude to our esteemed university, Dayananda Sagar University, for offering the first-rate facilities.

We are especially thankful to our Chairperson, **Dr. Shaila S G**, for providing necessary departmental facilities, moral support and encouragement. The largest measure of our acknowledgment is reserved for **Prof. Sindhu A,** whose guidance and support madeit possible to complete the project work in a timely manner.

We would want to thank everyone who has assisted us in successfully completing this project work, both directly and indirectly. The staff has provided us with a great deal of direction and cooperation.

Kumari Nainshi-ENG23DS0016
Anuja Suresh Shinde-ENG23DS0056

DECLARATION

We hereby declare that the project entitled **" EDA ON PALMER PENGUINS DATASET"** submitted to Dayananda Sagar University, Bengaluru, is a bonafide record of the work carried out by me under the guidance of **Prof. Sindhu A**, Assistant Professor in the Dayananda Sagar University School of Engineering's Department of Computer Science and Engineering (Data Science). This work is submitted toward the partial fulfillment of the requirements for the award of a Bachelor of Technology in Computer Science and Engineering (Data Science).

Kumari Nainshi-ENG23DS0016
Anuja Suresh Shinde-ENG23DS0056

ABSTRACT

The Palmer Penguins dataset provides a rich, multivariate collection of biological measurements for three penguin species observed in the Palmer Archipelago, Antarctica. This study conducts an extensive Exploratory Data Analysis (EDA) to uncover patterns, relationships, and insights across key numerical and categorical features.

The analysis begins with data cleaning, where missing values are identified, visualized, and handled using both deletion and imputation strategies. The dataset is then examined for data types, distributions, and potential anomalies. Histograms, bar charts, and boxplots are employed to visualize the spread and central tendencies of numerical variables such as bill length, bill depth, flipper length, and body mass across different species.

Categorical variables, including species, island, and sex, are analyzed using count plots to understand distributional imbalances. Correlation analysis is conducted among numeric features to identify strong linear relationships, which are visualized using heatmaps for better interpretability.

Furthermore, scatter plots and pairwise visualizations are used to explore inter-variable interactions, especially between morphological traits and species classification. Grouped statistics such as mean and standard deviation are calculated to observe interspecies differences in physical traits.

This EDA not only highlights the distinct characteristics of each species but also sets the foundation for further statistical modeling and machine learning tasks, offering valuable insights into ecological and morphological diversity within the penguin population.

TABLE OF CONTENTS

INTRODUCTION

Exploratory Data Analysis (EDA) is a fundamental step in the data science pipeline, aimed at understanding the structure, quality, and underlying patterns in a dataset before formal modeling begins. This study focuses on the Palmer Penguins dataset, a modern alternative to the widely used Iris dataset, offering a more nuanced and biologically meaningful set of features.

The dataset comprises morphological and demographic data for three species of penguins—Adélie, Chinstrap, and Gentoo—collected from three islands in the Palmer Archipelago, Antarctica. It includes both numerical variables (e.g., bill length, flipper length, body mass) and categorical attributes (e.g., species, island, sex), making it suitable for multivariate analysis.

The primary objective of this analysis is to gain insights into the characteristics of penguins across different species and locations through a detailed EDA. This includes identifying missing values, analyzing distributions, uncovering relationships between features, and visualizing the data using a variety of plots such as histograms, boxplots, and heatmaps.

By thoroughly exploring the dataset, we aim to uncover patterns that can inform ecological understanding, support classification tasks, and highlight the importance of data quality and structure in biological research.

OBJECTIVE AND SCOPE OF WORK

Objectives:

The primary objective of this study is to perform Exploratory Data Analysis (EDA) on the Palmer Penguins dataset to uncover patterns, identify relationships among features, and understand species-level distinctions in penguin morphology.

Scope of the Work:

- Handle and visualize missing data for data quality assessment
- Analyze the distribution of numerical and categorical variables
- Explore inter-variable relationships using correlation and scatter plots
- Compare species-specific characteristics through grouped statistics and boxplots
- Generate meaningful visualizations to support further predictive modeling or classification

DESCRIPTION OF WORK

The analysis includes the following steps:

1. **Data Loading and Inspection:**
   The dataset was loaded into a pandas DataFrame and initially examined to understand its structure, data types, and column values.

2. **Missing Value Analysis:**
   A detailed inspection was conducted to identify missing values across all columns. These were visualized using heatmaps and bar plots, and handled by either removing incomplete rows or applying suitable imputation techniques.

3. **Univariate Analysis:**
   Histograms and bar charts were used to examine the distribution of individual numerical variables (e.g., bill length, body mass) and categorical variables (e.g., species, island, sex).

4. **Bivariate and Grouped Analysis:**
   Boxplots and grouped mean plots were used to compare numerical features across different species. Relationships between pairs of variables were also explored using scatter plots.

5. **Correlation Analysis:**

   A correlation matrix was computed for numeric features to quantify linear relationships. The matrix was visualized using heatmaps with annotations to highlight strong positive or negative correlations.

6. **Data Visualization:**

   Multiple visualizations were generated throughout the analysis, including species-wise histograms, boxplots, pairwise scatter plots, and customized heatmaps to support deeper insight.

PROJECT MODEL DESCRIPTION

This project uses a systematic Exploratory Data Analysis (EDA) approach to understand the structure and patterns within the Palmer Penguins dataset. The model begins with data loading and preprocessing, where missing values are identified and either removed or imputed, and data types are standardized for analysis.

Next, univariate analysis is performed using histograms and bar plots to study the distribution of individual features. Multivariate analysis follows, using boxplots and scatter plots to explore relationships between features such as bill length, flipper length, and species.

A correlation matrix is generated to identify linear relationships between numeric variables, which is visualized using heatmaps for better clarity. Grouped summaries and visual comparisons are also made to highlight species-level differences in morphological traits.

The model emphasizes clear visualization and statistical interpretation to extract insights, detect trends, and prepare the dataset for further predictive modeling or ecological research.

REQUIREMENTS AND METHODOLOGY

Requirements:

Software:

- Python (v3.7 or later)
- Jupyter Notebook / Any Python IDE (e.g., VS Code, PyCharm)

Python Libraries Used:

- pandas – for data loading, cleaning, and manipulation
- numpy – for numerical operations
- matplotlib.pyplot – for basic plotting
- seaborn – for advanced and aesthetic data visualization

Methodology:

1. Data Import and Inspection

- The dataset was loaded using pandas.read_csv().
- Initial information such as column types, non-null counts, and sample records were displayed using .info() and .head().

2. Missing Values Handling

- The total and percentage of missing values per column were calculated and printed.
- A heatmap was generated using seaborn to visually assess the pattern of missing data.

- Missing values were handled by dropping incomplete rows using dropna(). Optional bar plots also helped highlight columns with missing values.

3. Univariate and Categorical Analysis

- Categorical features such as island and sex were visualized using bar charts to show frequency distributions.
- This helped assess category balance and potential class imbalances.

4. Numerical Feature Analysis by Species

- Key numeric columns (bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g) were analyzed across species using boxplots.
- This allowed for comparison of central tendency and variability between species, identifying outliers and interspecies differences.

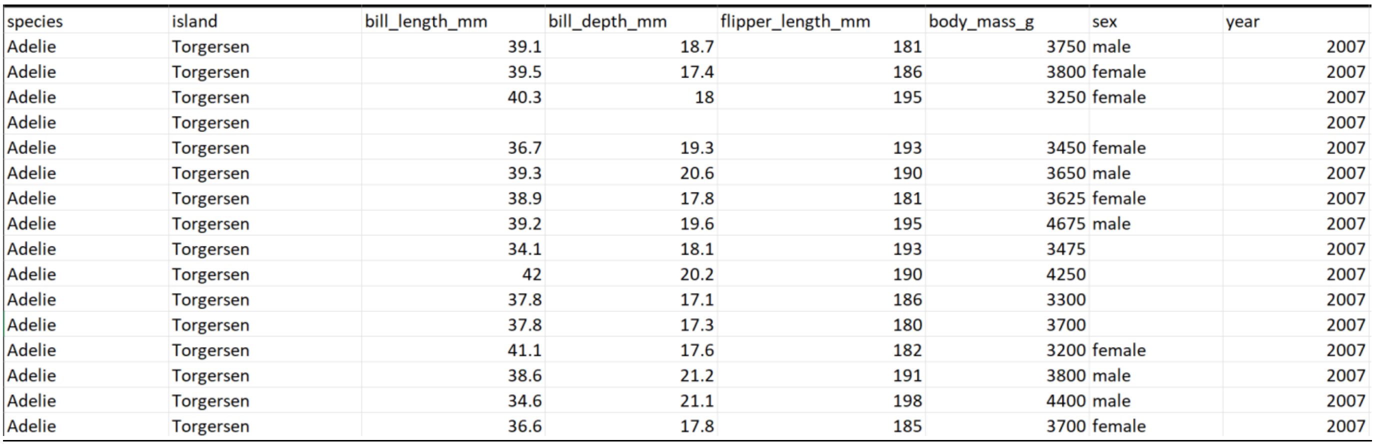
5. Visualization and Interpretation

- Plots were generated with clear titles, labels, and proper formatting using matplotlib and seaborn to aid interpretation.
- Layouts were adjusted with plt.tight_layout() to improve plot readability.

EXPECTED RESULT

1. Identification and removal or treatment of missing values.
2. Understanding the distribution of numerical and categorical variables.
3. Detection of outliers and data inconsistencies.
4. Visualization of feature patterns across penguin species.
5. Discovery of correlations between numeric features.
6. Comparison of species based on morphological traits.
7. Generation of clean, structured data ready for modeling or analysis.
8. Gaining insights to support ecological or predictive studies.

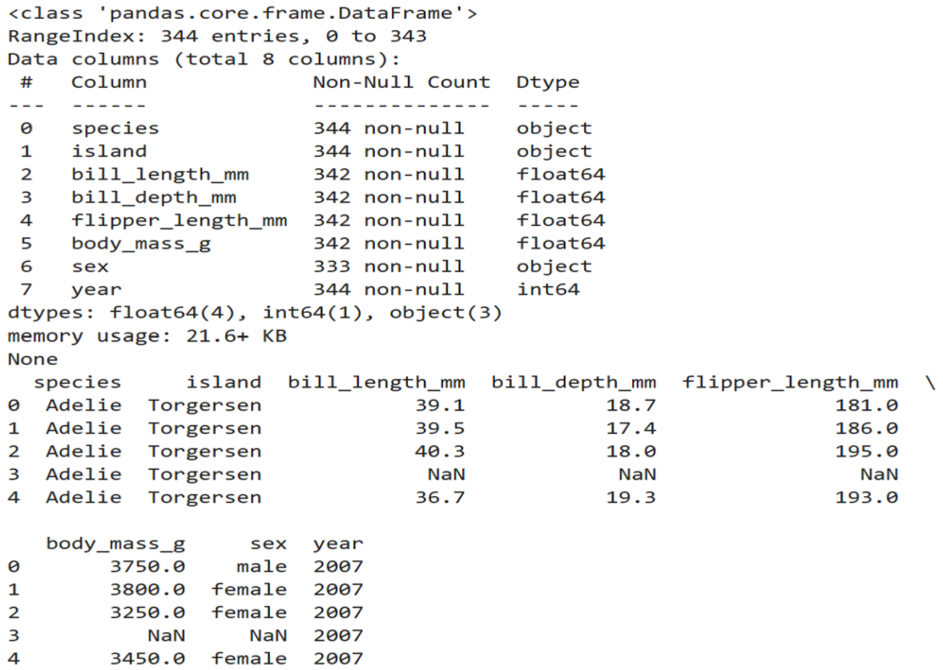
SAMPLE DATA



species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18	195	3250	female	2007
Adelie	Torgersen						2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007
Adelie	Torgersen	38.9	17.8	181	3625	female	2007
Adelie	Torgersen	39.2	19.6	195	4675	male	2007
Adelie	Torgersen	34.1	18.1	193	3475		2007
Adelie	Torgersen	42	20.2	190	4250		2007
Adelie	Torgersen	37.8	17.1	186	3300		2007
Adelie	Torgersen	37.8	17.3	180	3700		2007
Adelie	Torgersen	41.1	17.6	182	3200	female	2007
Adelie	Torgersen	38.6	21.2	191	3800	male	2007
Adelie	Torgersen	34.6	21.1	198	4400	male	2007
Adelie	Torgersen	36.6	17.8	185	3700	female	2007

SOURCE CODE WITH OUTPUT

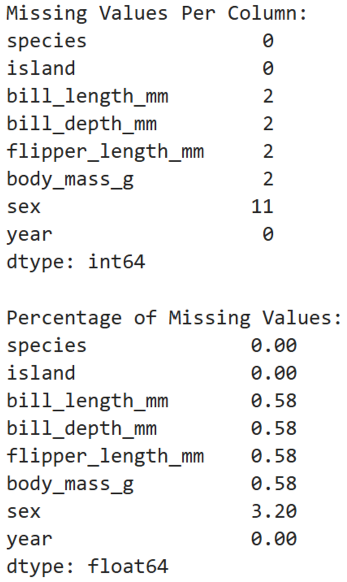
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
sns.set(style='whitegrid')

# Load data
data = pd.read_csv("C:/Users/Anuja Shinde/Desktop/PROJECT
MAT/palmerpenguins_original.csv")

# Display basic info
print(data.info())
print(data.head())

# Handle missing values
data.dropna(inplace=True)
```



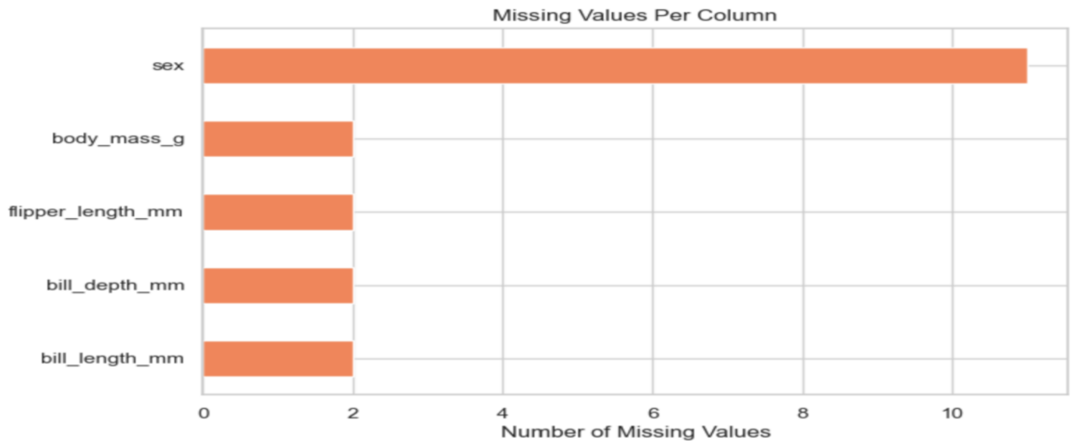
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 344 entries, 0 to 343
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   species            344 non-null    object
 1   island             344 non-null    object
 2   bill_length_mm     342 non-null    float64
 3   bill_depth_mm      342 non-null    float64
 4   flipper_length_mm  342 non-null    float64
 5   body_mass_g        342 non-null    float64
 6   sex                333 non-null    object
 7   year               344 non-null    int64
dtypes: float64(4), int64(1), object(3)
memory usage: 21.6+ KB
None
  species     island  bill_length_mm  bill_depth_mm  flipper_length_mm  \
0  Adelie  Torgersen            39.1           18.7              181.0
1  Adelie  Torgersen            39.5           17.4              186.0
2  Adelie  Torgersen            40.3           18.0              195.0
3  Adelie  Torgersen             NaN            NaN                NaN
4  Adelie  Torgersen            36.7           19.3              193.0

   body_mass_g     sex  year
0       3750.0    male  2007
1       3800.0  female  2007
2       3250.0  female  2007
3          NaN     NaN  2007
4       3450.0  female  2007
```

```
# Display total missing values per column
print("Missing Values Per Column:")
print(data.isnull().sum())

# Display percentage of missing values
print("\nPercentage of Missing Values:")
print((data.isnull().mean() * 100).round(2))
# Visualize missing data as a heatmap
```

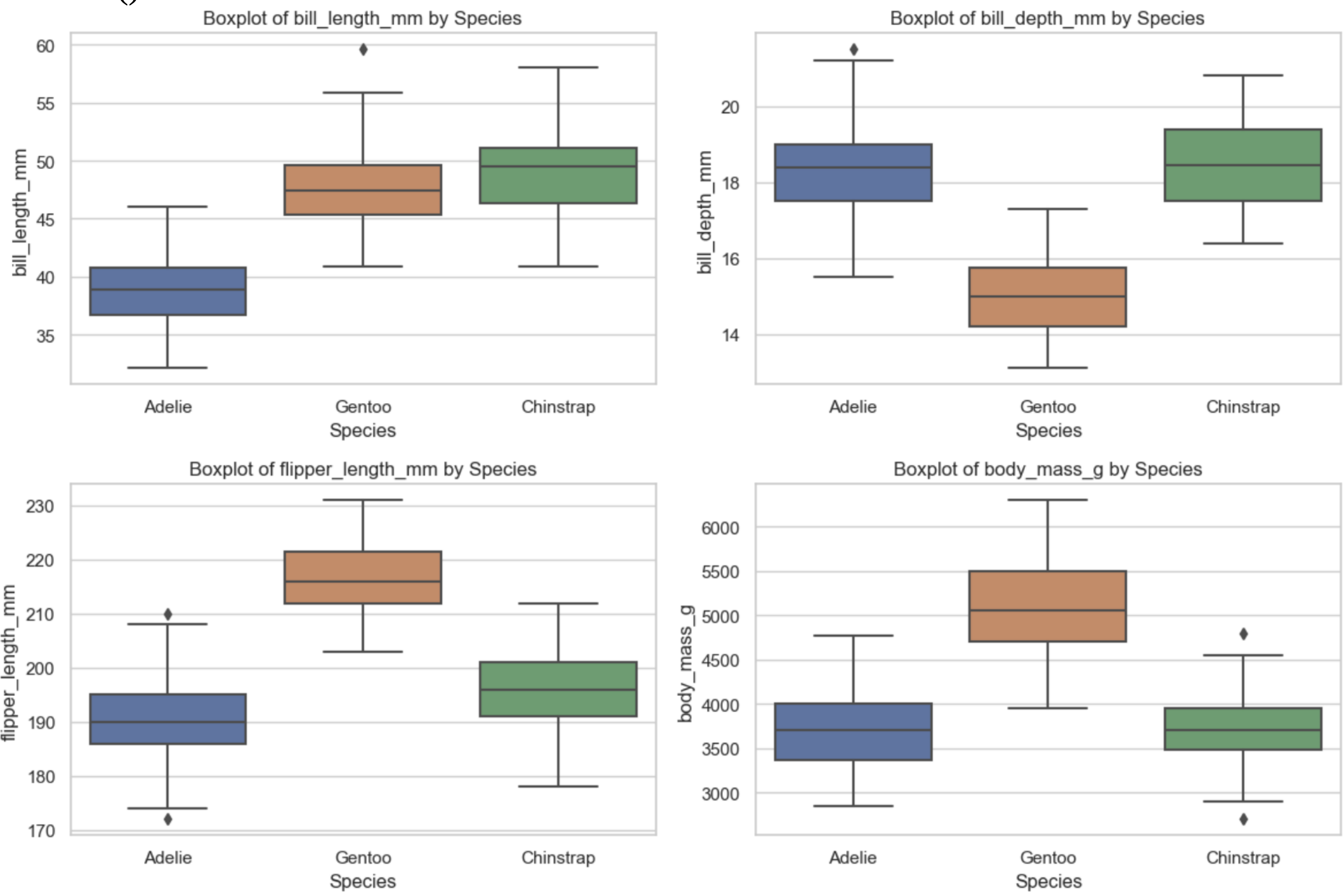
```
plt.figure(figsize=(10, 6))
sns.heatmap(data.isnull(), cbar=False, cmap='viridis', yticklabels=False)
plt.title('Missing Data Heatmap')
plt.show()

# Optionally: visualize using a bar chart
missing_counts = data.isnull().sum()
missing_counts = missing_counts[missing_counts > 0]
missing_counts.sort_values(inplace=True)

plt.figure(figsize=(8, 5))
missing_counts.plot(kind='barh', color='coral')
plt.xlabel('Number of Missing Values')
plt.title('Missing Values Per Column')
plt.show()

# Handle missing values - Option 1: Drop all rows with any missing values
clean_data = data.dropna()
```



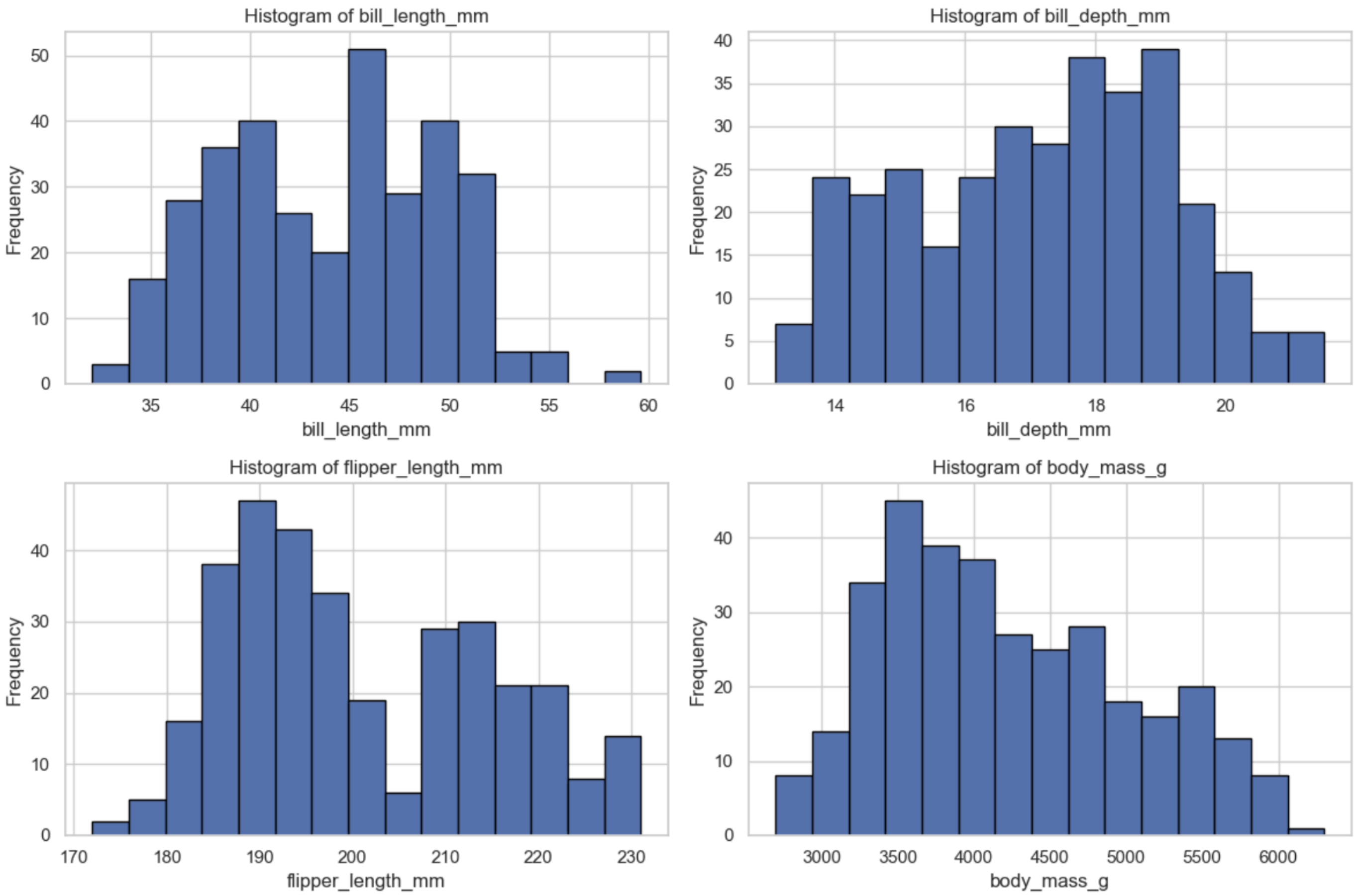
```
Missing Values Per Column:
species              0
island               0
bill_length_mm       2
bill_depth_mm        2
flipper_length_mm    2
body_mass_g          2
sex                 11
year                 0
dtype: int64

Percentage of Missing Values:
species             0.00
island              0.00
bill_length_mm      0.58
bill_depth_mm       0.58
flipper_length_mm   0.58
body_mass_g         0.58
sex                 3.20
year                0.00
dtype: float64
```



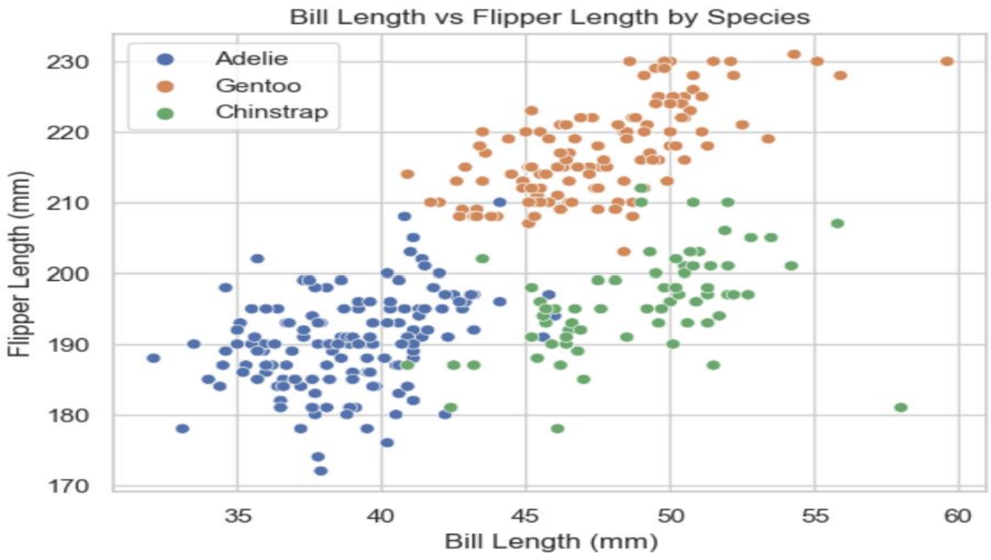
```
num_vars = ['bill_length_mm', 'bill_depth_mm', 'flipper_length_mm', 'body_mass_g']

plt.figure(figsize=(12, 8))
for i, var in enumerate(num_vars):
    plt.subplot(2, 2, i+1)
    sns.boxplot(x='species', y=var, data=data)
    plt.title(f'Boxplot of {var} by Species')
    plt.xlabel('Species')
    plt.ylabel(var)
plt.tight_layout()
plt.show()
```



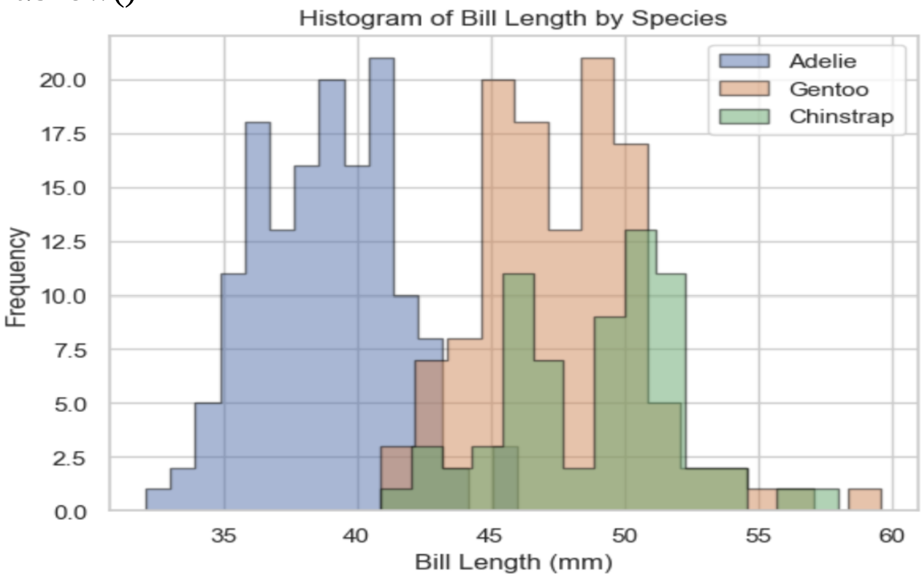
```
plt.figure(figsize=(12, 8))
for i, var in enumerate(num_vars):
    plt.subplot(2, 2, i+1)
    plt.hist(data[var], bins=15, edgecolor='black')
    plt.title(f'Histogram of {var}')
    plt.xlabel(var)
    plt.ylabel('Frequency')
plt.tight_layout()
plt.show()
```



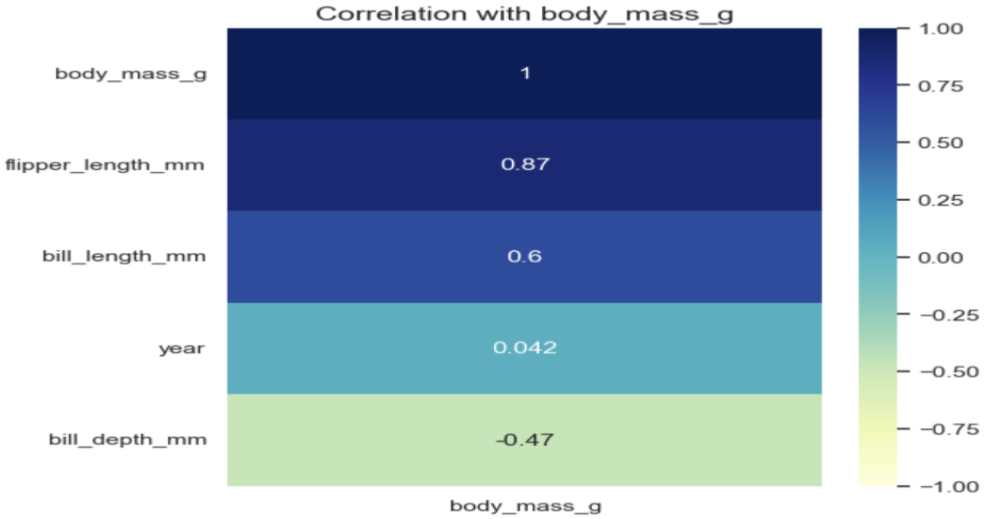
```
plt.figure()
sns.scatterplot(x='bill_length_mm', y='flipper_length_mm', hue='species', data=data)
plt.xlabel('Bill Length (mm)')
plt.ylabel('Flipper Length (mm)')
plt.title('Bill Length vs Flipper Length by Species')
plt.legend()
plt.show()
```



```
plt.figure()
for species in data['species'].unique():
    subset = data[data['species'] == species]
    plt.hist(subset['bill_length_mm'], bins=15, alpha=0.5, label=species, edgecolor='black',
histtype='stepfilled')
plt.xlabel('Bill Length (mm)')
plt.ylabel('Frequency')
plt.title('Histogram of Bill Length by Species')
plt.legend()
plt.show()
```



```
target = 'body_mass_g'
corr = data.select_dtypes(include=['float64', 'int64']).dropna().corr()
sorted_corr = corr[[target]].sort_values(by=target, ascending=False)
# Plot
plt.figure(figsize=(6, 6))
sns.heatmap(sorted_corr, annot=True, cmap='YlGnBu', vmin=-1, vmax=1)
plt.title(f'Correlation with {target}', fontsize=14)
plt.show()
```



CONCLUSION

This project successfully applied Exploratory Data Analysis techniques to the Palmer Penguins dataset to gain meaningful insights into penguin species and their physical characteristics. Through data cleaning, visualization, and statistical exploration, the study:

- Identified and handled missing data effectively.
- Analyzed and visualized the distribution of key numerical features such as bill length, flipper length, and body mass.
- Compared morphological traits across species using boxplots and grouped summaries.
- Uncovered correlations and interdependencies between features.
- Highlighted distinct patterns and differences among Adelie, Chinstrap, and Gentoo species.

REFERENCES

1. Allison Horst, Kristen Gorman, & Palmer Station LTER. (2020). *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. R package version 0.1.0.
2. Wes McKinney. (2010). *Data Structures for Statistical Computing in Python*. Proceedings of the 9th Python in Science Conference. Library used: pandas
3. NumPy Developers. (2024). *NumPy: Fundamental package for scientific computing with Python*.Library used: numpy
4. Python Software Foundation. (2024). *Python Language Reference*.