Final Project - Step 2 (20 Points)

PSTAT100: Data Science Concepts and Analysis

Ali Abuzaid

STUDENT NAME

- Dongzhen Huangfu (3137544)
- Harry He (7133085)
- Xihua Yao (5927777)
- Yue Zhu (A1F8C52)
- Tina Zhou (4876165)

Due Date

The deadline for this step is **November 8, 2024**.

Instructions

The goal of this step is to develop clear research questions and hypotheses based on your selected dataset and to conduct a thorough Exploratory Data Analysis (EDA). This process will set the foundation for your later analysis and insights.

1 Step 2: Research Questions, Hypotheses, and Exploratory Data Analysis (EDA)

1.1 Research Questions

Question 1 What is the correlation between each penguins' traits (body mass, flipper length, culmen length and culmen depth) of all three types of penguins?

Question 2 How do external factors (environment and food resource) and internal factor (gender difference) affect body mass? Which above factor influences it most?

1.2 Hypotheses

Hypothesis 1 Ho: There are no significant differences in body measurement(body mass, flipper length, culmen length and culmen depth)among the 3 penguin species. Ha: There is at least one body measurement that differs significantly among the 3 penguin species.

Hypothesis 2 Both external factors (food resource) and the internal factor (gender difference) significantly affect variations in body mass. External factor have a greater influence on body mass of penguins than internal factor.

1.3 Exploratory Data Analysis (EDA)

1.4 Data Cleaning

We delete rows that contain N/A.

```r
# We delete rows that contain N/A
library(dplyr)
library(palmerpenguins)
data <- penguins_raw %>% select(-Comments)
sorted_data <- na.omit(data)
```

1.5 Descriptive Statistics

```
[1] "Summary for Adelie Penguins (Selected Variables):"
```

```
 Culmen Length (mm) Culmen Depth (mm) Flipper Length (mm) Body Mass (g)
 Min.   :32.10      Min.   :15.50     Min.   :172.0       Min.   :2850
 1st Qu.:36.70      1st Qu.:17.45     1st Qu.:186.0       1st Qu.:3350
 Median :38.80      Median :18.40     Median :190.0       Median :3700
 Mean   :38.79      Mean   :18.32     Mean   :190.3       Mean   :3703
 3rd Qu.:40.65      3rd Qu.:19.00     3rd Qu.:195.0       3rd Qu.:4000
 Max.   :46.00      Max.   :21.50     Max.   :210.0       Max.   :4775
 Delta 15 N (o/oo) Delta 13 C (o/oo)
 Min.   :7.698     Min.   :-26.79
 1st Qu.:8.567     1st Qu.:-26.24
 Median :8.881     Median :-25.99
 Mean   :8.859     Mean   :-25.81
 3rd Qu.:9.166     3rd Qu.:-25.31
```

```
 Max.   :9.795     Max.   :-23.90
```

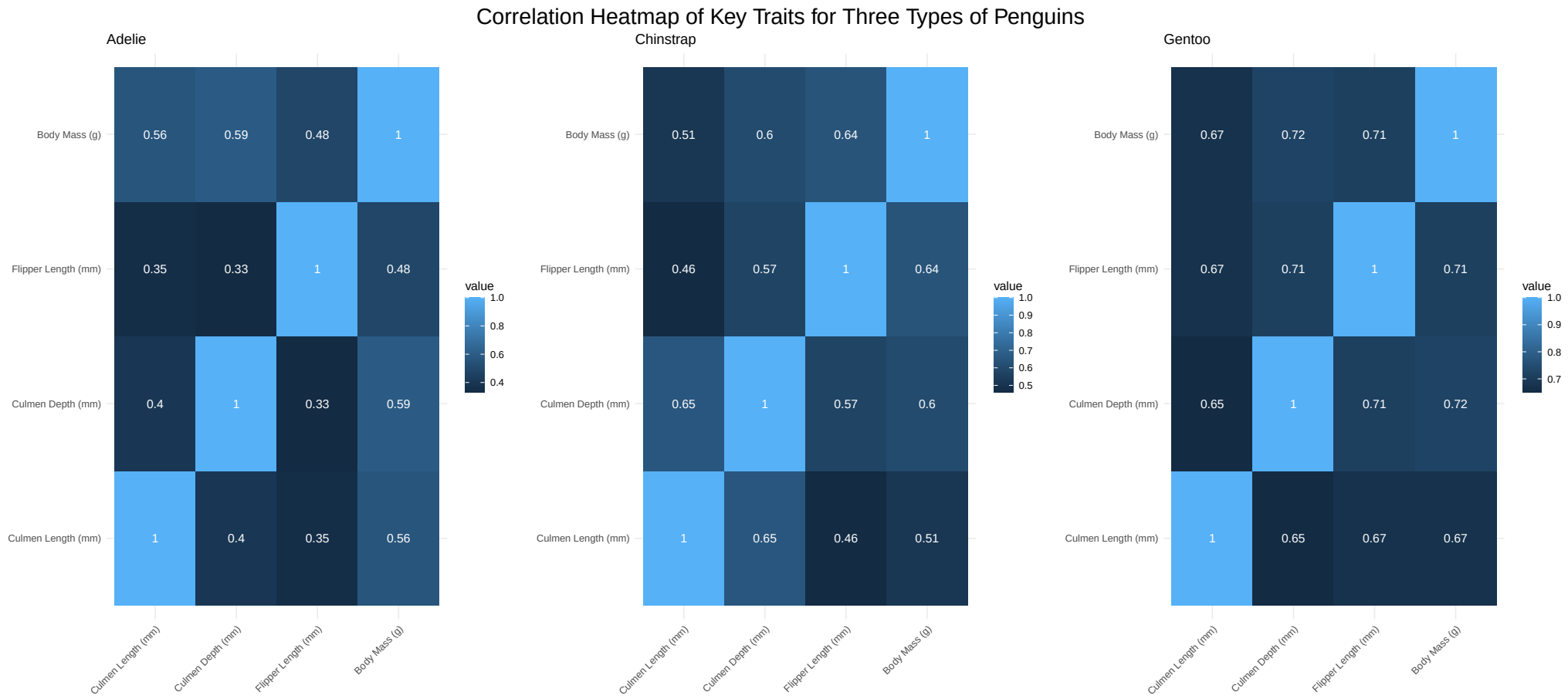
```
[1] "Summary for Chinstrap Penguins (Selected Variables):"
```

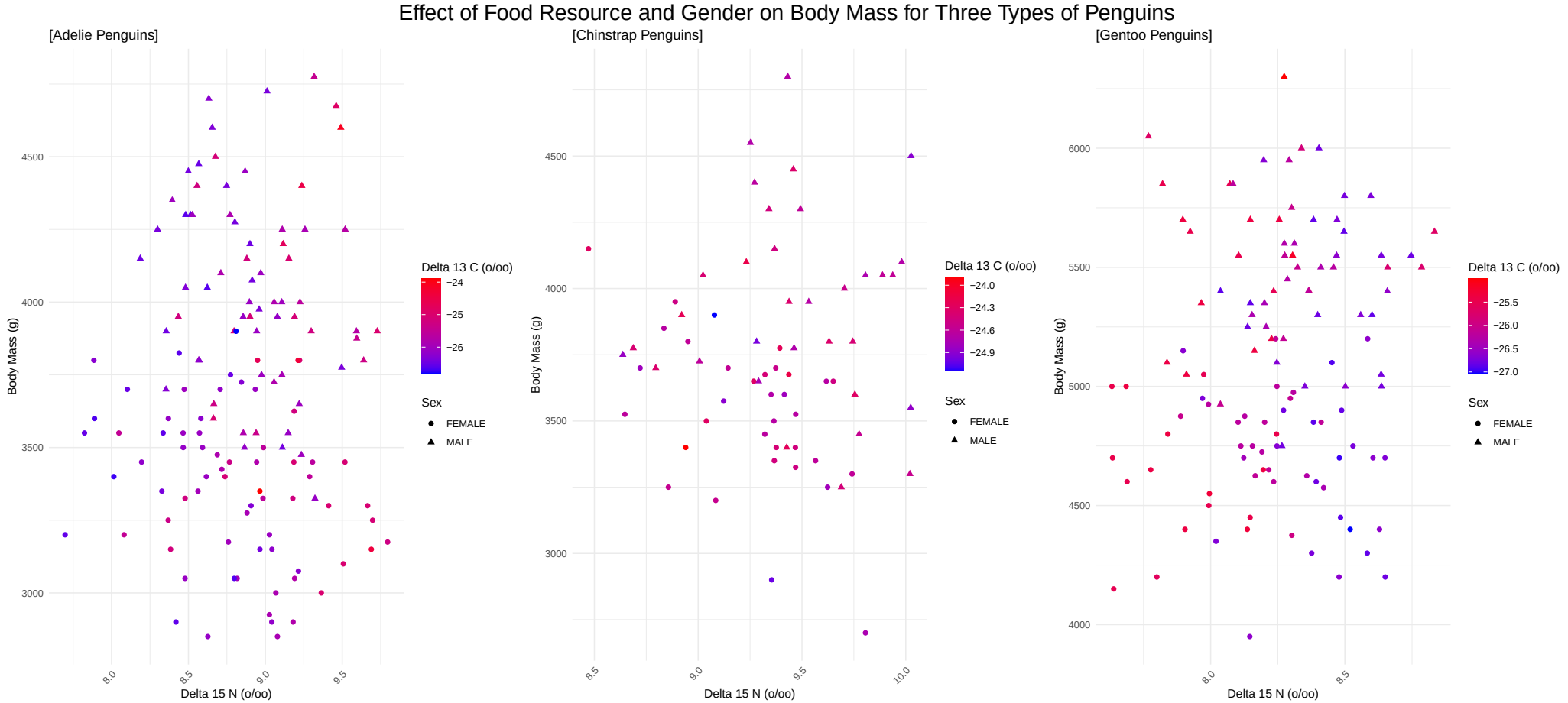
```
 Culmen Length (mm) Culmen Depth (mm) Flipper Length (mm) Body Mass (g)
 Min.   :40.90      Min.   :16.4      Min.   :178.0       Min.   :2700
 1st Qu.:46.30      1st Qu.:17.5      1st Qu.:191.0       1st Qu.:3475
 Median :49.50      Median :18.4      Median :196.0       Median :3700
 Mean   :48.79      Mean   :18.4      Mean   :195.7       Mean   :3730
 3rd Qu.:50.95      3rd Qu.:19.3      3rd Qu.:200.5       3rd Qu.:3950
 Max.   :58.00      Max.   :20.8      Max.   :212.0       Max.   :4800
 Delta 15 N (o/oo) Delta 13 C (o/oo)
 Min.   : 8.472    Min.   :-25.15
 1st Qu.: 9.104    1st Qu.:-24.69
 Median : 9.374    Median :-24.58
 Mean   : 9.356    Mean   :-24.56
 3rd Qu.: 9.620    3rd Qu.:-24.41
 Max.   :10.025    Max.   :-23.89
```

```
[1] "Summary for Gentoo Penguins (Selected Variables):"
```

```
 Culmen Length (mm) Culmen Depth (mm) Flipper Length (mm) Body Mass (g)
 Min.   :40.90      Min.   :13.10     Min.   :203.0       Min.   :3950
 1st Qu.:45.33      1st Qu.:14.20     1st Qu.:212.0       1st Qu.:4700
 Median :47.45      Median :15.00     Median :216.0       Median :5050
 Mean   :47.57      Mean   :14.99     Mean   :217.2       Mean   :5091
 3rd Qu.:49.60      3rd Qu.:15.78     3rd Qu.:221.0       3rd Qu.:5500
 Max.   :59.60      Max.   :17.30     Max.   :231.0       Max.   :6300
 Delta 15 N (o/oo) Delta 13 C (o/oo)
 Min.   :7.632     Min.   :-27.02
 1st Qu.:8.106     1st Qu.:-26.69
 Median :8.260     Median :-26.22
 Mean   :8.249     Mean   :-26.18
 3rd Qu.:8.444     3rd Qu.:-25.56
 Max.   :8.834     Max.   :-25.00
```

1.6 Data Visualization





Initial insights: The penguins dataset provides data on three penguin species—Adelie, Chinstrap, and Gentoo—found in Antarctica. By examining features like body mass and flipper length, we can explore physical differences between species and how they adapt to their environment. Nitrogen and carbon isotope values (Delta 15 N and Delta 13 C) offer clues about each species' diet and feeding areas, showing their roles in the ecosystem. Comparing these three species helps us understand what makes each one unique. This dataset supports insights into biodiversity and species-specific adaptations in their

EDA: First, we will conduct data cleaning by removing any missing values to ensure the accuracy of our analysis and selecting variables relevant to our research questions, such as body mass, culmen length, culmen depth, and flipper length.

Next, we will perform descriptive statistics, filtering data by penguin species (Adelie, Chinstrap, Gentoo) and calculating key statistics (e.g., minimum, maximum, median, mean) for each species to facilitate comparison.

For data visualization, we'll address the first research question by creating correlation heatmaps to examine relationships among penguin traits like body mass and flipper length across species. For the second research question, we'll use scatter plots to illustrate the effects of sex and food resources (using nitrogen and carbon isotopes as proxies) on body mass, using color gradients and shapes to differentiate between sex and isotope concentration.