Penguin Pursuits: Alice's Visual Adventure

Trishita Patra

MDS202440

Abstract:

"Curiouser and curiouser!" In a twist of fate, if Alice ventured into the Palmer Archipelago, she would encounter the fascinating world of penguins. With a dataset at her disposal, she'd embark on a quest to explore the unique identities of three species: Adélie, Chinstrap, and Gentoo penguins. This study examines their physical characteristics, distribution, and the intriguing correlations among their measurements—an adventure through the lens of data!

Introduction:

The Palmer Penguins dataset provides a set of body size measurements for three penguin species—Adélie, Gentoo, and Chinstrap—collected from 2007 to 2009 in the Palmer Archipelago, Antarctica. It consists of 344 observations with variables such as body mass, flipper length, and culmen dimensions. This analysis aims to uncover differences between species and sex while exploring correlations among these physical measurements.Through the visual journey of data analysis, we hope to illuminate the unique characteristics of these remarkable birds, providing insights that might even help Alice make educated guesses about the species of her newfound friends. Who knows, she might even encounter some statistical paradox along the way!

Data description:

The dataset, available here, contains biological and geographical data on the three penguin species. It includes 17 columns and 344 observations, focusing on 7 variables (3 categorical and 4 numerical) corresponding to 333 observations after excluding missing values. The selected variables provide insights into the physical characteristics of the penguins and are straightforward to interpret. Additionally, the correlation matrix reveals significant correlations among these variables.

Table 1: The Penguin Dataset

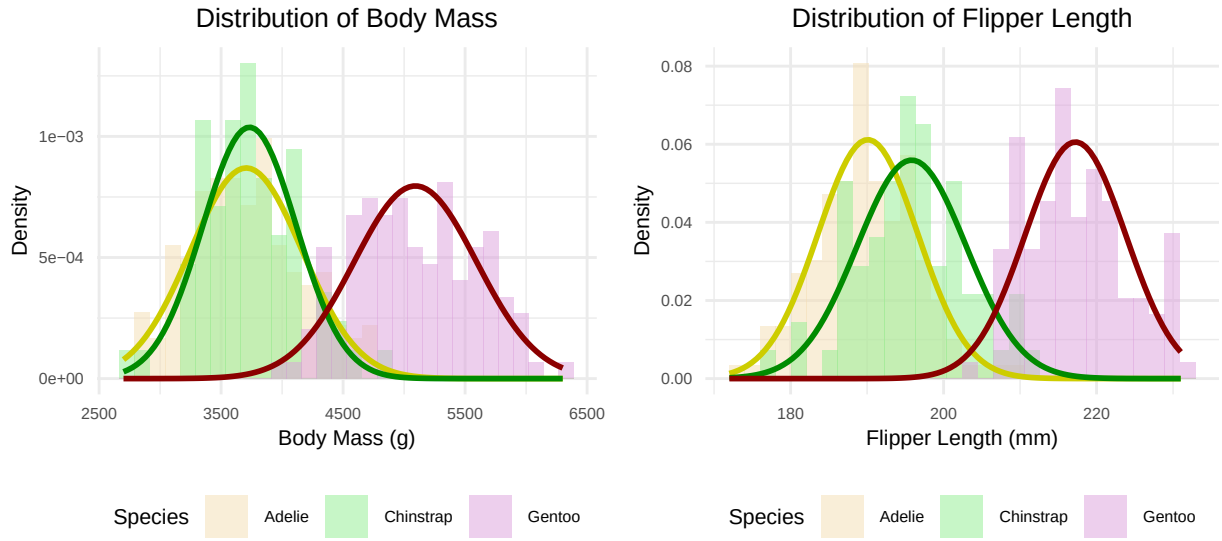
species	island	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g	sex
Adelie	Torgersen	39.1	18.7	181	3750	MALE
Adelie	Torgersen	39.5	17.4	186	3800	FEMALE
Adelie	Torgersen	40.3	18.0	195	3250	FEMALE
Adelie	Torgersen	36.7	19.3	193	3450	FEMALE
Adelie	Torgersen	39.3	20.6	190	3650	MALE
Adelie	Torgersen	38.9	17.8	181	3625	FEMALE

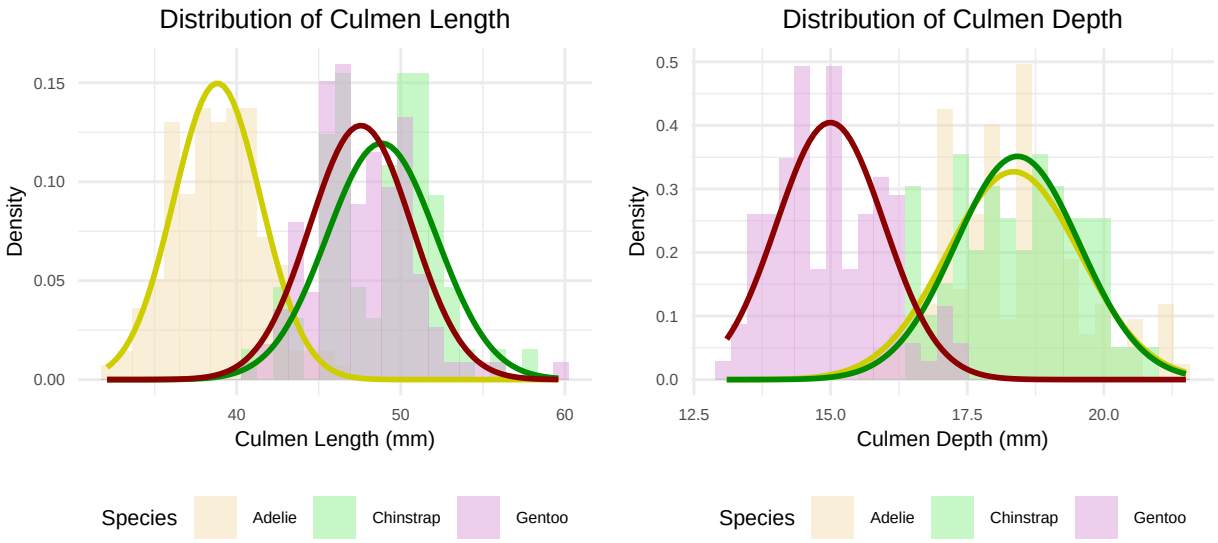
The Species variable identifies the penguin species, while the Island variable indicates its specific island in the archipelago (Biscoe, Dream, or Torgersen). The Sex variable denotes the sex of the penguin (female

or male). The dataset also includes numerical measurements: Culmen Length and Culmen Depth (both in millimeters) represent the bill's size, and the Flipper Length (in millimeters) measures the flipper's length. Finally, the Body Mass variable indicates the penguin's weight in grams. al): a factor denoting penguin sex (female, male).

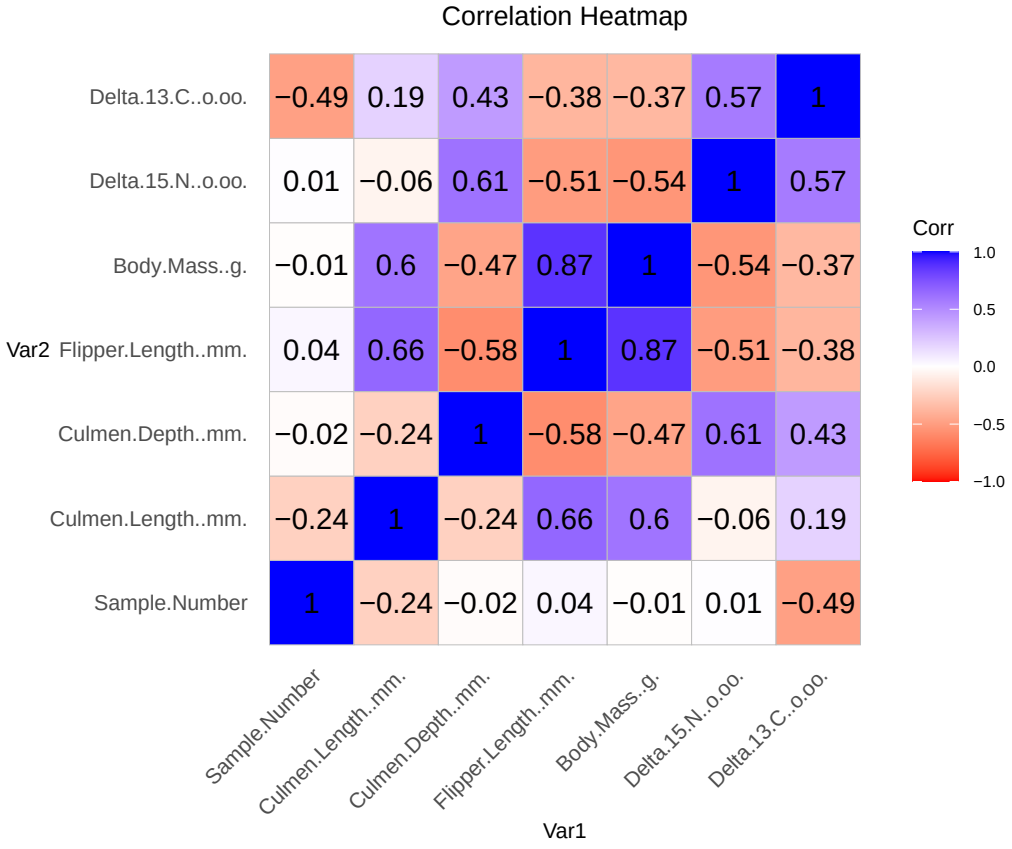
Exploratory Data Analysis:

We look at the distribution of the four numerical variables for the three species, which exhibit centralization tendencies. The peak of the Gaussian curve represents the average (mean) value for each species, indicating where most penguins cluster regarding traits like body mass and flipper length.



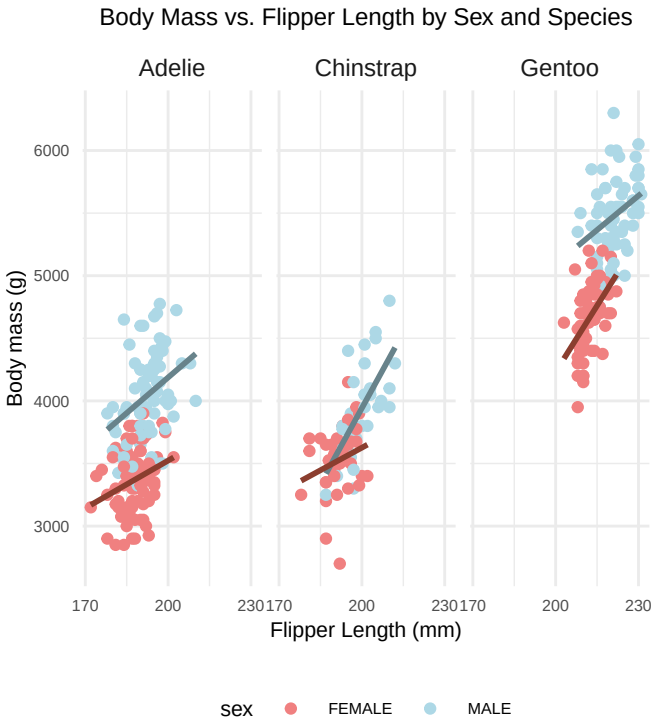


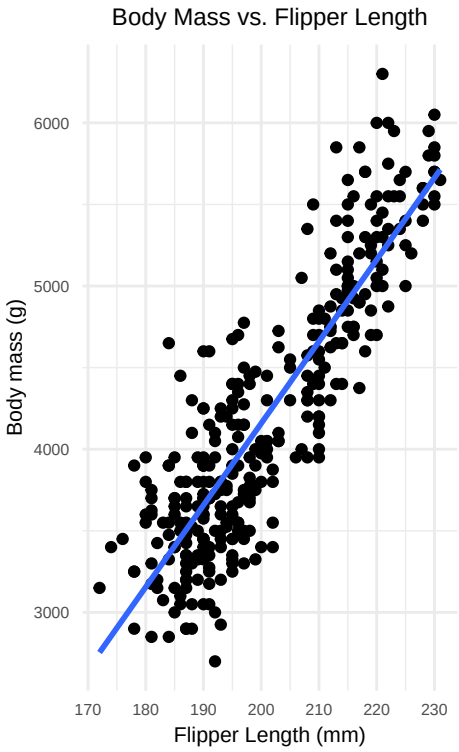
We examine the correlation between numeric variables, specifically focusing on easily interpretable body measurements. From the following matrix, Strong positive correlations are observed among flipper length ~ body mass, as well as flipper length ~ culmen length, and culmen length ~ body mass. Additionally, there is a notable negative correlation between flipper length ~ culmen depth.

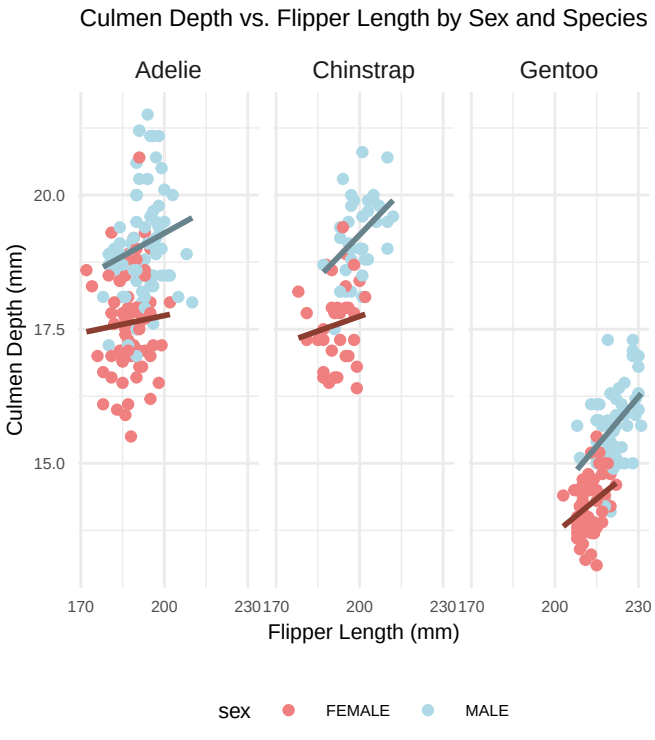


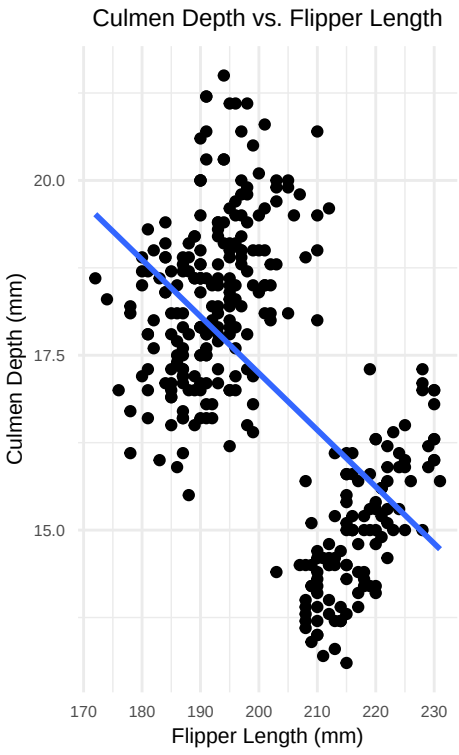
We will visualize the relationships between body mass and flipper length, as well as flipper length and culmen depth, both overall and faceted by species. Coloring by sex helps differentiate between the characteristics of males and females.Faceting by species allows for easy comparison between subsets of the data, showing how

the traits differ across groups (Adelie, Gentoo, Chinstrap).









Results and Discussion:

From the EDA, we observe the following:

- The sex ratio across species is nearly 1:1.
- The Adélie species is present on all three islands, while the Chinstrap and Gentoo are found only on Dream and Biscoe islands, respectively.
- The centralization tendency of the numerical variables is insightful. For example, body mass distribution shows that most individuals are around an average mass, with fewer being very light or heavy; one species tends to be heavier (i.e., a higher mean).
- In our sample, the number of penguins is: Adélie > Gentoo > Chinstrap.
- Average flipper length order: Adélie < Chinstrap < Gentoo.
- Average body mass order: Adélie Chinstrap < Gentoo.
- Average culmen length order: Adélie < Gentoo < Chinstrap.
- Average culmen depth order: Gentoo < Adélie Chinstrap.
- The correlation matrix reveals the highest positive correlation between body mass and flipper length, and the highest negative correlation between flipper length and culmen depth. The plots illustrate these correlations, but when faceted by species (and applicable for each sex), a positive correlation between flipper length and culmen depth emerges, illustrating Simpson's paradox.

Conclusion:

The analysis shows that Gentoo penguins are larger in terms of both flipper length and body mass compared to Adelie and Chinstrap penguins. Through visualizations, we confirmed a correlation between body measurements. Interestingly, we observe Simpson's Paradox in the scatter plot for Culmen Depth vs. flipper length when viewed across species and sex. While individual species show positive trends, the overall correlation across all penguins might give the false impression of a weaker relationship due to differences between species. This emphasizes the importance of analyzing data by subgroups to avoid misleading conclusions. Hopefully both Alice and the reader enjoyed the exploration!

Referances:

1. Gorman, K. B., Williams, T. D., & other collaborators. (2019). Palmer Penguins: A dataset for the study of penguins in the Antarctic. Retrieved from here.
2. Palmer Archipelago Penguins Data in the palmerpenguins R Package - An Alternative to Anderson's Irises.