ROBO 5008 - Intro to Research Assignment: Data & Figures

Steve Gillet

December 7, 2025

Instructions

What you turn in should be neatly prepared, and concisely address all parts of the assignment in a clearly organized manner. Be sure to explain your analysis and document all decisions (particularly for data cleaning). Code-related questions should be done in Python. Do not include in your final report, but rather include a link to a github repository You are encouraged to discuss your work with other students, but submitted work must be your own. Please indicate on the first page of your answer sheet who your collaborators for this assignment are. All figures require axes labels (including units where applicable), legends where appropriate, and a clear title or caption. Points will be deducted if these elements are not included.

1 Data Cleaning (25%)

Prepare a cleaned version of the dataset. Your cleaning process should:

- *Identify and handle missing values. Support your approach to handling missing data in your report.*
- *Standardize categorical variables (species, sex, island names).*
- *Identify outliers and decide whether to remove, adjust, or retain them (justify your choice).*

Deliverables:

- *Cleaned CSV file.*
- *Brief written description (≤200 words) of all cleaning steps and your rationale.*

1.1 Cleaning Summary

So for missing values my strategy was to drop any rows where all of the key numerical measurements were missing as these don't add anything useful to the dataset, where only some of the data is missing I imputed the missing data with the mean for that species group so that that particular measure isn't changed while still being able to use the other useful data. Where the sex was missing I kept that as NaN since that data might still be useful for giving information about the species, imputing a sex would skew sex information unecessarily. For category variables I standardized the species names so that they are all either simply 'Adelie', 'Gentoo', and 'Chinstrap' to make them easier to handle and read. With outliers I just left them, the data is meanful and changing it would bias the data towards average penguins and weird penguins are the ones that change the world.

2 Exploratory Data Analysis (25%)

Explore relationships in the cleaned dataset:

- *Compare measurements (bill length, flipper length, body mass, etc.) by species and sex.*
- *Explore correlations between numerical measurements.*
- *Visualize distributions and relationships with appropriate plots.*

Deliverables:

- *2–5 exploratory figures with captions. These figures need not be as polished as your final figures.*
- *2–3 sentences of interpretation for each figure.*

2.1 Exploratory Figures and Interpretations

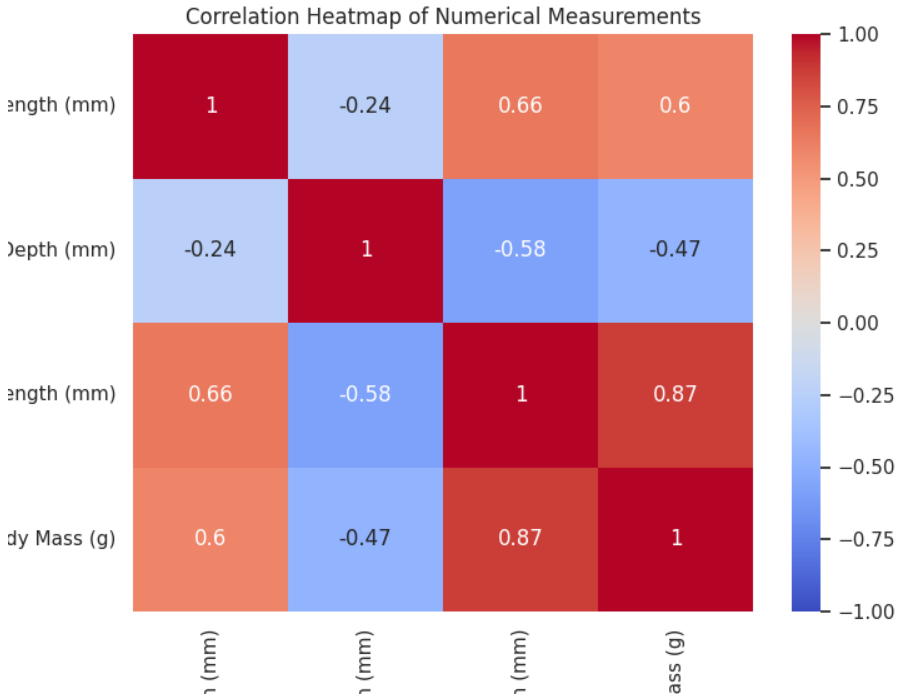

Figure 1: Data Correlation Heatmap.

I started with the heatmap to see how much correlation there is amongst the different variables. You can see some promising ones to start off with like culmen depth and flipper length.

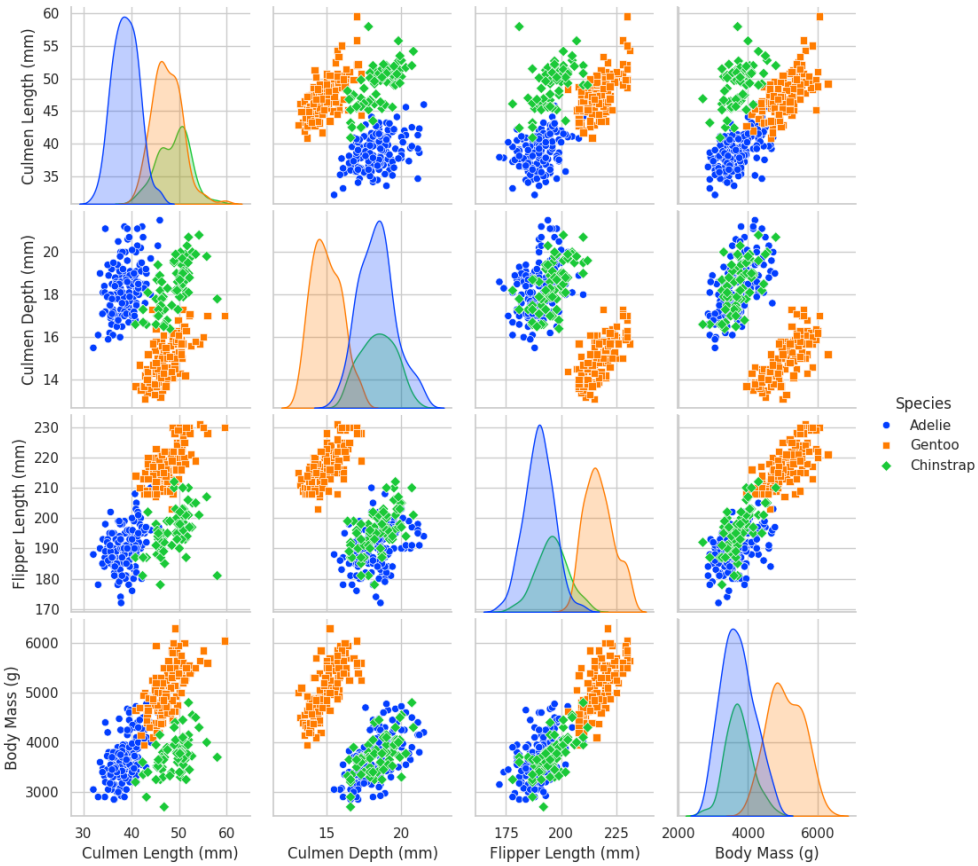

Figure 2: Data Comparison by Species.

I then plotted all of the variables against eachother and colored it by species and you can see that some of the data that had the lowest correlation is actually very correlated by species so I went for culmen length versus culmen depth for further investigation though culmen length versus flipper length is also pretty uncorrelated by species.

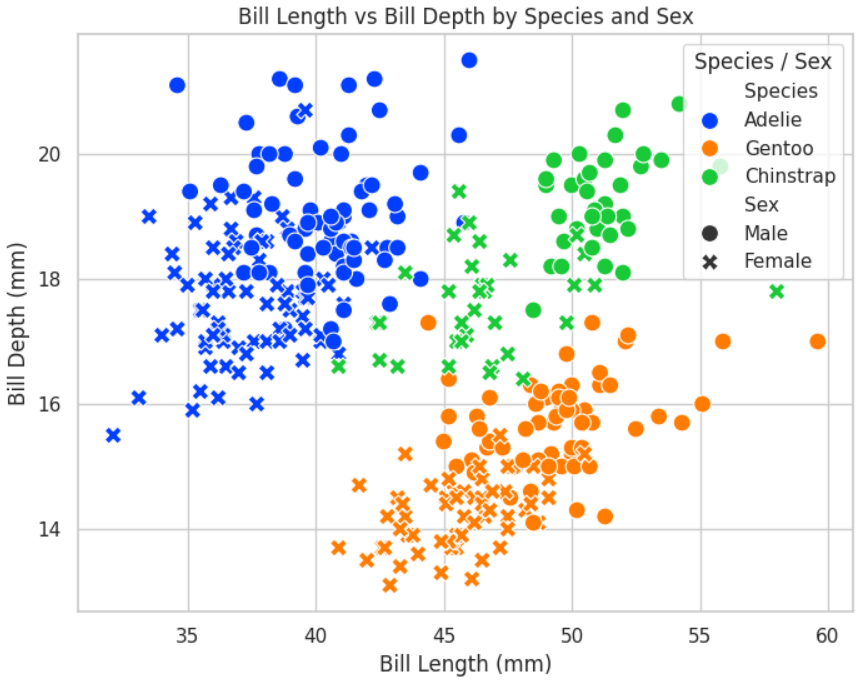

Figure 3: Culmen Length vs. Culmen Depth by Species and Sex.

Then I plotted culmen length versus culmen depth by species and sex and you can see some pretty clear grouping. I'm fairly confident I could train a ML model to identify a penguins species and gender based on their honker with some degree of accuracy.

3 Basic Modeling (25%)

Choose one modeling path:

- *Regression: Predict body mass from bill length, flipper length, and sex.*
- *Classification: Predict species from numerical measurements.*

Deliverables:

- *Model summary table (coefficients, accuracy, and/or R^2 as appropriate).*
- *A paragraph discussing the modeling approach and potential shortcomings.*
- *1–2 sentence interpretation of results.*

Metric	Value
Accuracy	0.8507
Average Precision (weighted)	0.86
Average Recall (weighted)	0.85
Average F1-Score (weighted)	0.85

Table 1: Model summary metrics for combined species/sex classification.

Class	Culmen Length Coef	Culmen Depth Coef	Intercept
Adelie Female	-1.274	1.888	22.87
Adelie Male	-0.823	3.095	-16.98
Chinstrap Female	0.081	-0.242	3.630
Chinstrap Male	0.338	0.227	-17.09
Gentoo Female	0.619	-3.006	22.13
Gentoo Male	1.059	-1.962	-14.57

Table 2: Logistic regression coefficients per class.

3.1 Discussion

I used multinomial logistic regression here for simplicity to classify the species/sex from culmen length/depth. The approach involves label-encoding the target, splitting the data 80/20 (train/test), and fitting the model. Potential shortcomings include overly simplistic model, data correlation, and small sample size. These could be mitigated by incorporating features that capture sex differences better.

3.2 Interpretation

The model achieved 85% accuracy which tells you that using culmen dimensions is effective for distinguishing these penguin species and even sex to a slightly lesser extent.

4 Final Figures (25%)

Select two figures from your analysis that best communicate your findings. Refine them for publication-quality presentation (this may require post-processing in 3rd party software such as Adobe Illustrator or Affinity Designer):

- *Clear, consistent labeling and color scheme.*
- *Informative captions.*
- *Axes labeled with units where applicable.*

Deliverables:

- *Two final polished figures with captions.*
- *Brief paragraph explaining why these figures were chosen, what data they portray and what your conclusions are from the figures.*
- *2–3 sentences on your approach and workflow to generate these final figures. Are there any improvements you envision for the future?*

4.1 Polished Figures

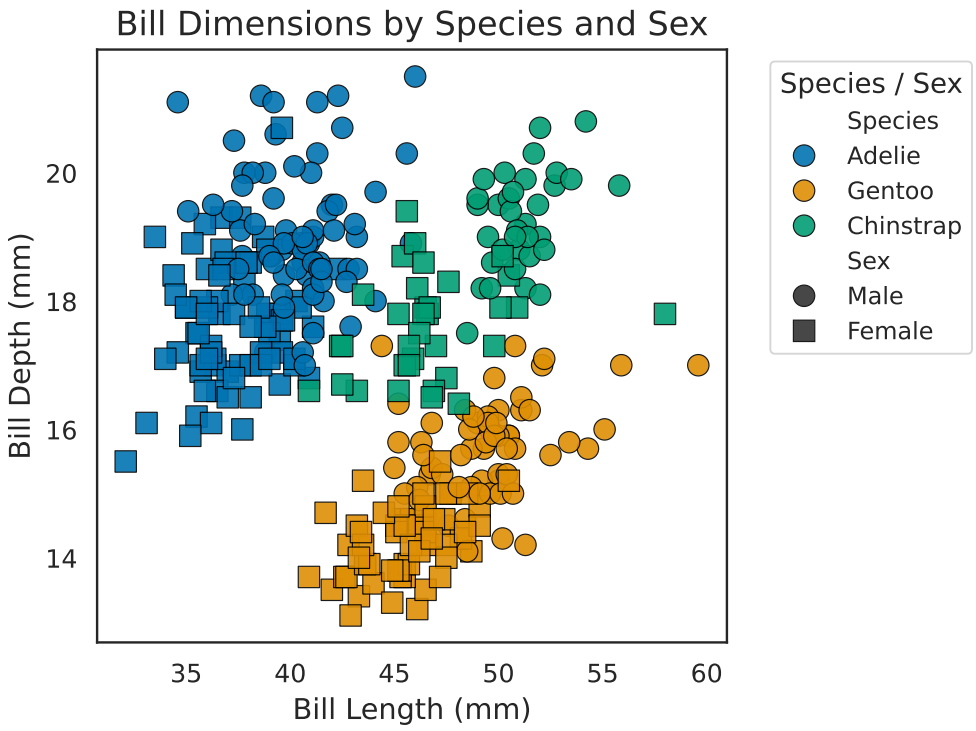

Figure 4: Bill Length v Bill Depth.

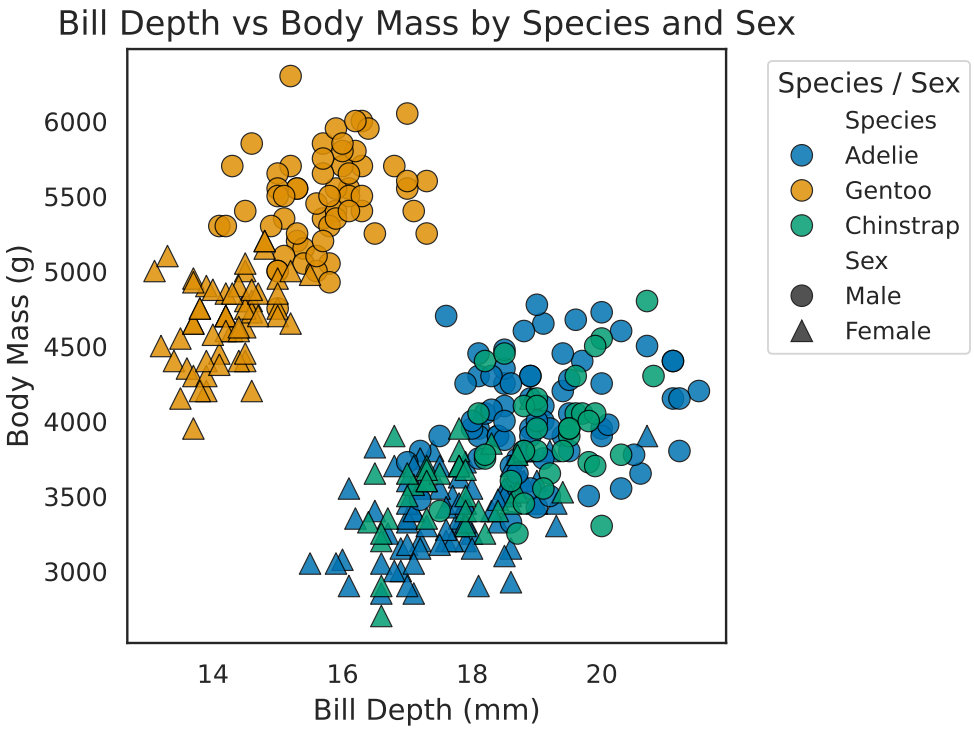

Figure 5: Bill Depth v. Body Mass.

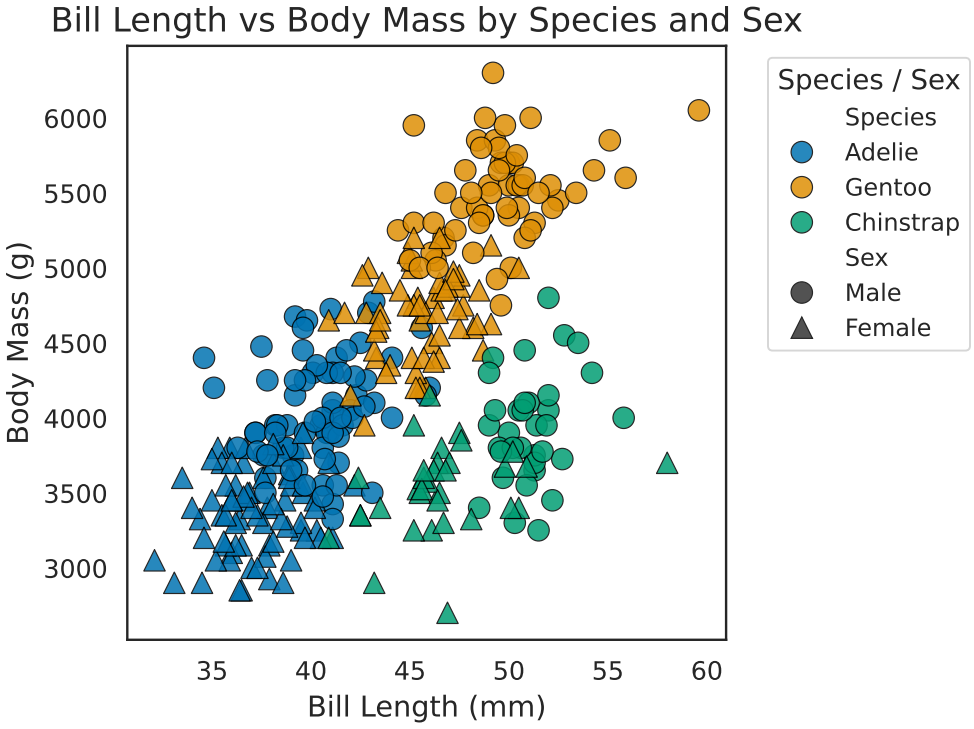

Figure 6: Bill Length v. Body Mass.

4.2 Explanation

So I improved my bill length/depth plot and then I added a plot for bill length vs body mass and bill depth vs body mass because I wanted to show how adding body mass adds a dimension by which you can differentiate the sex of the penguins better and improve the accuracy of our model.

4.3 Approach and Workflow

I used seaborn context and styles to get the plots a bit more official looking based on what other researchers did. I really wish there was a way to show the 3D plot of body mass vs bill length vs bill depth because then you really get a sense for how adding body mass makes that sex difference pop out, maybe there's other plots and I should have approached it from a different angle but I'd have to think about it. I also want to play with the colors more because the plots look really nice and clean but I just think they need some little bit more background or lines or shadows or something to make them pop, the next thing I would look into is how to add a background color while maintaing the colorblind color scheme.

5 GitHub Repository

The code used for data cleaning, analysis, modeling, and figure generation, along with the cleaned CSV file, is available at: `https://github.com/yourusername/yourrepo`