MAS8403: Palmer Achipelago Penguins

210431461 | 21/10/22

Introduction

To the West of the Antarctic Peninsula, extending North and South of the Palmer Basin, is the Palmer Long Term Ecological Research (LTER) (Rutgers and LTER (2022)) study area. Midway down the the Antarctic Peninsula, on Anvers Island, is Palmer Station. Researchers are staffed there to monitor the polar marine biome, including the local penguin population (Foundation (n.d.)).

The LTER researchers are using penguin-borne sensors to inform long-term studies on penguin population dynamics and improve our understanding of how Antarctic penguins are adjusting to rapid climate changes (Program (n.d.)). The most dramatic effects of climate change are being observed in our polar regions (Program (n.d.)).

This report is informed by a dataset called `penguins` from the `pamlerpenguins` R package; it is one of two packages provided by Palmer LTER researchers (Alison Hill and Gorman (2020)). The dataset is pre-processed, so accuracy and quality are assumed. The dataset provides 333 observations across 8 variables; it is assumed that this is a sample of the data collected by Palmer Station. The `set.seed()` and `sample` functions in R were to generate a random, representative sample to inform this report.

Objectives

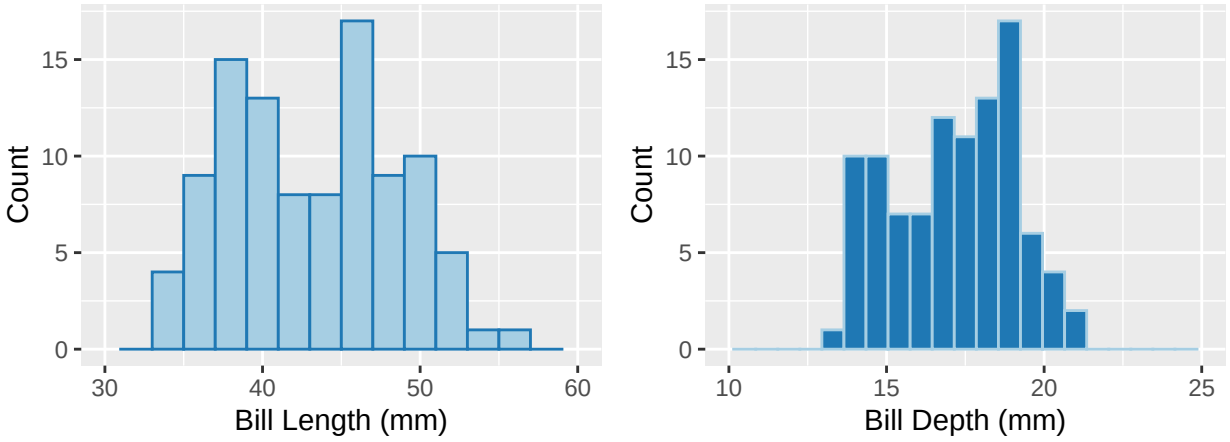
The exploratory data analysis outlined in this report explores a sample of 100 penguins from the `penguins` dataset. There are four objectives for this analysis:

1. identify an appropriate probability distribution to represent at least one measurement variable (bill length, bill depth, flipper length and body mass);
2. find estimates for the parameters of the distribution of your data;
3. identify which variables are likely to reliably estimate the sex of a penguin; and
4. identify if the penguins' location (island) appears to have a significant impact on any of its physical characteristics.

Data Exploration

Objective 1 (Distribution)

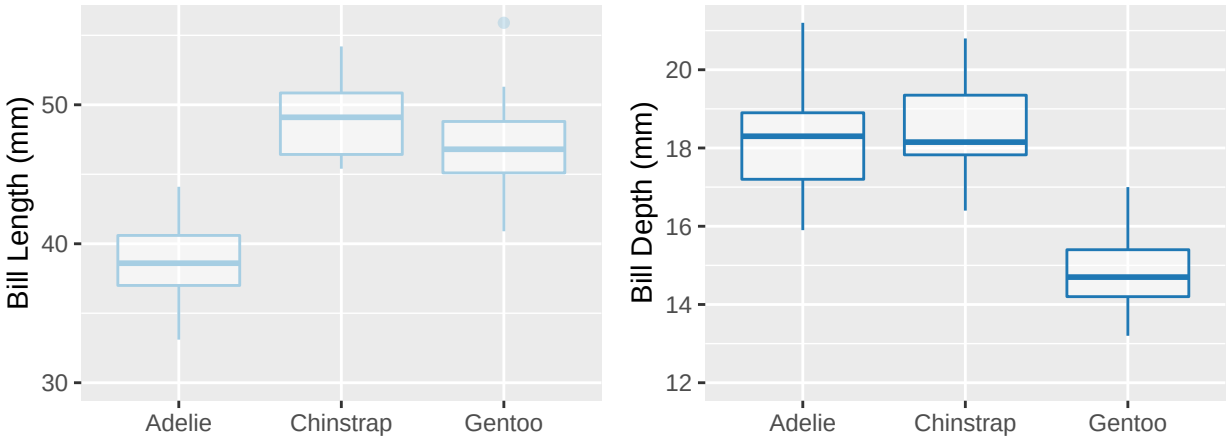
The data sample includes 8 variables that provide information relating to the 100 penguins. The species (Adelie, Chinstrap or Gentoo), island (Biscoe, Dream or Torgerson), and sex (male or female) are nominal, qualitative values. The year is discrete quantitative data that identifies when the variables were recorded (2007, 2008 or 2009). The variables bill length (mm), bill depth (mm), flipper length (mm) and body mass (g) are quantitative, numerical measures of each penguin. These 4 variables of measurement data are continuous and random; the variables adopt a smooth range of values (Newcastle University (2022)). The plots below show this (all histograms are provided in `Appendix A`).



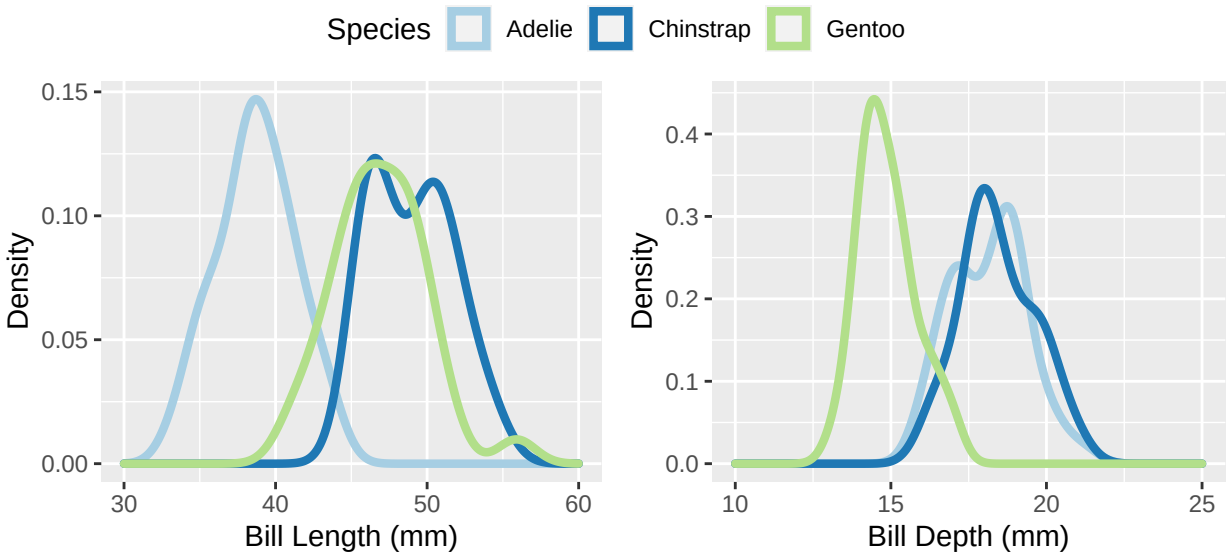
These histograms visualise the distribution of penguins' bill length (BL), bill depth (BD), flipper length (FL), and body mass (BM) data. The measurement data appears to be multimodal with some variance.

These histograms suggest that there may be elements, such as the penguins' species, which are influencing the distribution of the data.

Filtering the data demonstrates that the BL, BD, FL, and BM changes by penguin species. The boxplots for BL and DP are provided below (all boxplots are provided in `Appendix B`).



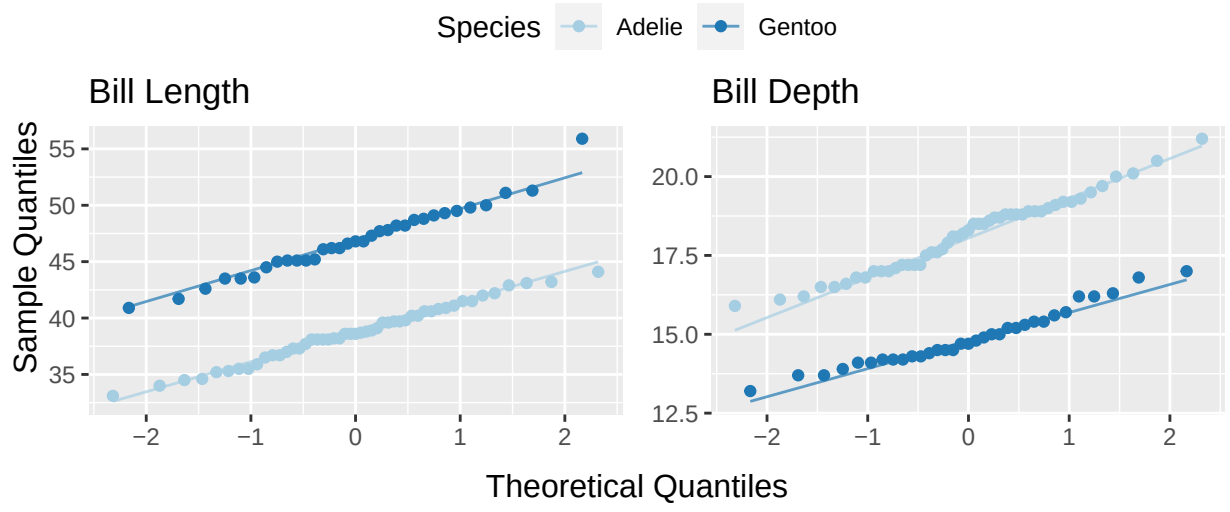
It is likely to be more significant, therefore, to observe the distribution, and estimate the parameters of the population, separately for each penguin species. For each species, the distribution of BL and BD are visualised below (all density distributions are provided in `Appendix C`).



Measurement data (e.g. weight or height) of a population frequently follows a normal distribution. Initially, however, the density distributions of BL, BD, FL, and BM of the penguin species do not appear to follow a normal distribution. On closer inspection, some variables do appear to approximate a normal distribution, especially the BL, FL and BM of Adelie penguins.

It is important to remember that our data sample is small. In the sample of 100 penguins, there are 49 Adelie, 33 Gentoo, and 18 Chinstrap. It is likely that there is too little measurement data for Chinstrap penguins to accurately confirm its data distribution.

The Q-Q plots below, therefore, test whether or not the BL and BD data of Adelie and Gentoo penguins are normally distributed (all Q-Q plots are provided in `Appendix D`).



These Q-Q plots demonstrate that the BL, BD, FL and BM of Adelie and Gentoo penguins does approximate a normal distribution. It is important to note that the density distributions suggest there is another element influencing the data distribution. `Appendix E` presents Q-Q plots that clearly demonstrate the data distributions for male and female penguins are different for all species, and this will be explored further in `Objective 2`.

Objective 2 (Distribution Parameters)

A normal distribution is characterised by two paramenters; these are the 'mean' and 'standard deviation.'

The 'mean,' its 95% 'confidence interval' (95% CI), and 'standard deviation' (SD) of BL and BD from the sample data for each penguin species are presented in the tables below (all parameters are provided in `Appendix F`).

Table 1: Bill Length Summary Statistics

Species	Mean (mm)	95% CI (mm)	SD
Adelie	38.70	37.98-39.42	2.58
Gentoo	46.88	45.83-47.93	3.09
Chinstrap	48.99	48.08-49.90	2.68

Table 2: Bill Depth Summary Statistics

Species	Mean (mm)	95% CI (mm)	SD
Adelie	18.15	17.81-18.49	1.21
Gentoo	14.88	14.57-15.19	0.91
Chinstrap	18.44	18.03-18.85	1.19

The 'mean' and 'standard deviation' as parameters of our BL, BD, FL and BM distributions will be limited as estimators for the population parameters. There is only one sample to inform the estimators and, when filtered by species, the sample set is small. The 95% 'confidence interval,' therefore, provides us with the interval that indicates how close the estimated 'mean' is likely to be to the true value; we can be 95% confident that the population mean will be between this interval.

In accordance with the central limit theorem, however, as we increase the sample size the sample 'mean' will approach the population 'mean.' Similarly, if there were more sample sets, the 'mean' and 'standard deviation' of the samples' parameters would better estimate the population parameters.

Objective 3 (Estimate Penguin Sex)

LTER researchers at Palmer station would like to estimate the sec of a penguin from measurement data to avoid the need for invasive procedures that cause penguin distress.

The Q-Q plots in `Appendix E` appear to suggest that the data distribution of male and female penguins differ, regardless of their species. The difference between the BL and BM is most visually obvious. Two-sample t-tests were used to test whether this difference is statistically significant.

These tests tell us that in our sample of Adelie and Gentoo penguins, the difference between males and females BM 'mean' is more statistically significant than BL 'mean.' In our sample of Chinstrap penguins, the difference between males and females BL 'mean' is more statistically significant than BM 'mean.'

The BM two-sample t-test for Adelie penguins is set out below (all two-sample t-tests are provided in `Appendix G`.)

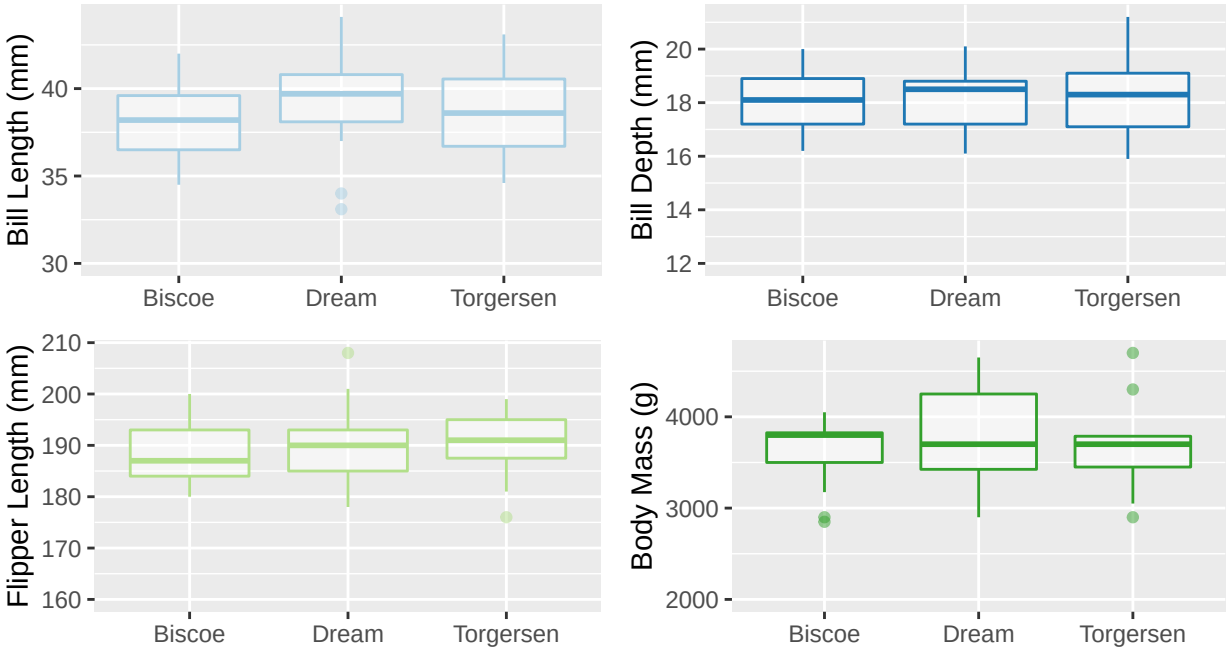
Adelie two-sample t-test for BM: Of the 49 Adelie penguins, there are 26 females and 23 males. We would like to compare BL observations from Adelie females (F) to determine that their 'mean' is different to observations from Adelie males (M).

$$\begin{aligned} &\mathrm{H}_0 : \mu_F = \mu_M \\ &\mathrm{H}_1 : \mu_F \neq \mu_M \end{aligned}$$

The `barlett.test` function for equal variances confirms that the assumption of equal variance is valid (p-value = 0.42). The `t.test` function provides us with a p-value of 2.926e-06. This suggests that there is very strong evidence against H_0 and reject it in favour of H_1.

Objective 4 (Island Impact)

Only one species of penguin is found on all three islands (the bar chat is in `Appendix H`); this is the Adelie penguin. The boxplots below show the BL, BD, FL and BM for the Adelie penguins on each island.

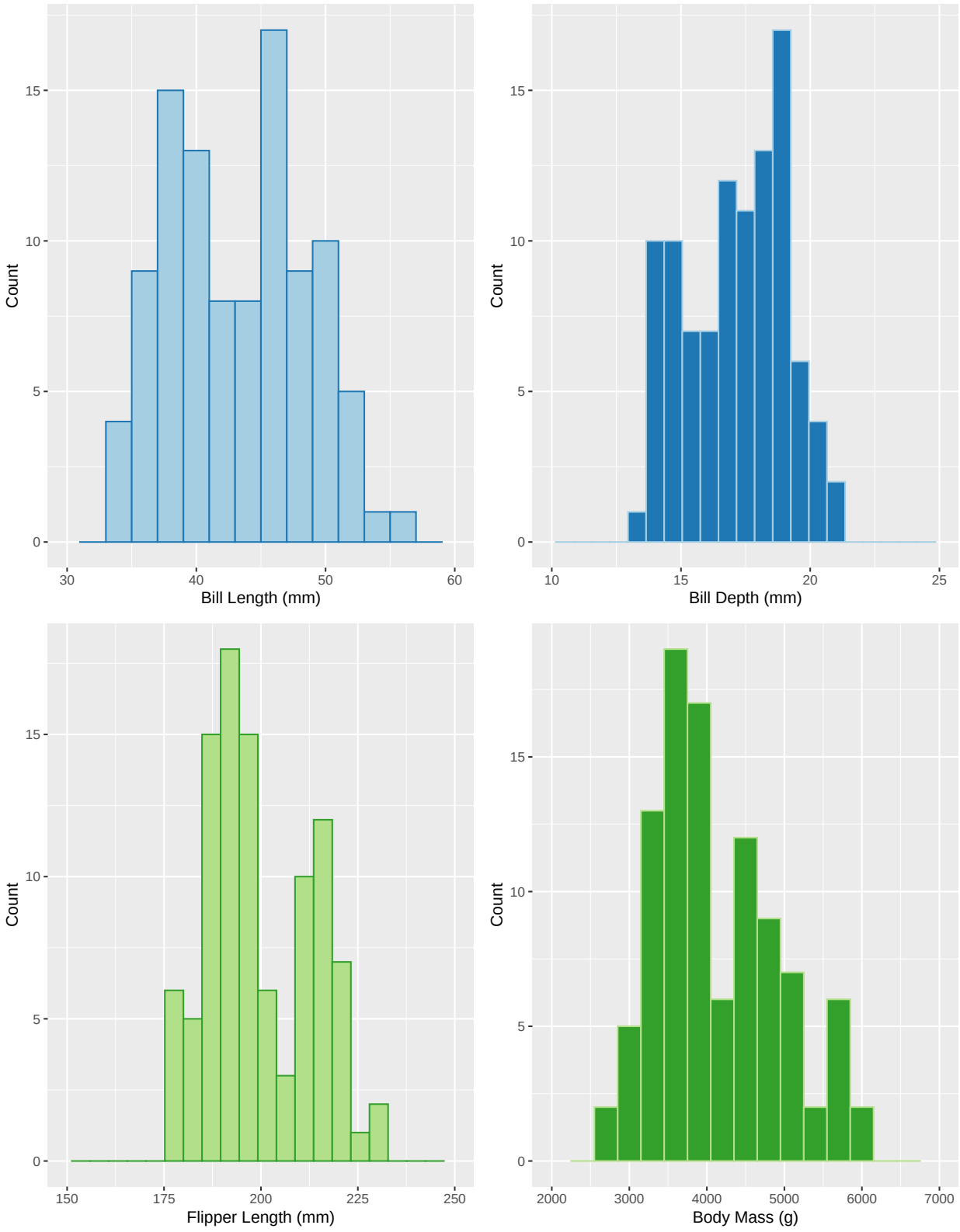


The boxplots suggest that location may impact the phsyical characteristics of the Adelie penguins. To test if this impact is statistically significant, two-sample t-tests were used (and are provided in `Appendix I`).

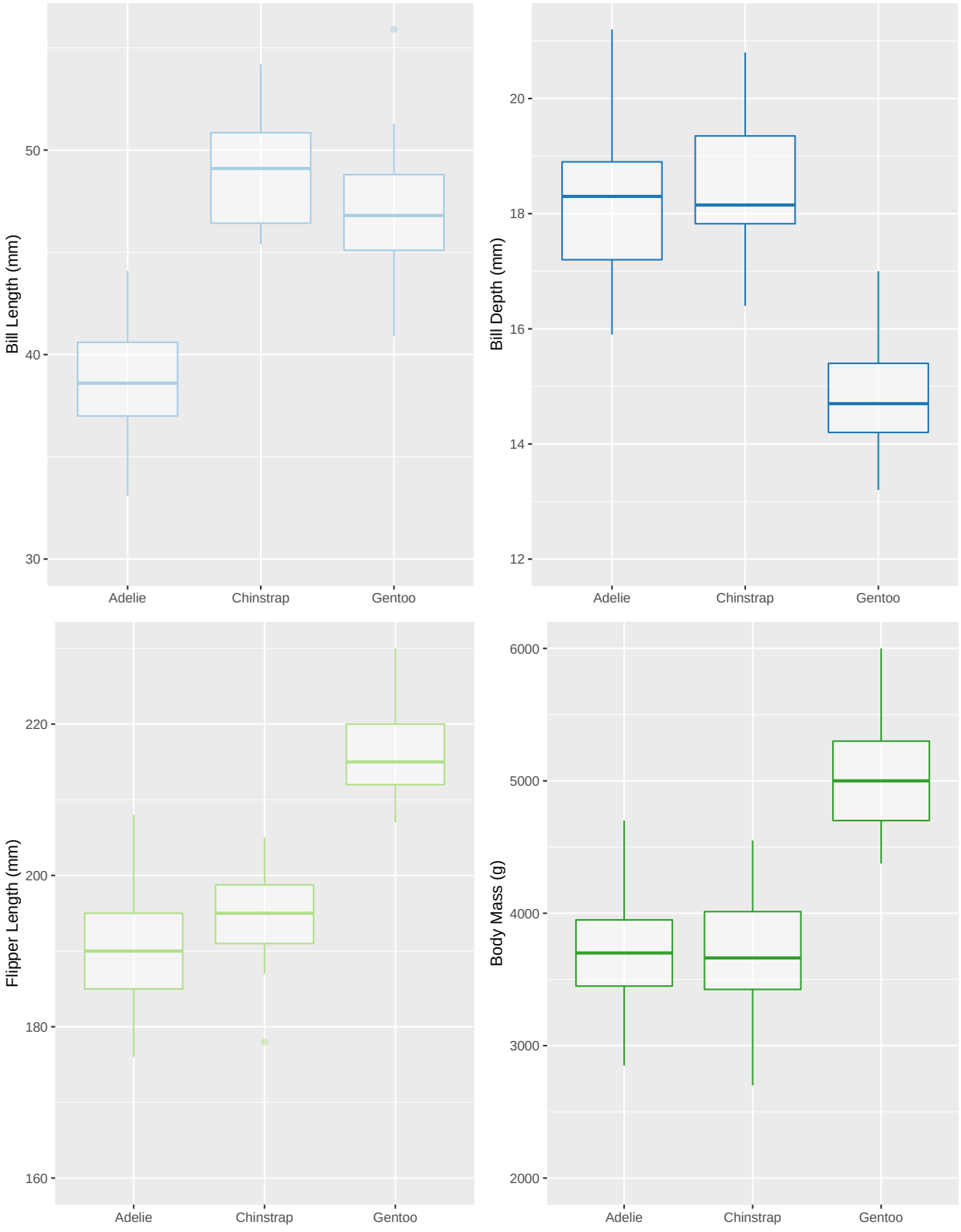
(Show the conclusion and significance)

Evalation

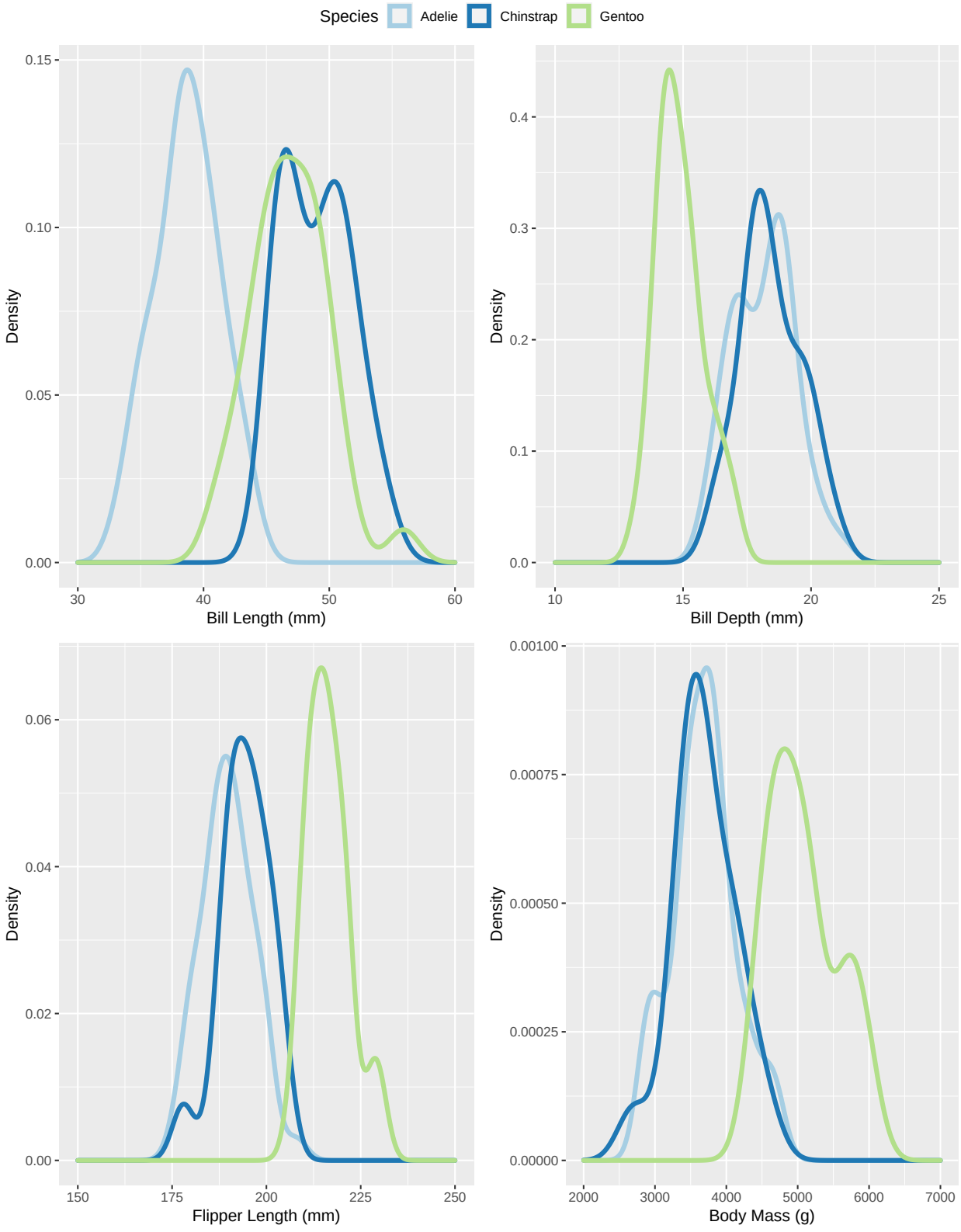
Appendix A: Histograms of Measurement Variables



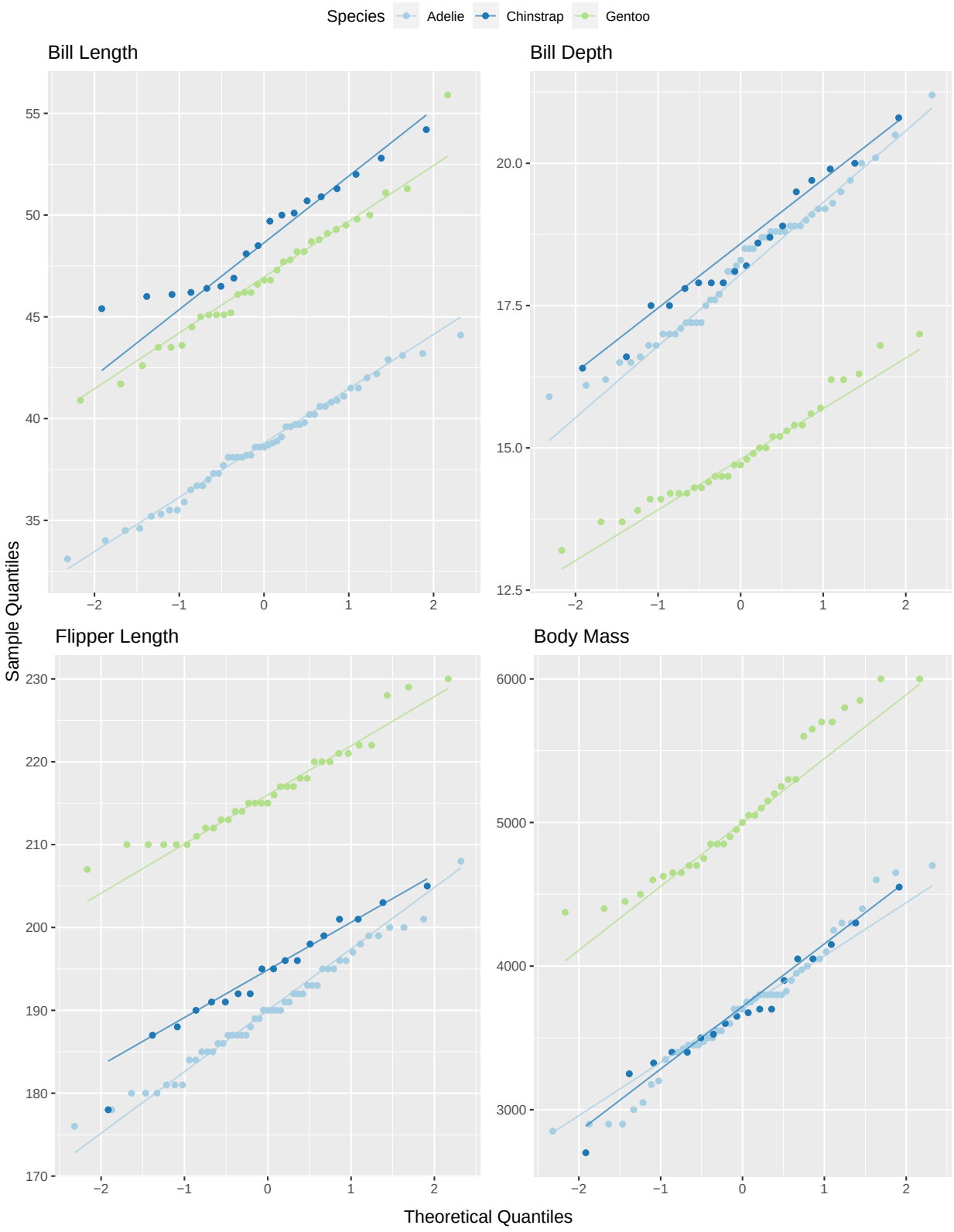
Appendix B: Boxplots of Measurement Variables by Species



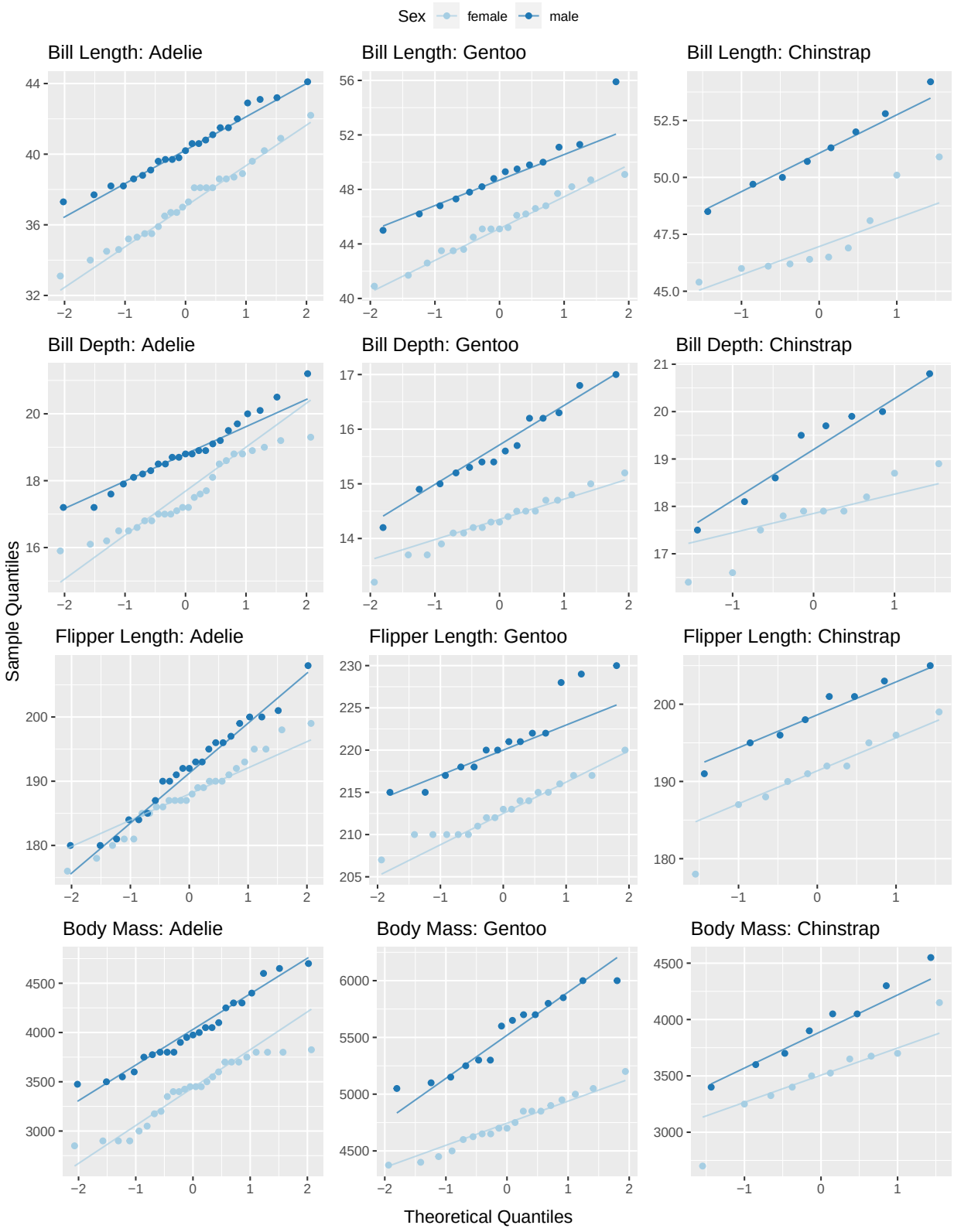
Appendix C: Density Distributions of Measurement Variables by Species



Appendix D: Q-Q Plots of Measurement Variables of Adelies and Gentoos



Appendix E: Q-Q Plots of Measurement Variables by Species



Appendix F: Confidence Intervales for Parameter Estimates

Table 3: Bill Length Summary Statistics

Species	Mean (mm)	95% CI (mm)	SD
Adelie	38.70	37.98-39.42	2.58
Gentoo	46.88	45.83-47.93	3.09
Chinstrap	48.99	48.08-49.90	2.68

Table 4: Bill Depth Summary Statistics

Species	Mean (mm)	95% CI (mm)	SD
Adelie	18.15	17.81-18.49	1.21
Gentoo	14.88	14.57-15.19	0.91
Chinstrap	18.44	18.03-18.85	1.19

Table 5: Flipper Length Summary Statistics

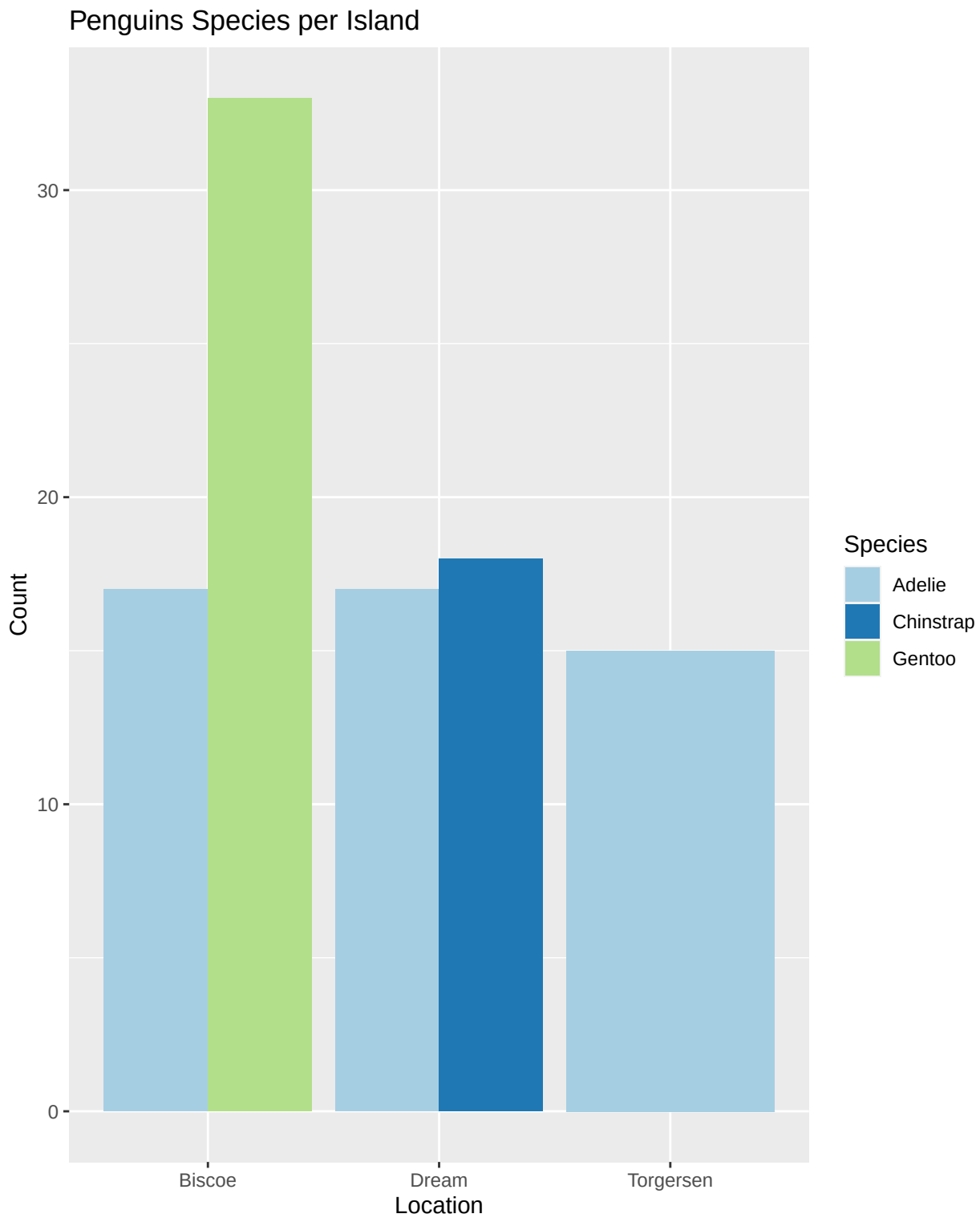
Species	Mean (mm)	95% CI (mm)	SD
Adelie	189.78	187.88-191.68	6.80
Gentoo	216.42	214.48-218.36	5.68
Chinstrap	194.33	192.06-196.60	6.65

Table 6: Body Mass Summary Statistics

Species	Mean (g)	95% CI (g)	SD
Adelie	3691.84	3565.18-3818.50	452.38
Gentoo	5075.76	4911.39-5240.13	481.76
Chinstrap	3690.28	3542.83-3837.73	432.17

Appendix G: Two-sample t-tests for Penguin Sex

Appendix H: Bar chart of Penguin Species Location



Appendix I: Two-sample t-tests for Adelie location

Bibliography [CHECK AS NOT ALL ARE APPEARING]

Alison Hill, Allison Horst, and Krissten Gorman. 2020. "Release the Penguins." https://education.rstudio.com/blog/2020/07/palmerpenguins-cran/.

Foundation, Nationa Science. n.d. "Palmer Station Webcams." https://www.usap.gov/videoclipsandmaps/palwebcam.cfm?t=1.

Newcastle University, Joe Matthews on behalf of. 2022. "Mas8403 1: Introduction."

Program, Long Term Ecological Research. n.d. "Penguin Science." https://penguinscience.com/.

Rutgers, The State University of New Jersey, and Palmer Station Antarctica LTER. 2022. "Palmer Station Antarctica LTER." https://pallter.marine.rutgers.edu/.