Exploratory Data Analysis of the *Palmer Penguins* Dataset

GDAA 1001 Fundamentals of Data Analysis: Assignment 2 - Exploratory Data Analysis

Mohit Francis (W0476572)

2022-11-13

Contents

List of Figures

List of Tables

Introduction

Exploratory data analysis (EDA) is a way of assessing datasets through data transformation and data visualisation. There is no set method for EDA but it may involve: calculating summary statistics for the variables within the dataset; examining the variation within variables using bar graphs, density plots, histograms, and box plots; examining covariation between variables using scatterplots and heat maps.

For this assignment, I perform exploratory data analysis on a dataset of my choice. Per assignment parameters, this dataset has to be an external dataset (i.e. not a built-in dataset with `R`), either spatial or non-spatial in nature, with at least one hundred observations, and five variables, two of which have to be categorical. Next, I ask general questions about my data in terms of summary statistics, distributions and correlations. These questions then inform the various ways I explore my dataset. Finally, I revisit these questions after thoroughly exploring my dataset and I produce this report containing all of my work.

After setting my working directory, I load the following libraries into my session of `R` to accomplish the above tasks. I use the `tidyverse`, `data.table` and `kableExtra` libraries for data transformations. I use the `ggplot2`, `GGally`, `RColorBrewer`, `grid`, and `gridExtra` libraries for data visualisation.

```
library(tidyverse)
library(data.table)
library(kableExtra)
library(ggplot2)
library(RColorBrewer)
library(GGally)
library(grid)
library(gridExtra)
```

Data Selection

I perform exploratory data analysis (as well as some data tidying beforehand) on the `Palmer Penguins` dataset. I found the `Palmer Penguins` dataset on the *R For Data Science: Tidy Tuesday Github*[1]. The original authors of the data, who first collected and made this dataset available, are: Dr. Kristen Gorman, Dr. Allison Horst, Dr. Alison Hill, and Palmer Station, Antarctica Long Term Ecological Research (LTER)[2].

This dataset is a fairly simple, beginner-friendly dataset specifically for novices such as myself to practise their data tidying, exploration, and visualisation skills. Gorman et al., describe this dataset as an alternative to the built-in `R` dataset `iris`, which is also beginner friendly (2020).

Many of the datasets on Kaggle did not meet the above specifications. Additionally, the Kaggle datasets have specific, more advanced purposes such as clustering, regression, or classification, and I did not feel comfortable using these datasets yet. On the other hand, while census data and weather data meet the above requirements, I thought it best to use a simple but 'fun' dataset.

As `Palmer Penguins` is a dataset containing measurements of penguins, I expect to see differences in body characteristics in terms of sex and species. But I am interested in the extent of intra-species and inter-species variation. I also wonder if there is a difference in body traits in penguins living on different islands, but which are of the same species.

There are two versions of the dataset: a *clean* and a *raw* version. The *clean* version is already tidy (though not completely) whereas the *raw* version is the raw data the authors collected. I use the *raw* version to further practise my data tidying skills, but I use the *clean* dataset[3] as a reference for my data tidying.

[1] https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-07-28/readme.md
[2] https://allisonhorst.github.io/palmerpenguins/
[3] https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-07-28/penguins.csv

Data Preparation

The *raw* dataset is stored as a comma separated values (CSV, .csv) file online in the *R For Data Science: Tidy Tuesday Github* repository[4]. I import the file into my session of `R` using the `dplyer` function `read_csv()` (from the `tidyverse` library) and I store it in a tibble in `R` I name `penguins`.

```
penguins <- read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/20
```

The dataset contains 344 observations and 17 variables. Using the built-in `R` function `str()` I examine the structure of the tibble (I do not display my results, however). There is an even mix of numeric and string data types within this dataset, along with a single `Date` data type. The following list describes each variable:

- *Study Name* refers to the codename of the field study undertaken to sample the penguins.
- *Sample Number* refers to the sample cohort within a specific study.
- *Species* refers to the species of penguin sampled - Adelie, Chinstrap, and Gentoo, along with their genus and species in binomial nomenclature.
- *Region* refers to the area in Antarctica where the study was conducted.
- *Island* refers to the island within the region in Antarctica where the study was conducted.
- *Stage* refers to the the life and breeding stage of each adult penguin.
- *Individual ID* is the number denoting the individual penguin and penguin couple sampled.
- *Clutch Completion* is a boolean masked as a string, signifying whether a penguin individual or couple successfully hatched an egg.
- *Date Egg* is the date the penguin laid an egg.
- *Culmen Length (mm)* is a measure of the length of a penguin's bill in millimetres, from the dorsal ridge atop the bill down to the tip (Gorman et al., 2020).
- *Culmen Depth (mm)* is a measure of the depth of a penguin's bill in millimetres, from the dorsal ridge atop the bill to the ventral surface of the bill (Gorman et al., 2020).
- *Flipper Length (mm)* is a measure of the length of the penguin's flipper in millimetres.
- *Body mass (g)* signifies the weight or body mass in grams of a penguin,
- *Sex* signifies whether an individual is male or female.
- *Delta 15 N (o/oo)* refers to Blood isotopic Nitrogen and is used used for dietary comparison (Gorman et al., 2020).
- *Delta 13 C (o/oo)* refers to Blood isotopic Carbon and is used used for dietary comparison (Gorman et al., 2020).
- *Comments* are notes made by researchers.

```
str(penguins)
```

To tidy the *raw* dataset, first I determine which variables in the *raw* dataset contain values which are either non-descriptive, or contain single values, i.e. unnecessary values. I use the built-in `R` function `unique()` to do so. While I did not list the results of my output, my code is below.

```
unique(penguins$studyName)
unique(penguins$`Sample Number`)
unique(penguins$Species)
unique(penguins$Region)
unique(penguins$Island)
unique(penguins$Stage)
unique(penguins$`Individual ID`)
unique(penguins$`Clutch Completion`)
unique(penguins$`Date Egg`)
```

[4] https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-07-28/penguins_raw.csv

```
unique(penguins$`Culmen Length (mm)`)
unique(penguins$`Culmen Depth (mm)`)
unique(penguins$`Flipper Length (mm)`)
unique(penguins$`Body Mass (g)`)
unique(penguins$Sex)
unique(penguins$`Delta 15 N (o/oo)`)
unique(penguins$`Delta 13 C (o/oo)`)
unique(penguins$Comments)
```

Several columns contain single values: *Region*, *Stage*, *Clutch Completion*. Other columns contain non-descriptive categorical values : *StudyName*, *Sample Number*, *Individual ID*, *Date Egg*, and *Comments*. Further, there are many `NULL` values present in *Delta 13 C (o/oo)* and *Delta 15 N (o/oo)*. I subset my dataset and store it in a new variable I name `penguins2` to select the variables I deem necessary, i.e. all other variables not listed above.

```
penguins %>% select(
  Species,
  Island,
  `Culmen Length (mm)`,
  `Culmen Depth (mm)`,
  `Flipper Length (mm)`,
  `Body Mass (g)`,
  Sex,
) -> penguins2
```

Next I add an extra column I name *Penguin Number* which numbers each observation. This is not a necessary step, but I decide to do so anyway.

```
penguins2$`Penguin Number` <- 1:nrow(penguins2)
```

I split *Species* into two columns since it does not hold an atomic value, but instead contains two values: the species name, and the name of each species in binomial nomenclature. Though not necessary, I retain the names of each species in binomial nomenclature and I store these names in a separate column I name *Binomial Nomenclature*.

```
penguins2 <- separate(penguins2,
                      Species,
                      into = c("Species", "Binomial Nomenclature"))
```

```
## Warning: Expected 2 pieces. Additional pieces discarded in 344 rows [1, 2, 3, 4,
## 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

Unfortunately, since there are several white spaces within the variable names for *Species*, I lose the binomial names. I use nested `ifelse()` functions to add them to the dataset.

```
penguins2$`Binomial Nomenclature` <- ifelse(penguins2$Species == "Adelie",
                                            "Pygoscelis adeliae",
                                            ifelse(penguins2$Species == "Gentoo",
                                                   "Pygoscelis papua",
                                                   "Pygoscelis antarctica"))
```

I convert the remaining string data types into factors, except the variable *Sex*, which I convert later because of errors which pop-up if I convert them now.

```
penguins2$Species <- as_factor(penguins2$Species)
penguins2$`Binomial Nomenclature` <- as_factor(penguins2$`Binomial Nomenclature`)
penguins2$Island <- as_factor(penguins2$Island)
```

I convert the values within *Sex* from uppercase to lowercase, and I capitalise the values so the data looks cleaner. I use an `ifelse()` function to do so.

```
penguins2$Sex <- ifelse(penguins2$Sex == "MALE", "Male", "Female")
```

I reorder the table so *Penguin Number* displays first. I store my new tidy dataset into a tibble I name `penguins_tidy`, and I display the last five rows using `knitr::kable()` and `tail()`. I also specify a table caption. I use `kableExtra::kable_styling()` to increase the width of the kable, and I position the output after the code chunk (see **Table 1** below).

```
penguins2 %>% relocate(`Penguin Number`, .before = Species) -> penguins_tidy

penguins_tidy %>%
  tail(5) %>%
  knitr::kable(caption = "The last five observations of the newly tidied Palmer Penguins dataset.") %>%
  kableExtra::kable_styling(latex_options = "hold_position",
                            full_width = TRUE)
```

Table 1: The last five observations of the newly tidied Palmer Penguins dataset.

Penguin Number	Species	Binomial Nomenclature	Island	Culmen Length (mm)	Culmen Depth (mm)	Flipper Length (mm)	Body Mass (g)	Sex
340	Chinstrap	Pygoscelis antarctica	Dream	55.8	19.8	207	4000	Male
341	Chinstrap	Pygoscelis antarctica	Dream	43.5	18.1	202	3400	Female
342	Chinstrap	Pygoscelis antarctica	Dream	49.6	18.2	193	3775	Male
343	Chinstrap	Pygoscelis antarctica	Dream	50.8	19.0	210	4100	Male
344	Chinstrap	Pygoscelis antarctica	Dream	50.2	18.7	198	3775	Female

Data Summary

I first convert the variable *Sex* in `penguins_tidy` into a factor. Next I use `summary()` to explore the base summary statistics of my dataset. I summarise each variable separately because when I summarise all the variables in the dataset together, the output kable looks terrible.

Note I did not create a frequency table for *Penguin Number* because it is simply a count of the number of observations in this data set, i.e. the number of penguins. I also did not create a frequency table for *Binomial Nomenclature* because a frequency table displaying the number of penguins by *Binomial Nomenclature* is the same as the frequency table displaying the number of penguins by *Species*.

Values with 'NA' are `NULL` or blank values. I include these values in the frequency tables since it is necessary to count the number of `NULL` values. But the `summary()` function by definition does not include `NULL` values when calculating summary statistics. I eventually remove these values because they interfere with the visualisation of my data.

I display the kables using the same functions and similar parameters as above. However, for the quantitative data, I use an additional function, the built-in `R` function `as.array()` because the kables do not render without this extra function.

```
penguins_tidy$Sex <- as_factor(penguins_tidy$Sex)

penguins_tidy$Species %>%
  summary() %>%
  knitr::kable(caption = "Frequency of penguins by Species.") %>%
  kableExtra::kable_styling(latex_options = "hold_position")
```

Table 2: Frequency of penguins by Species.

	x
Adelie	152
Gentoo	124
Chinstrap	68

```
penguins_tidy$Island %>%
  summary() %>%
  knitr::kable(caption = "Frequency of penguins per Island.") %>%
  kableExtra::kable_styling(latex_options = "hold_position")
```

Table 3: Frequency of penguins per Island.

	x
Torgersen	52
Biscoe	168
Dream	124

```
penguins_tidy$Sex %>%
  summary() %>%
  knitr::kable(caption = "Frequency of penguins by Sex.") %>%
  kableExtra::kable_styling(latex_options = "hold_position")
```

```
penguins_tidy$`Culmen Length (mm)` %>%
  summary() %>%
  as.array() %>%
  knitr::kable(caption = "Summary statistics of penguins in terms of Culmen Length (mm)") %>%
  kableExtra::kable_styling(latex_options = "hold_position")
```

Table 4: Frequency of penguins by Sex.

	x
Male	168
Female	165
NA's	11

Table 5: Summary statistics of penguins in terms of Culmen Length (mm)

Var1	Freq
Min.	32.10000
1st Qu.	39.22500
Median	44.45000
Mean	43.92193
3rd Qu.	48.50000
Max.	59.60000
NA's	2.00000

```
penguins_tidy$`Culmen Depth (mm)` %>%
  summary() %>%
  as.array() %>%
  knitr::kable(caption =
                 "Summary statistics of penguins in terms of Culmen Depth (mm)") %>%
  kableExtra::kable_styling(latex_options = "hold_position")
```

Table 6: Summary statistics of penguins in terms of Culmen Depth (mm)

Var1	Freq
Min.	13.10000
1st Qu.	15.60000
Median	17.30000
Mean	17.15117
3rd Qu.	18.70000
Max.	21.50000
NA's	2.00000

```
penguins_tidy$`Flipper Length (mm)` %>%
  summary() %>%
  as.array() %>%
  knitr::kable(caption =
                 "Summary statistics of penguins in terms of Flipper Length (mm)") %>%
  kableExtra::kable_styling(latex_options = "hold_position")
```

```
penguins_tidy$`Body Mass (g)` %>%
  summary() %>%
  as.array() %>%
  knitr::kable(caption =
                 "Summary statistics of penguins in terms of Body Mass (g)") %>%
  kableExtra::kable_styling(latex_options = "hold_position")
```

Table 7: Summary statistics of penguins in terms of Flipper Length (mm)

Var1	Freq
Min.	172.0000
1st Qu.	190.0000
Median	197.0000
Mean	200.9152
3rd Qu.	213.0000
Max.	231.0000
NA's	2.0000

Table 8: Summary statistics of penguins in terms of Body Mass (g)

Var1	Freq
Min.	2700.000
1st Qu.	3550.000
Median	4050.000
Mean	4201.754
3rd Qu.	4750.000
Max.	6300.000
NA's	2.000

In terms of *Sex*, penguins are evenly distributed, though there are several penguins for whom *Sex* is not recorded (see **Table 2**). However, the distribution of this sample of penguins varies quite heavily by *Species* and *Island*: there are almost double the number of Adelie and Gentoo penguins than the number of Chinstrap penguins (see **Table 3**). Additionally, penguins on Biscoe and Dream Islands are represented much more than penguins on Torgersen Island, by at least a factor of two (see **Table 4**).

There is a lot of variation in the values for *Culmen Length*, *Culmen Depth*, *Flipper Length*, and *Body Mass* in this data set both in terms of the overall ranges and the inter-quartile ranges (see **Tables 5 to 8**). I expect a large part of this variation to be caused by the effect of *Sex* and *Species*. However, the best way of verifying this is to visualise the variation and covariation of the variables in the dataset.

Exploration of Variation

While the tidied `Palmer Penguins` dataset is not as large or as extensive as some other datasets, I only include a select few plots of the variation within my variables. First, I visualise variation within the categorical variables of this dataset using bar charts. Second, I visualise variation within the numeric variables of this dataset using box plots, density plots, and histograms.

From calculating summary statistics, there are eleven penguins with no recorded values for *Sex*, including two penguins with no recorded values in all other numeric variables. Therefore, I use the `!is.na()` argument within a `filter()` function, specifying *Sex*, to remove these eleven penguins. This leaves 333 observations to work with, which I store in `penguins_tidy2`. This is a necessary step because I separate many of the plots by *Sex* and *Species*.

```
penguins_tidy %>%
  filter(!is.na(`Sex`)) -> penguins_tidy2
```

Plotting categorical variables

The categorical (or qualitative) variables within the tidied `Palmer Penguins` dataset are *Species*, *Sex*, and *Island*. I plot bar charts of the means of the numeric variables on the Y-Axis and some of the categorical variables on the X-Axis.

For the first plot (**Figure 1**), I group my data by *Island*, *Species*, and *Sex*, and I calculate *Mean Culmen Length (mm)* for each sub-group. I store this new summary in `plot1`. I use `plot1` to plot a grouped column graph with *Mean Culmen Length (mm)* on the Y-Axis, *Species* on the X-Axis, and *Island* as the variable by which I group the bars. I separate the bars instead of stacking them using the `position = position_dodge()` argument in `geom_col()`. I include this plot despite its non-standard nature because it is informative.

There are minor differences in *Mean Culmen Length (mm)* between Chinstrap and Gentoo penguins, but much greater differences in *Mean Culmen Length (mm)* between Adelie penguins and the other penguins. Differences in *Mean Culmen Length (mm)* of Adelie penguins spread across all three islands are minor.

```
penguins_tidy2 %>%
   group_by(Island, Species, Sex) %>%
   summarise(`Culmen Length (mm)` = mean(`Culmen Length (mm)`)) -> plot1

ggplot(plot1, aes(x=Species, y=`Culmen Length (mm)`, fill = fct_rev(Island))) +
  geom_col(position = position_dodge()) +
  xlab("Penguin Species") +
  labs(fill='Island') +
  scale_y_continuous(breaks=seq(0,60,10), limits = c(0,60))-> fig1

fig1
```

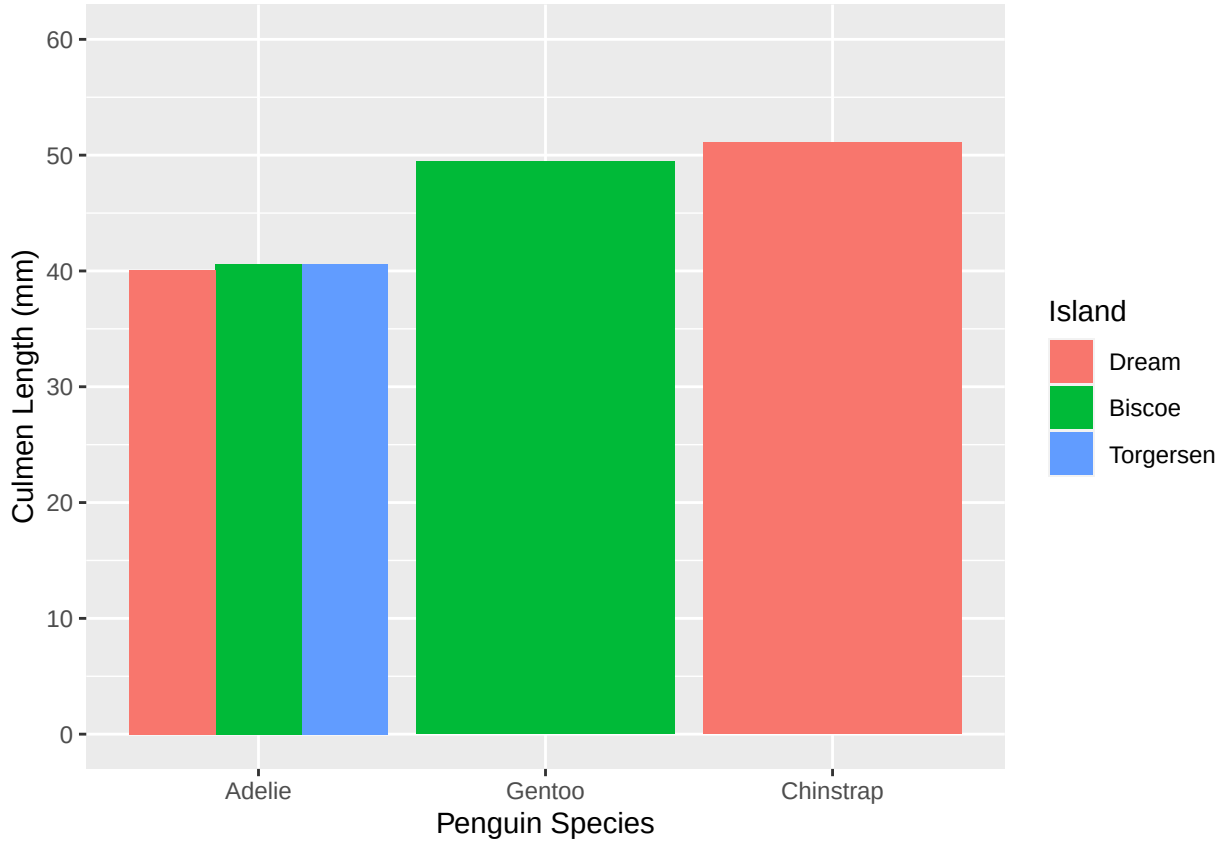

Figure 1: Bar chart of Mean Culmen Length (mm) of three Penguin Species across three Antarctic islands

For the second plot (**Figure 2**), I plot a more formal column graph. Again, I group my data by *Species*, and *Sex*, but not *Island*. I calculate *Mean Body Mass (g)* for each sub-group, and I store this summary in `plot2`. I use `plot2` to plot a grouped column graph with *Mean Body Mass (g)* on the Y-Axis, *Species* on the X-Axis, and *Sex* as the variable by which I group the bars.

There are major differences in *Mean Body Mass (g)* between sexes for all penguin species - Males are heavier than females. Gentoo penguins are heavier than Adelie or Chinstrap penguins. Adelie and Chinstrap penguins are of similar average body masses.

```
penguins_tidy2 %>%
   group_by(Island, Species, Sex) %>%
   summarise(`Mean Body Mass (g)` = mean(`Body Mass (g)`)) -> plot2

ggplot(plot2, aes(x=Species, y=`Mean Body Mass (g)`, fill =Sex)) +
  geom_col(position = position_dodge())  +
  xlab("Penguin Species") +
  labs(fill='Sex') +
  scale_x_discrete(name ="Penguin Species", limits=c("Gentoo","Adelie","Chinstrap")) +
  scale_y_continuous(breaks=seq(0,6000,500), limits = c(0,6000), labels = scales::comma)-> fig2

fig2
```

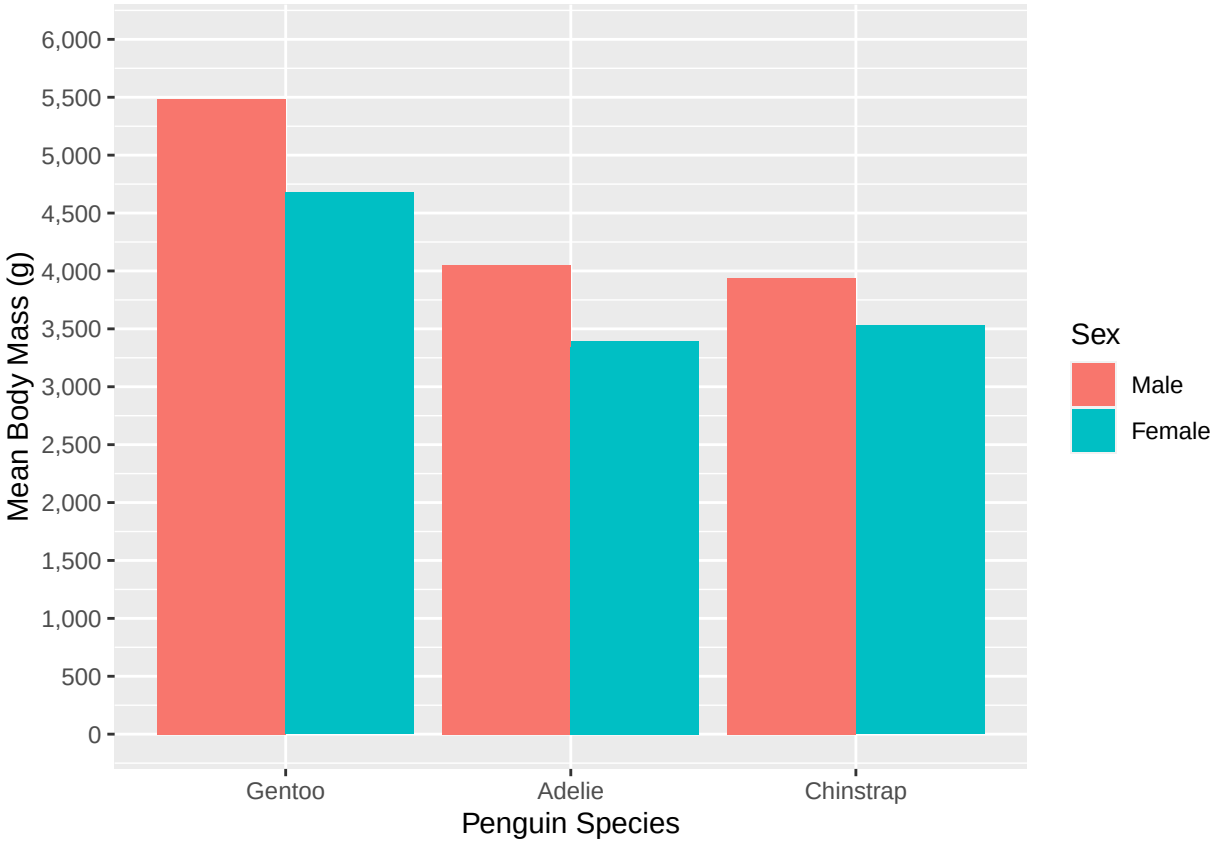

Figure 2: Bar chart of Mean Body Mass (g) of three Penguin Species separated by Sex

Plotting numeric variables

The numeric (or quantitative) variables within the tidied `Palmer Penguins` dataset are *Culmen Length (mm)*, *Culmen Depth (mm)*, *Flipper Length (mm)*, and *Body Mass (g)*. I plot a box plot of *Species* on the Y-Axis and *Flipper Length (mm)* on the X-Axis. I plot density plots of *Culmen Depth (mm)* by *Species*. I plot histograms of *Culmen Length (mm)* by *Species*. I separate all plots by *Sex*.

For the box plots in **Figure 3**, I include error bars and remove outliers to make the data cleaner and more presentable. I organise the plot from shortest *Flipper Length (mm)* to longest.

Generally, male penguins have longer flippers than their female counterparts. Males from all species display the greatest variability in flipper lengths than their female counterparts, with the exception of Chinstrap penguins, where I observe the opposite. Adelie penguins have by far the shortest flipper lengths, but the range of male Adelie flipper lengths includes and exceeds the flipper lengths of female Chinstraps.

```
penguins_tidy2 %>%
ggplot(aes(x = Species, y = `Flipper Length (mm)`, fill = Sex)) +
  stat_boxplot(geom ='errorbar') +
  geom_boxplot(outlier.shape = NA) +
  scale_x_discrete(name ="Penguin Species", limits=c("Gentoo","Chinstrap", "Adelie")) +
  scale_y_continuous(breaks=seq(170,240,10), limits = c(170,240)) +
  coord_flip() +
  scale_fill_brewer(limits = c("Female", "Male"), aesthetics = "fill", palette = "Set1") -> fig3
```

```
fig3
```

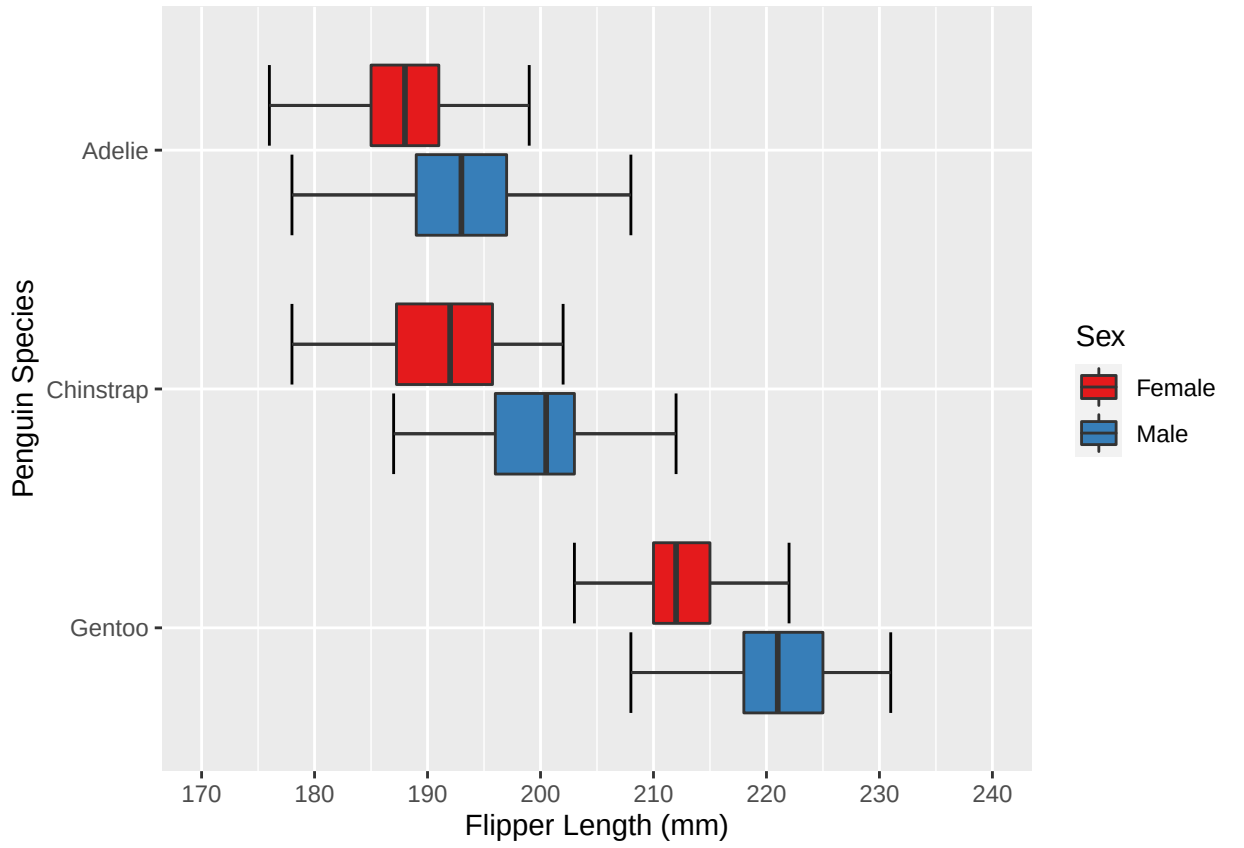

Figure 3: Box plot of Flipper Length (mm) of three Penguin Species separated by Sex

For the density plots in **Figure 4**, I plot three separate plots for *Culmen Depth (mm)* for each *Species*. By doing so, I made the data cleaner and more presentable. I organise the plot from shortest *Culmen Depth (mm)* to longest - I had to rearrange the order of factor levels within *Species* to do so.

Overall, the distributions are fairly normal. However, Gentoo penguins display a near bimodal distribution and short tails. Adelie penguins have the longest distributions, but appear to be somewhat left-skewed. Chinstrap penguins have an almost multimodal distribution, considering the large plateaus at peak densities. Adelie and Chinstrap penguins are broadly similar in terms of *Culmen Depth (mm)*. But surprisingly, Gentoo penguins have the lowest values for *Culmen Depth (mm)*, since they are the largest penguins in the data set.

```
arrange(transform(penguins_tidy2, Species=factor(Species,levels=c("Gentoo", "Adelie","Chinstrap")), Spe

plot3 %>%
ggplot(aes(x=`Culmen.Depth..mm.`, fill=Sex)) +
  geom_density(alpha=.4) +
  scale_x_continuous(breaks=seq(10,25,5), limits = c(10,25)) +
  scale_y_continuous(breaks=seq(0,0.8,0.2), limits = c(0,0.8)) +
  ylab("Density") +
  xlab("Culmen Depth (mm)") +
  facet_wrap(~Species, nrow =3) -> fig4
```

```
fig4
```

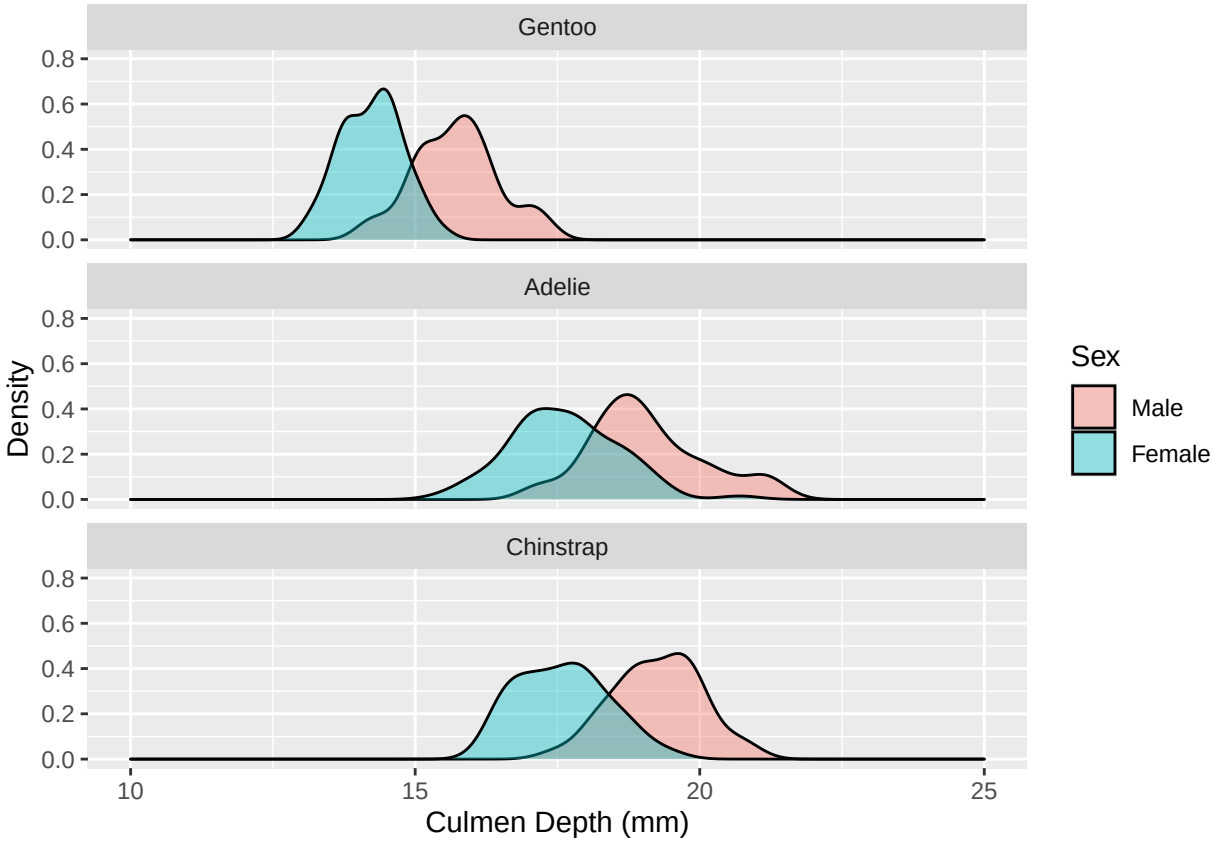

Figure 4: Density Plots of Culmen Depth (mm) of three Penguin Species separated by Sex

In **Figure 5**, I plot two histograms for *Culmen Length (mm)* by *Sex* for each *Species*. Instead of overlapping bars, I use the `position = "dodge"` argument in `geom_histogram()` to separate the bars. There are areas in the plot where all three *Species* overlap in their distributions. Unlike the box plot, I did not remove outliers.

Unsurprisingly, Adelie penguins have the shortest lengths. However, this time, Chinstrap penguins have long culmen lengths similar to their Gentoo counterparts and not their Adelie counterparts, even though Chinstrap penguins are closer to Adelie penguins in terms of all other measurements. In fact, one female Chinstrap penguin has a culmen almost the same length as a male Gentoo penguin. Even more surprising is that Gentoo penguins have long culmen lengths despite their shallow culmen depths. I specifically explore these relationships in the next section due to my findings here.

```
penguins_tidy2 %>%
  ggplot(aes(x=`Culmen Length (mm)`, fill = Species)) +
  geom_histogram(alpha = 1, bins = 50, size = 3, position = 'dodge') +
  facet_wrap(~Sex, nrow=2) +
  ylab("Count") +
  scale_x_continuous(breaks=seq(30,60,5), limits = c(30,60)) +
  scale_y_continuous(breaks=seq(0,10,1), limits = c(0,10)) +
  scale_fill_brewer(aesthetics = "fill", palette = "Set1") -> fig5
```

```
fig5
```

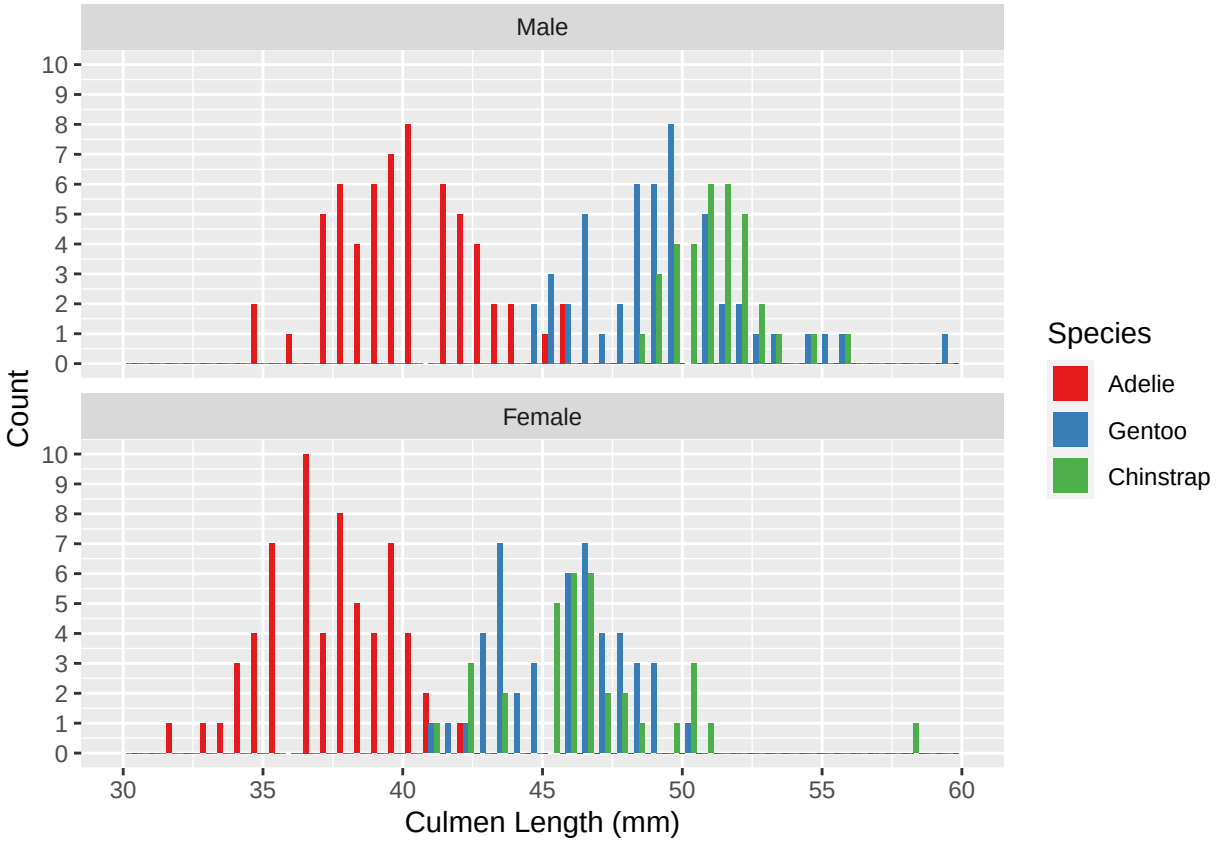

Figure 5: Histograms of Culmen Length (mm) of Male and Females across three Penguin Species

Exploration of Covariation

To examine covariation between all of my variables, I plot heat maps, a scatter plot matrix, and two scatter plots. I first plot two heat maps - one examining covariation between my categorical data, and one where I bin one of my numeric variables into deciles to transform it into categorical data. Next I plot a scatter plot matrix of my numeric variables. I select two of the scatter plots from the matrix to examine these plots further.

Plotting categorical variables

For the heat maps in **Figure 6**, I plot a relationship I first observed in **Figure 1** - the spatial distribution of penguin species in this data. The heat maps in **Figure 6** are like abstract chloropeth maps. I plot the counts of penguins of all three *Species* by *Sex* and *Island*.

As expected, Gentoo penguins are only found on Biscoe Island, Chinstrap penguins are only found on Dream Island, and Adelie penguins are evenly distributed across all three islands. More importantly, these heatmaps show that on their respective islands, Gentoo and Chinstrap penguins outnumber their Adelie counterparts, even though Gentoo and Adelie penguins outnumber Chinstrap penguins as a whole (as seen in **Table 3**).

```
penguins_tidy2 %>%
  select(Species, Sex, Island) %>%
  xtabs(~., data=.) %>%
  as_tibble() %>%
  ggplot(aes(Species, Island, fill = n)) +
  geom_tile(col = "grey35") +
  scale_fill_gradient(name="Count", low = "yellow", high = "darkred") +
  scale_x_discrete(limits=c("Gentoo","Chinstrap", "Adelie")) +
  facet_wrap(~Sex, nrow=2) -> fig6

fig6
```

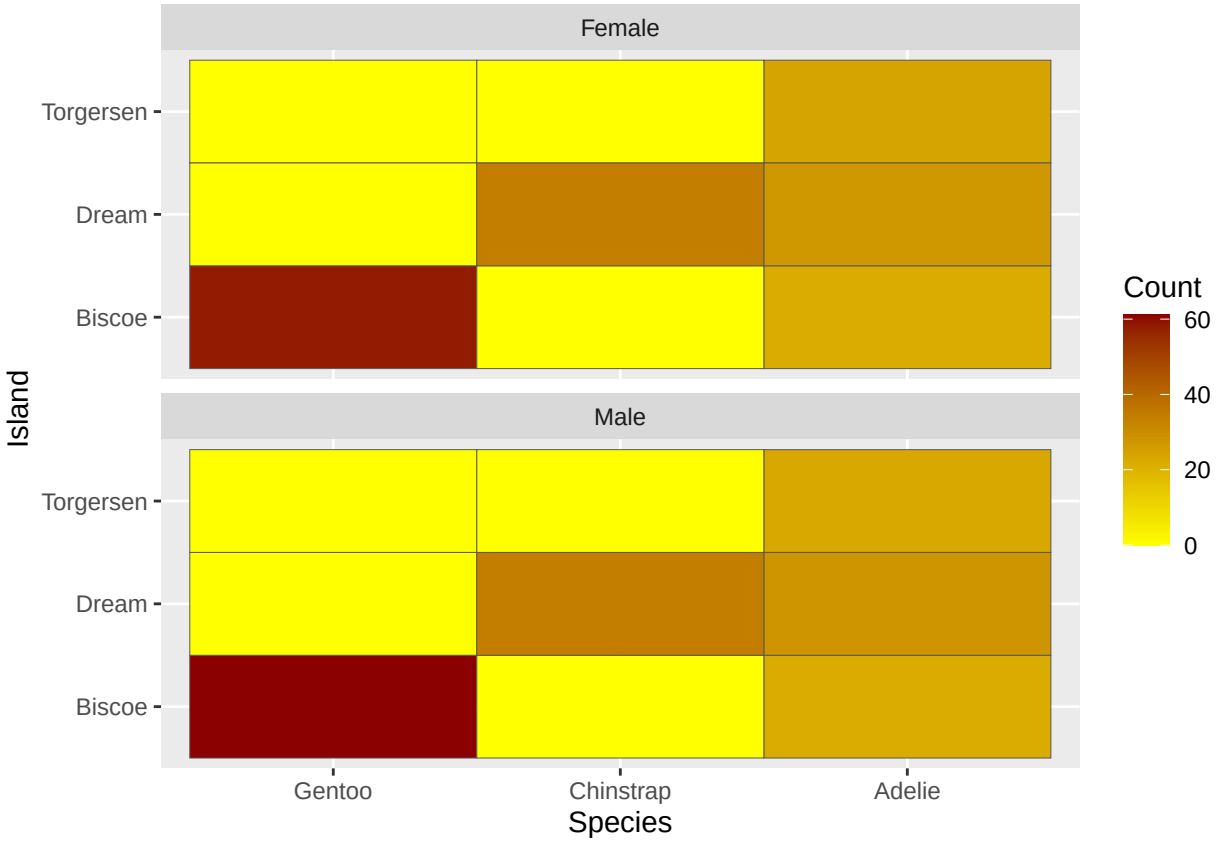

Figure 6: Heat map of Species and Island, separated by Sex

For the heat map in **Figure 7**, since I have no other categorical variables to plot, I create my own. I bin the *Culmen Depth (mm)* data into deciles, and I plot the counts of all penguins per decile by *Species*, as another way of visualising the distribution of *Culmen Depth (mm)* values.

As seen previously, Gentoo penguins are heavily over-represented in the lower deciles, whereas Adelie and Chinstrap penguins are represented more in the middle deciles.

```
order = 1:10

penguins_tidy2 %>%
  mutate(`Culmen Depth (mm) Deciles` = ntile(`Culmen Depth (mm)`, 10)) %>%
```

```
  select(Species, `Culmen Depth (mm) Deciles`, Sex) %>%
  xtabs(~., data=.) %>%
  as_tibble() %>%
  ggplot(aes(Species, factor(`Culmen Depth (mm) Deciles`, level = order), fill = n)) +
  geom_tile(col = "grey35") +
  scale_fill_gradient(name = "Count", low = "yellow", high = "darkred") +
  ylab("Culmen Depth (mm) Deciles")-> fig7

fig7
```

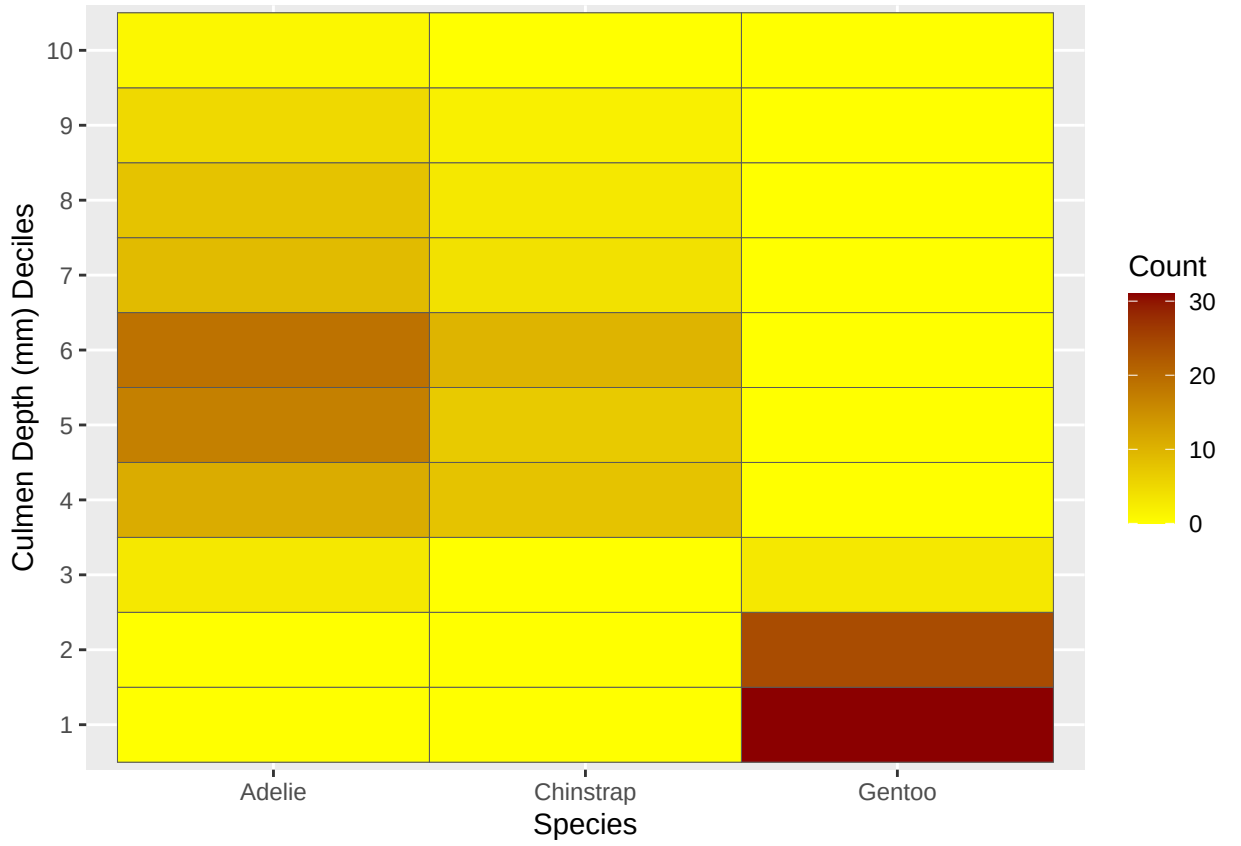

Figure 7: Heat map of Species and Culmen Depth (mm) Deciles

Plotting numeric variables

For **Figure 8**, I plot a scatterplot matrix of my four numeric variables: *Culmen Length (mm)*, *Culmen Depth (mm)*, *Flipper Length (mm)*, and *Body Mass (g)*. I colour the points by *Species* to better differentiate between my points, and to better visualise the relationships between the variables.

In a general sense, none of the linear positive correlations across and within *Species* are particularly surprising. What is interesting is the extent to which the larger Gentoo penguins differentiate themselves from the other smaller penguin species when *Culmen Depth (mm)* is plotted. Another interesting relationships is the extent to which the smaller Chinstrap penguins differentiate themselves from the smaller Adelie penguins and the larger Gentoo penguins when *Culmen Length (mm)* is plotted.

```
penguins_tidy2 %>%
select(`Culmen Length (mm)`, `Culmen Depth (mm)`, `Flipper Length (mm)`, `Body Mass (g)`) %>%
ggpairs(aes(colour = penguins_tidy2$Species, alpha = 0.3),
        lower = list(continous=wrap("points", alpha=0.3, size=0.1)))+
theme() -> fig8

fig8
```

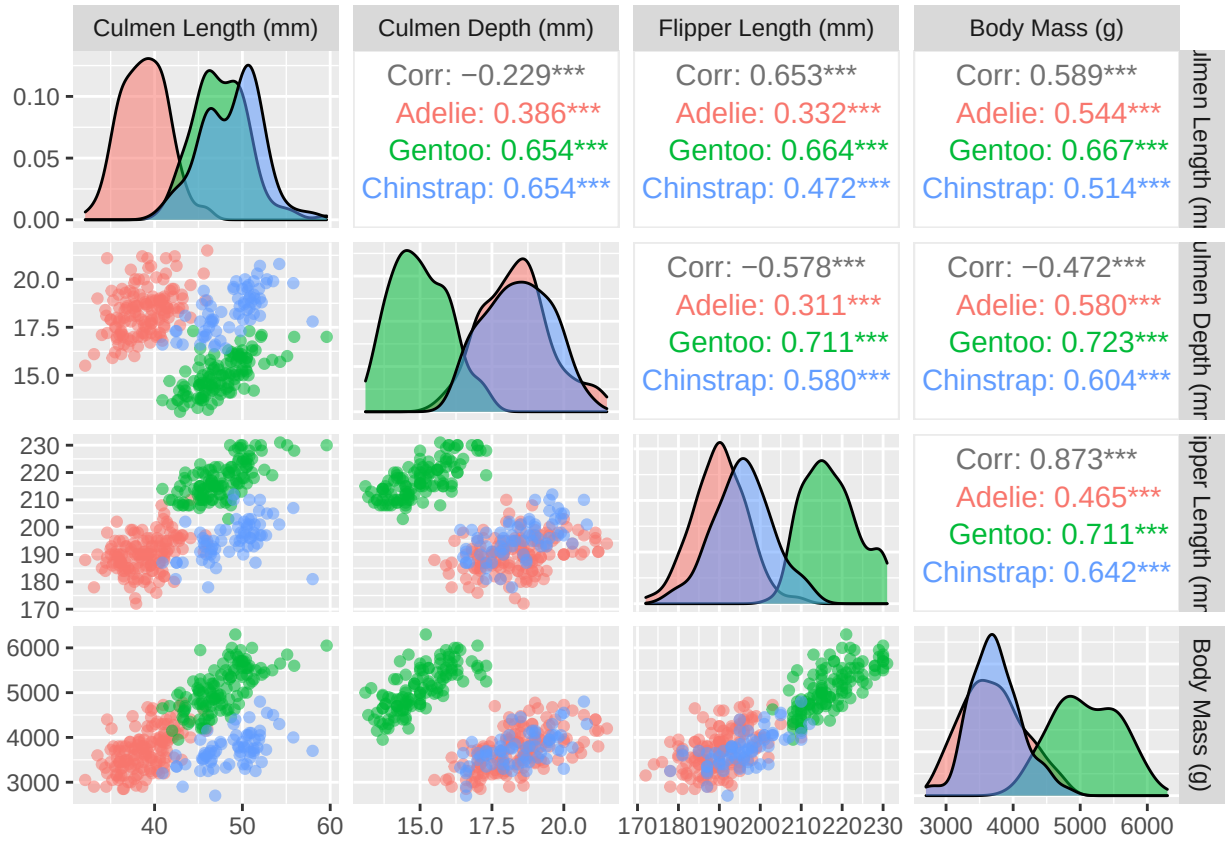

Figure 8: Scatter plot matrix of the numeric variables of interest and their correlations coloured by Species

As seen in **Figure 4** and **Figure 5**, Gentoo penguins have smaller values for *Culmen Depth (mm)* despite their larger size (and larger values for all other numeric variables in general). To examine this relationship further, I graph a scatter plot of *Culmen Depth (mm)* on the Y-Axis and *Culmen Length (mm)* on the X-Axis, and I separate the data by *Sex* (**Figure 9**).

Within each *Species* and across both *Sexes*, there is a general positive correlation, and a linear increase in *Culmen Length (mm)* and *Culmen Depth (mm)* values, with the exception of male Adelie penguins, which display a slight negative correlation and slight decrease in these values. The scatter-plot matrix in **Figure 8** obscures this particular sex-specific relationship, since it suggests an overall positive, though weak, correlation between the two variables for Adelie penguins.

```
penguins_tidy2 %>%
ggplot(aes(`Culmen Length (mm)`, `Culmen Depth (mm)`, colour = Species)) +
  geom_point(alpha = 0.4,
```

```
            size = 2, aes(shape=Species, colour = Species)) +
  geom_smooth(method = "lm", level = FALSE) +
  facet_wrap("Sex", ncol = 2) +
  scale_x_continuous(breaks=seq(30,60,5), limits = c(30,60)) +
  scale_y_continuous(breaks=seq(10,25,5), limits = c(10,25)) -> fig9

fig9
```

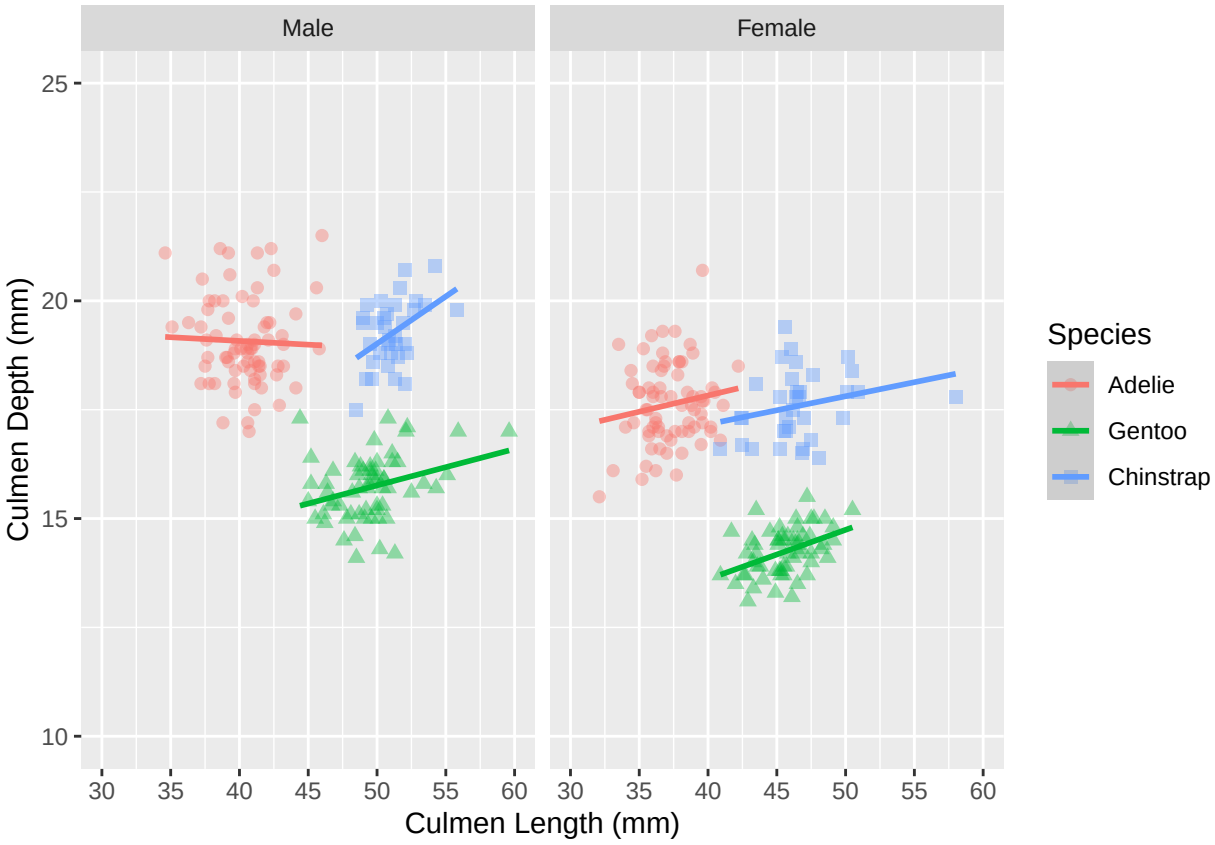

Figure 9: Scatter plots of Culmen Length (mm) and Culmen Depth (mm) of three penguin species separated by Sex

To further highlight the disparity between the larger sized Gentoo penguins and the smaller culmen depths, for **Figure 10** I plot *Culmen Depth (mm)* on the Y-Axis and *Body Mass (g)* on the X-Axis, again separating the data by *Sex*.

Within each *Species* and across both *Sexes*, there is a general positive correlation, and a linear increase in *Culmen Depth (mm)* and *Body Mass (g)* values. What is interesting about this plot is how similar Adelie and Chinstrap penguins are in terms of their *Culmen Depth (mm)* and *Body Mass (g)*. This is unsurprising: Adelie and Chinstrap penguins tend to group together across all variables. However, they do not group together when *Culmen Length (mm)* is plotted, as seen in **Figure 9**. The same can be seen in the scatter plot matrix in **Figure 8** as well. Therefore, not only do Gentoo penguins have smaller culmen depths despite their large size, Chinstrap penguins have longer culmens despite their smaller size. This suggests a useful way of differentiating between the three penguins using body measurements.

```
penguins_tidy2 %>%
ggplot(aes(`Body Mass (g)`, `Culmen Depth (mm)`, colour = Species)) +
  geom_point(alpha = 0.4,
             size = 2, aes(shape=Species, colour = Species)) +
  geom_smooth(method = "lm", level = FALSE) +
  facet_wrap("Sex", nrow = 2) +
  scale_x_continuous(breaks=seq(2500,6500,500), limits = c(2500,6500), labels = scales::comma) +
  scale_y_continuous(breaks=seq(5,25,5), limits = c(10,25)) -> fig10

fig10
```

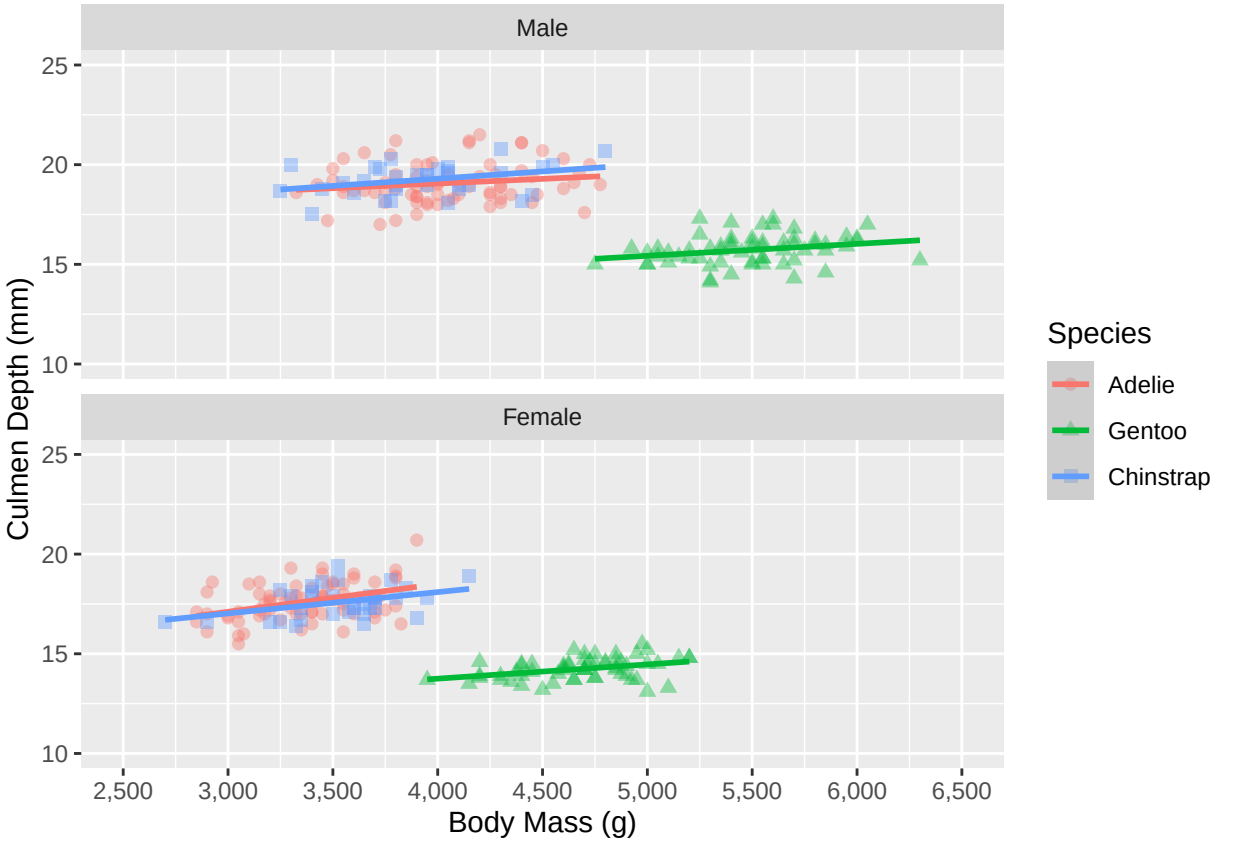

Figure 10: Scatter plots of Body Mass (g) and Culmen Depth (mm) of three penguin species separated by Sex

Discussion

Exploratory data analysis of the `Palmer Penguins` was quite useful in finding hidden distributions and relationships in the dataset. I expected to see differences in body measurements in terms of *Sex* and *Species*, and I largely found these same differences.

I wondered if there is a difference in body traits in penguins of the species living on different islands, but as seen in **Figure 1**, there were differences in the spatial distributions of all three penguin species. Adelie penguins were found across all three islands, whereas Gentoo penguins were only found on Biscoe Island, and Chinstrap penguins were only found on Dream Island. Essentially, the variation in measurements of

penguins between *Islands* was nearly the same as the variation between *Species* due to these differing spatial distributions. Visually, there was a large difference between *Mean Culmen Length (mm)* of Adelie penguins and the *Mean Culmen Length (mm)* of the other penguin species. Determining whether these differences inter-species and intra-species were statistically significant would require a paired t-test.

The distribution of certain measurements across certain *Species* was particularly interesting. As was the correlations between these measurements. Gentoo penguins had lower values for *Culmen Depth (mm)* despite their larger size (and larger values for all other numeric variables). Chinstrap penguins and Adelie penguins were incredibly similar in morphology as they generally had smaller values for almost all body measurements. However, Chinstrap penguins had much higher values for *Culmen Length (mm)* despite their smaller size.

There are probably evolutionary explanations, selective pressures behind these differences in morphology, most likely differences in diet. But explaining these differences was beyond the scope of this assignment, but may be the subject of a future analysis.

The `Palmer Penguins` dataset is a prime dataset for a classification exercise, or as a reference dataset for a classification exercise. Using *Culmen Depth (mm)*, *Culmen Length (mm)*, and *Body Mass (g)* values for penguins of unknown *Species* and *Sex*, a statistician could create and train a model to classify the unknown data by *Species* and *Sex*.

List of References

Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer Archipelago (Antarctica) |penguin data. R package version 0.1.0. https://allisonhorst.github.io/palmerpenguins/. doi: |10.5281/zenodo.3960218.