Assignment Markdown

Question 1

1a)

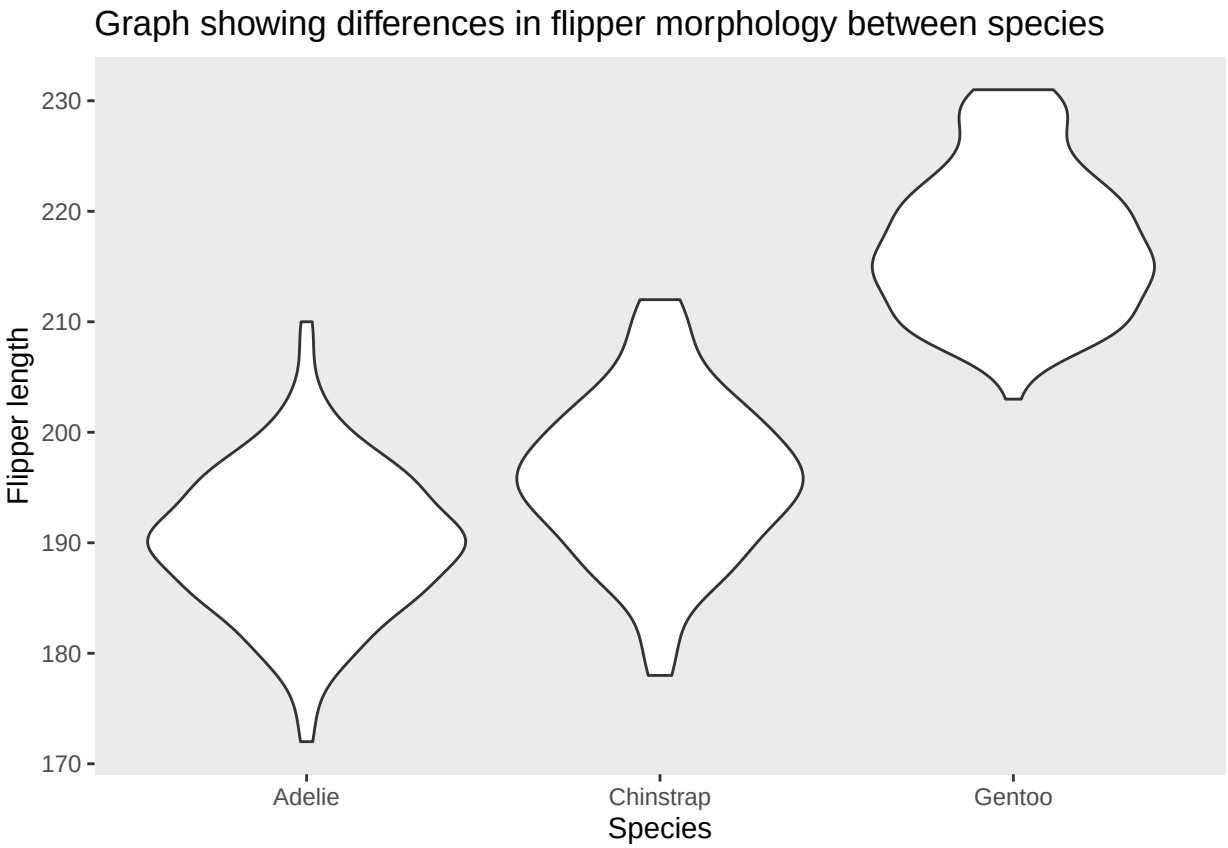

Figure 1. A "bad" figure showing flipper length distribution by species in the Palmer penguins data set.

1b) Write about how your design choices mislead the reader about the underlying data (200-300 words).

The graph does not include any units for the flipper length, meaning the readers have guess the units based on the flipper lengths they would expect to see in penguins. However this estimation is difficult as the genus of these species or the common name (penguins) is not stated anywhere in the graph.

Furthermore, the y axis is truncated, meaning it does not start at 0. This makes differences seem a lot bigger than they actually are as we have effectively zoomed into part of the graph and blown it out of proportion.

The data points are not shown, which means we don't know how many data points are present in each of the species. Though the width of the violins show the density and consequently the distribution of flipper lengths, there is no axis or scale for the width of the violin. This means we don't know how many data points lie in each part of the violin. Overlaying the data points would make this more clear. Furthermore,

without data points we cannot see whether there are any outliers and how strongly they influence the shape of the violin.

The choice of a violin plot also makes it hard to compare distributions laterally, merely due to the gap between each of the violins. This is less influential in boxplots as boxplots only differ in the y axis due to the lack of curves. Finally, there are no grid lines, meaning it is hard to see where the median and quartiles lie on the y axis.

References:

Molina, E., Viale, L., and Vázquez, P. (2022) "How should we design violin plots?," 2022 IEEE 4th Workshop on Visualization Guidelines in Research, Design, and Education (VisGuides). doi: 10.1109/VisGuides57787.2022.00006.

Question 2 - Data pipeline

Introduction.

The current analysis investigates the characteristics of different species of penguin in the dataset in the Palmerpenguins package. Specifically, we will look at whether body mass varies between three species: Chinstrap penguins (Pygoscelis antarcticus), Gentoo penguins (Pygoscelis papua) and Adelie penguins (Pygoscelis adeliae). These three species all habit Antartica and are closely related. It may be interesting to see whether their diversification involves changes in body shape. We hypothesise that the mean body masses of the species are not all the same, with differences in body shape being a product of their diversification.

Firstly, the raw data from the Palmer penguins data set was cleaned to make the column names uniform, concise, and both computer and human readable. The data set was filtered to only include body mass and species.

```
#load the function definitions
source("functions/cleaning.r")

#save the raw data as an object
write.csv(penguins_raw, "data/penguins_raw.csv")

#visualise the raw data
head(penguins_raw)
```

```
## # A tibble: 6 x 17
##   studyName `Sample Number` Species          Region Island Stage `Individual ID`
##   <chr>               <dbl> <chr>            <chr>  <chr>  <chr> <chr>
## 1 PAL0708                 1 Adelie Penguin ~ Anvers Torge~ Adul~ N1A1
## 2 PAL0708                 2 Adelie Penguin ~ Anvers Torge~ Adul~ N1A2
## 3 PAL0708                 3 Adelie Penguin ~ Anvers Torge~ Adul~ N2A1
## 4 PAL0708                 4 Adelie Penguin ~ Anvers Torge~ Adul~ N2A2
## 5 PAL0708                 5 Adelie Penguin ~ Anvers Torge~ Adul~ N3A1
## 6 PAL0708                 6 Adelie Penguin ~ Anvers Torge~ Adul~ N3A2
## # i 10 more variables: `Clutch Completion` <chr>, `Date Egg` <date>,
## #   `Culmen Length (mm)` <dbl>, `Culmen Depth (mm)` <dbl>,
## #   `Flipper Length (mm)` <dbl>, `Body Mass (g)` <dbl>, Sex <chr>,
## #   `Delta 15 N (o/oo)` <dbl>, `Delta 13 C (o/oo)` <dbl>, Comments <chr>
```

```
#create a new object with cleaned data
clean.penguins <- penguins_raw %>%
```

```
    clean_columns() %>%
  # function to remove capital letters from the column names, remove spaces in column names
  shorten_species()
#function to shorten the species names to "Adelie", "Chinstrap", and "Gentoo"

#put the clean data into a csv file
write.csv(clean.penguins, "Data/clean.penguins.csv")

#create a new object filtering out the variables that aren't needed
penguins.mass<- clean.penguins %>%
    subset_columns(c("body_mass_g", "species")) %>% #keep these variables
    remove_NA() #remove rows (samples) that have any NAs

#visualise the filtered object
head(penguins.mass)
```

```
## # A tibble: 6 x 2
##   body_mass_g species
##         <dbl> <chr>
## 1        3750 Adelie
## 2        3800 Adelie
## 3        3250 Adelie
## 4        3450 Adelie
## 5        3650 Adelie
## 6        3625 Adelie
```

A violin plot was constructed to show the distribution of body masses in the three species. This was done by creating a plot function, called from the script plotting.r present in the 'functions' folder of the project directory.

```
#plot data using a violin plot using function from plotting.r
exploratory.plot<-plot_function(penguins.mass)
exploratory.plot
```

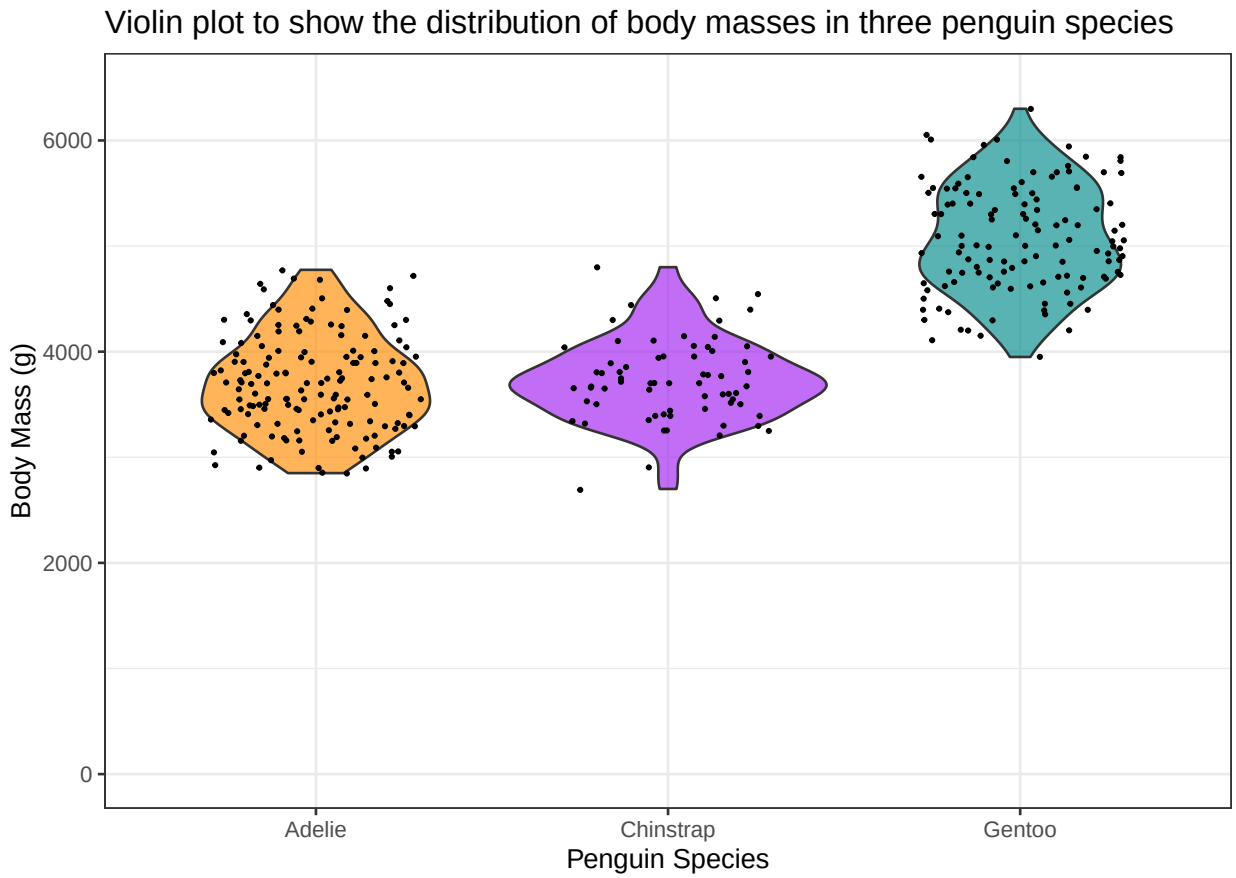

Figure 2. Exploratory violin plot showing the distribution of body masses in three species of penguins in the Palmer penguins data set. This plot shows that there is both intra- and inter-species variation of body mass in the data set. Though there is overlap in the distributions of Gentoo and Chinstrap, the widest parts of the violins (the medians) seem to be quite different. Statistical tests can be conducted to test the significance of the inter-species variation.

This plot was then saved as an .svg file as this file type has a vector format, which is resolution-independent and will thus will not lose resolution with changes in size. This can be found in the 'figures' folder in the project directory.

```
#save the figure in vector format using function called from plotting.r
save_function(penguins.mass,
              "figures/save.plot.svg",
              size=size_inches,scaling=scaling)
```

```
## pdf
##   2
```

Statistical tests

A one-way ANOVA statistical test was chosen to determine whether the difference in mean body masses between species was significant. ANOVA was used as it is a method of comparing the mean of a continuous variable across more than two categories.

Null hypothesis (H0) : the mean body mass is equal across all three species. Alternative hypothesis (H1) : the mean body mass is significantly different in at least one of the species.

```r
#fit a linear model to the data
penguins.lm<-lm(body_mass_g~species,penguins.mass)

#use ANOVA function
anova.result<-anova(penguins.lm) # run ANOVA on the linear model
print(anova.result) # print ANOVA results table
```

```
## Analysis of Variance Table
##
## Response: body_mass_g
##            Df    Sum Sq  Mean Sq F value    Pr(>F)
## species     2 146864214 73432107  343.63 < 2.2e-16 ***
## Residuals 339  72443483   213698
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3. Figure showing the results of the one-way ANOVA test. P<0.05, meaning there is sufficient evidence to reject the null hypothesis. This means that at least one of the species has a significantly different mean body mass than the others.

A post-hoc Tukey Kramer was conducted to determine which of the three species had a significant difference in body mass.

```r
#conduct post hoc Tukey Kramer test
tukey.test <- TukeyHSD(aov(penguins.lm))
print(tukey.test)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = penguins.lm)
##
## $species
##                         diff       lwr       upr     p adj
## Chinstrap-Adelie   32.42598 -126.5002  191.3522 0.8806666
## Gentoo-Adelie    1375.35401 1243.1786 1507.5294 0.0000000
## Gentoo-Chinstrap 1342.92802 1178.4810 1507.3750 0.0000000
```

Figure 4. Figure showing the results of a post-hoc Tukey Kramer test. There is insignificant difference in the mean body mass of Chinstrap and Adelie, and a significant difference between Gentoo and Adelie, and Gentoo and Chinstrap.

Discussion

From these results, we ask the question of why Chinstrap and Adelie species are significantly lighter than Gentoo. Such differences could be due to environmental effects, such as the availability of food in the different populations of species sampled. Higher food availability in the habitat of the Gentoo penguins could have led to this larger body mass. The differences could be a result of genetic divergence between species. If the latter is true, we might question whether smaller body mass is the ancestral state, given that two of the three species sampled have a similarly small body mass.

It is interesting to note that the distribution of body masses in the three species differed. The Chinstrap sample had a fairly concentrated distribution around the mean, relative to the other two species that had a broader distribution. This might suggest that there is stronger stabilising selective pressure on the Chinstrap sample, causing a more uniform body mass within the sample.

To conclude, we see a significantly larger body mass in the sample of Gentoo penguin species, compared to Chinstrap and Adelie samples, which had similarly small body masses.

Question 3 - Reflection on mine and partner's code

a) My GitHub Link. Link : https://github.com/palmpeng1/assignment.git

b) Partner's GitHub Link. Link : https://github.com/Mynewgithubabc/Penguin.Assignment.git

c) Running partner's code. My partner's code ran all the way through. The code was readable and easy to follow. Firstly, the names of their data frames, tibbles, and files were all self-explanatory, and human and computer readable, and before each code chunk they briefly explained the purpose of the chunk. They could have also named the chunks for even more clarity, however this is not crucial. In their data cleaning they created three functions to use in a pipe. Cleaning the data in one step makes it more clear to read as there is not lots of messy code that could confuse readers. The created functions were stored and called from separate R files in their repo, and they successfully communicated through hashes that these were created functions and where they were stored. Generally, the files that were loaded into the markdown were well labelled, however in their functions file, plotting.R is actually for saving the plots and the Graph.function.R is for plotting the graph which was a bit confusing. They also visualized the data before and after cleaning, so I could see what their functions did. The data were also saved as .csv files in the repo, which I could go and look at in more detail if the output from the chunk was hard to interpret. All of this makes it more readable.

They could have modified a few things to make it more readable and reproducible. When installing packages, they could have used an 'if' loop to only install the packages if not already installed. This saves time and computer energy when running the code, and reduces the number of lines of code. It makes it more reproducible as the packages will only be installed on the device of the person running the code if they don't already have it. The package versions used could also have been stored in another file as they might update before the reader accesses the analysis. This increases reproducibility in the future. They could also make it more readable by adding figure legends in the R markdown to describe what each figure is showing. Finally, they could have also made a function for their 'bad' plot, as they did for their scatter plot.

It would be easy to change their figure by altering their code, as they created a function for their figure, called scatter.plot.flippers(), and stored it in a separate R file. This saves the need for retyping out the entire plot code, and reduces the risk of changing the figure too much without being able to go back. This separate code could be edited to alter the figure in a version controlled way, so that changes can be reverted if they do not work. Furthermore, having a function means that the same figure could be created for a new, updated data set.

d) Reflection on my own code. My partner mentioned that though I explained what my created cleaning functions did, I did not explicitly state that I had created them; instead instead I just called the cleaning.R file at the beginning of the code. I agree that this could be confusing to the reader, as if they were unfamiliar with library R functions they might mistake them for functions that come in the packages I installed at the beginning of the code. This makes the code less readable, and makes troubleshooting a lot harder for readers who are working with the code in the future. Nonetheless, it does not make the code less reducible. This is because as long as they have called the file containing the created functions, the functions will work.

My code could also have been more readable and clear if I showed the head for the clean data frame. Though it would not affect the reproducibility of the analysis, it would help demonstrate the effect of the cleaning function on my data and why these steps needed to be taken. This is what my partner did and it was a good way of visualizing the different steps.

Through this process of writing code for other people, I learned that it is important to have a concise, clear code and to explain what I am doing in the code chunks by using the hash comments. This way my code is not confusing but is transparent, readable and understood. This means my code can be used and applied to future analysis (meaning it is reproducible), as well as being critically evaluated or developed to be more effective. I found that storing functions and figures in separate files is very beneficial. Not only does this make the code more concise, it also acts effectively as a stored, version controlled copy of the figure function that is separate from the main code.