Question2PenguinProject

2023-11

QUESTION 2: Data Pipeline

Write a data analysis pipeline in your .rmd RMarkdown file. You should be aiming to write a clear explanation of the steps as well as clear code.

Introduction:

A data analysis pipeline is where raw data is integrated from different sources, and then a number of tasks can be carried out including cleaning, filtering, transforming and moving the data, so that it can be analysed. As well as this, the data is stored and protected so that it is accessible in the future, and each stage of the data can be recorded and stored, so that previous versions can be recovered if mistakes are made.

The data used in this analysis pipeline is from the Palmer Penguins dataset, and includes information gathered on three different species of penguin. The data that will be analysed from Palmer Penguins in this pipeline includes the mass and sex of these three penguin species, not only to look at whether they differ significantly from one another, but also whether these explanatory variables of sex and species interact to have an effect on the body mass of penguins.

Loading the data

The first step in a data analysis pipeline is to load the data by using the library function: "library()". This is done because the packages that have been previously installed need to be loaded. One of these packages is "palmerpenguins" and this holds a dataset with information on three different species of penguin: Adelie, Chinstrap and Gentoo.

```
library(ggplot2) #a package used to visualise data.
library(palmerpenguins) #holds the palmerpenguin data.
library(janitor) # used for cleaning data.
```

```
##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(dplyr) #helps to efficiently manipulate data.
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Before cleaning the data, a raw data file needs to be saved. This is important because it ensures preservation for future use and also protection of the data from damage that might be made when cleaning and manipulating it.

To do this, a separate folder labelled "data" needs to be created within this project. This will allow you to save the raw data within this folder, and then work on a copy of it instead. The words within the green speech marks in this code below instructs R to save the "penguins_raw.csv" inside a the data folder.

```
write.csv(penguins_raw, "data/penguins_raw.csv")
```

Appropriately clean the data

The next step is to clean the data. Cleaning is an important process of the pipeline before data analysis as it makes the data more consistent and reliable. It does this in a number of ways including the removal of incorrect information, changing the format of the column names to make them more readable, and allowing for edits to be made to the "tabulation", which is the way the data is represented in rows and columns. The "janitor" package is useful to clean data.

First load the data from the saved raw data file:

```
penguins_raw <- read.csv("data/penguins_raw.csv")
```

It is important to look at the raw version in order to look for aspects that need changing, and identify how to appropriately clean it. To do this, the "head()" function can be used:

```
head(penguins_raw)
```

```
##   X studyName Sample.Number                             Species Region
## 1 1   PAL0708             1 Adelie Penguin (Pygoscelis adeliae) Anvers
## 2 2   PAL0708             2 Adelie Penguin (Pygoscelis adeliae) Anvers
## 3 3   PAL0708             3 Adelie Penguin (Pygoscelis adeliae) Anvers
## 4 4   PAL0708             4 Adelie Penguin (Pygoscelis adeliae) Anvers
## 5 5   PAL0708             5 Adelie Penguin (Pygoscelis adeliae) Anvers
## 6 6   PAL0708             6 Adelie Penguin (Pygoscelis adeliae) Anvers
##      Island              Stage Individual.ID Clutch.Completion   Date.Egg
## 1 Torgersen Adult, 1 Egg Stage          N1A1               Yes 2007-11-11
## 2 Torgersen Adult, 1 Egg Stage          N1A2               Yes 2007-11-11
## 3 Torgersen Adult, 1 Egg Stage          N2A1               Yes 2007-11-16
## 4 Torgersen Adult, 1 Egg Stage          N2A2               Yes 2007-11-16
## 5 Torgersen Adult, 1 Egg Stage          N3A1               Yes 2007-11-16
## 6 Torgersen Adult, 1 Egg Stage          N3A2               Yes 2007-11-16
##   Culmen.Length..mm. Culmen.Depth..mm. Flipper.Length..mm. Body.Mass..g.    Sex
## 1               39.1              18.7                 181          3750   MALE
## 2               39.5              17.4                 186          3800 FEMALE
## 3               40.3              18.0                 195          3250 FEMALE
## 4                 NA                NA                  NA            NA   <NA>
## 5               36.7              19.3                 193          3450 FEMALE
## 6               39.3              20.6                 190          3650   MALE
##   Delta.15.N..o.oo. Delta.13.C..o.oo.                     Comments
## 1                NA                NA Not enough blood for isotopes.
```

```
## 2           8.94956         -24.69454                           <NA>
## 3           8.36821         -25.33302                           <NA>
## 4                NA                NA             Adult not sampled.
## 5           8.76651         -25.32426                           <NA>
## 6           8.66496         -25.29805                           <NA>
```

Cleaning this dataset involves changing the column names to make them more readable, and also removing two columns that are not needed.

Cleaning can be done using piping, where multiple steps can be carried out at once. Piping makes these processes easier to read, and by doing all these steps at once, there is less chance of making mistakes.

This cleaning of the column names can also be made into a function which is beneficial as it can be saved for future use, and also then incorporated using piping with other functions that already exist for cleaning data.

```
clean_column_names <- function(penguins_data) {
    penguins_data %>%
        select(-starts_with("Delta")) %>%
        select(-Comments) %>%
        clean_names()
}
```

"Clean names()" is a function from the janitor package, and it removes the spaces and capital letters that are problematic for plotting figures.

The "clean_column_names" function that has been created can then be saved in a file so that it can be used in multiple analysis. This can be achieved by creating a functions sub-folder in your project, and other functions can be saved in this file including shortening of the species names and removing any empty columns or rows.

Once this is saved, the "functions" Rscript can be loaded into the RMarkdown file, so that these functions can be used on the dataset.

```
source("functions/cleaning.r")
```

All of these cleaning functions can be combined together through piping, and the new clean dataset can be labelled as "penguins_clean".

```
penguins_clean <- penguins_raw %>%
    clean_column_names() %>%
    shorten_species() %>%
    remove_empty_columns_rows()
```

The cleaned dataset "penguins_clean" can also then be saved in the data folder as its own csv file.

```
write.csv(penguins_clean, "data/penguins_clean.csv")
```

Last in this cleaning process, the data can be subsetted and filtered so that the mass, species and sex data are extracted for easy analysis. This can be done by using functions and piping as well.

The na.omit() function has been included after subsetting the data, because if it was included in the penguins_clean pipeline, this would have discarded all data with any NAs, rather than those that only lacked data for sex, species or mass.

```
mass_data <- penguins_clean %>%
    subset_columns(c("body_mass_g", "species", "sex")) %>%
  na.omit()
```

You can then use the head function to check and make sure your data for body mass and species has been correctly filtered.

```
head(mass_data)
```

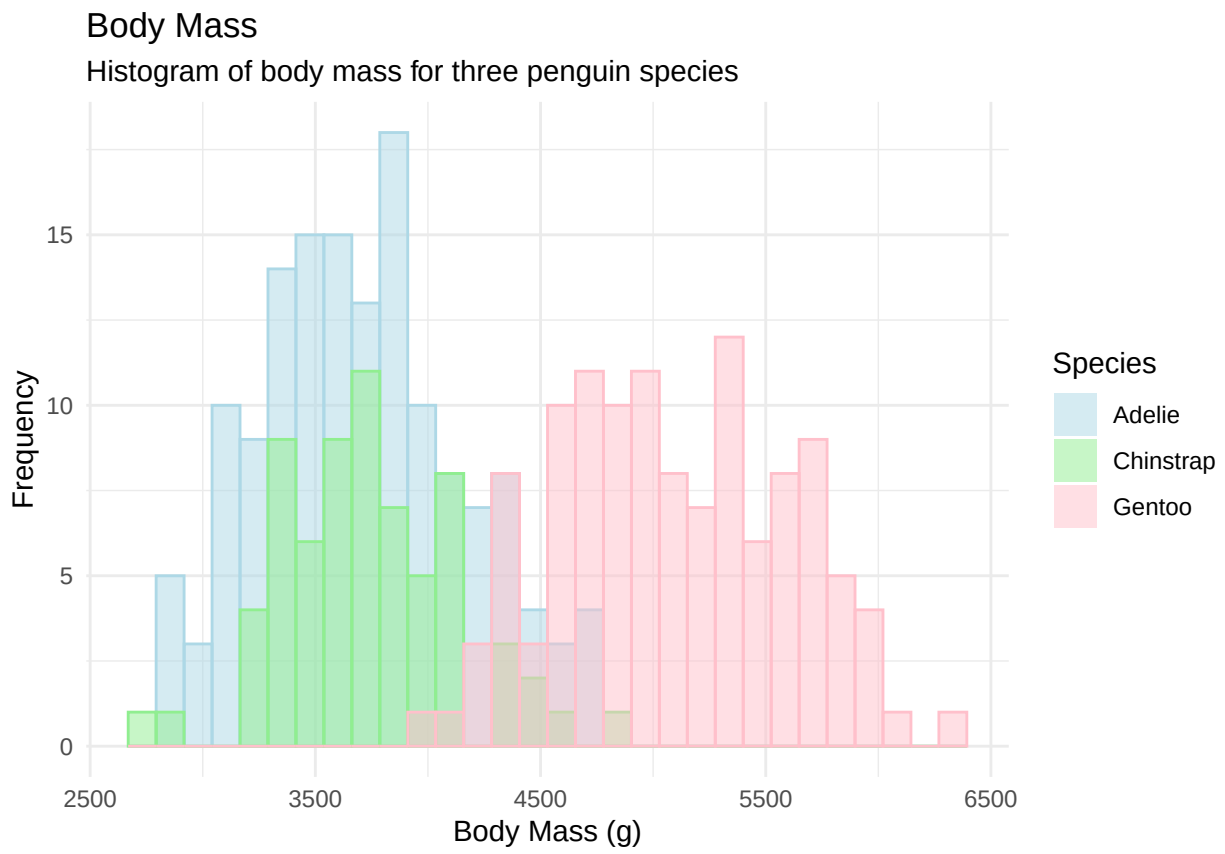
```
##   body_mass_g species    sex
## 1        3750  Adelie   MALE
## 2        3800  Adelie FEMALE
## 3        3250  Adelie FEMALE
## 5        3450  Adelie FEMALE
## 6        3650  Adelie   MALE
## 7        3625  Adelie FEMALE
```

Create an Exploratory Figure

Exploratory data analysis uses figures to investigate and show raw datasets, and one useful exploratory figure is the histogram. Histograms are useful for visually representing and summarising large datasets such as this one (Chatfield 1986). It can be useful to show the distribution of the body mass of each of these three species, and also the distribution for each sex.

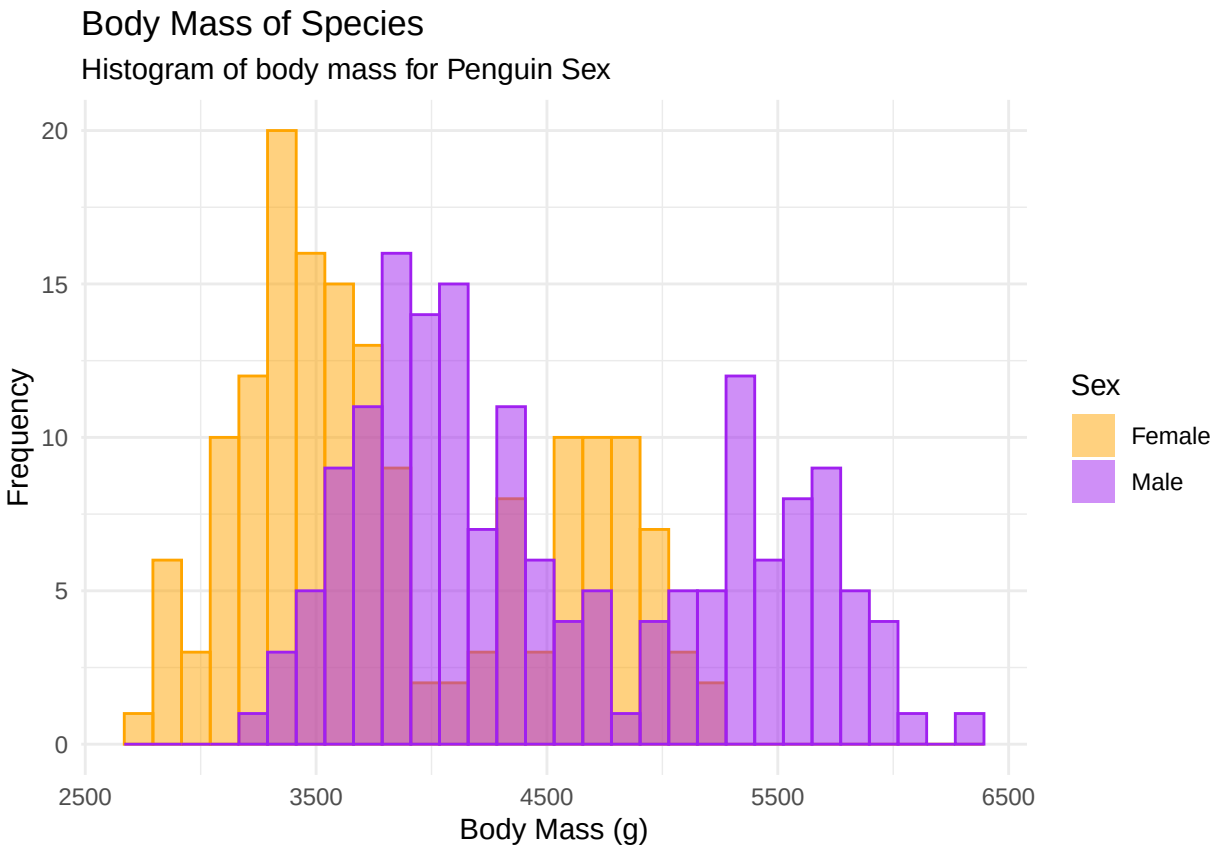
The first histogram demonstrates the distribution of the body mass for each of these species, with Adelie in blue, Chinstrap in green, and Gentoo in pink.

```
Histo_mass_species <- ggplot(data = mass_data, aes(x = body_mass_g)) +
  geom_histogram(bins = 30, aes(colour = species, fill = species),
                 alpha = 0.5, position = "identity") +
  scale_fill_manual(name = "Species",
                    values = c("lightblue", "lightgreen", "pink")) +
  scale_color_manual(values = c("lightblue", "lightgreen", "pink"))+
  labs(x = "Body Mass (g)", y = "Frequency",
       title = "Body Mass",
       subtitle = "Histogram of body mass for three penguin species") +
  theme_minimal() +theme(legend.position = "right") +
  guides(color = "none")

Histo_mass_species
```



A second histogram can be made to show how the body mass is distributed for female and male penguins (for all species). In this figure, females are shown in orange and males in purple.

```
Histo_mass_sex <- ggplot(data = mass_data, aes(x = body_mass_g)) +
  geom_histogram(bins = 30, aes(colour = sex, fill = sex),
                 alpha = 0.5, position = "identity") +
  scale_fill_manual(name = "Sex", labels = c("Female", "Male"),
                    values = c("orange", "purple")) +
  scale_color_manual(values = c("orange", "purple"))+
  labs(x = "Body Mass (g)", y = "Frequency",
       title = "Body Mass of Species",
       subtitle = "Histogram of body mass for Penguin Sex") +
  theme_minimal() +theme(legend.position = "right") +
  guides(color = "none")

Histo_mass_sex
```



It is important to save the code for these figures in a separate file, so that it can be reused in other pieces of work, and it also protects this figure from any edits made later in this script, which might damage it. Saving in multiple locations also allows for version control, so you can see how files and code are being changed over time.

To do this, first the code for these figures need to be made into a function, and then this should be saved into a separate file called "plotting.R" within the functions folder that was previously created.

```
plot_mass_species <- function(mass_data){
  mass_data %>%
  ggplot(aes(x = body_mass_g)) +
  geom_histogram(bins = 30, aes(colour = species, fill = species),
                 alpha = 0.5, position = "identity") +
  scale_fill_manual(name = "Species", values = c("lightblue", "lightgreen", "pink")) +
    scale_color_manual(values = c("lightblue", "lightgreen", "pink"))+
  labs(x = "Body Mass (g)", y = "Frequency", title = "Body Mass of Sex",
       subtitle = "Histogram of body mass for three penguin species") +
    theme_minimal() +theme(legend.position = "right") + guides(color = FALSE)
}
```

```
plot_mass_sex <- function(mass_data){
  mass_data %>%
    ggplot(aes(x = body_mass_g)) +
  geom_histogram(bins = 30, aes(colour = sex, fill = sex),
                 alpha = 0.5, position = "identity") +
  scale_fill_manual(name = "Sex", labels = c("Female", "Male"),
                    values = c("orange", "purple")) +
```

```
    scale_color_manual(values = c("orange", "purple"))+
  labs(x = "Body Mass (g)", y = "Frequency", title = "Body Mass of Species",
       subtitle = "Histogram of body mass for Penguin Sex") +
    theme_minimal() +theme(legend.position = "right") +
    guides(color = FALSE)
}
```

Saving the Figures

It is now important to save these figures, and to do this a new folder should be made called "figures".

The best way to save the figures is as .svg or .pdf figures, as they are vectors. This means that the graph can be zoomed in on and it will not go blurry. This is unlike the png. files, which are created by lots of pixels and do go blurry. The dimensions of the figure will also differ depending on how it is going to be presented.

First the library svglite needs to be loaded and the figure saved as an svg. The width and height of the figure have been specified as 6 inches (as the units for this library are inches). 6 inches has been chosen because the figure size recommended in large-sized journals with single-column text areas such as Springer cannot be larger than 17.4cm width x 23.4cm height. 6 inches is 15.24cm, so it stays below these limits.

```
library(svglite)

svglite("figures/mass_species_histogram.svg", width = 7 , height = 7 )
Histo_mass_species
dev.off()
```

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

```
svglite("figures/mass_sex_histogram.svg", width = 7 , height = 7 )

Histo_mass_sex
dev.off()
```

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

Hypothesis:

There are three null hypotheses in this analysis including:

1. When sex is controlled, the mean body mass for all three species is the same.
2. When the species is controlled, the mean body mass for both sexes is the same.
3. There is no interaction between species and sex explanatory variables. The relationship between species and body mass is not different between male and female penguins.

There are three alternative hypotheses:

1. The mean body mass for the three species of penguins differs even when sex is controlled.
2. The mean body mass for the male and female penguins differs when the species is controlled.
3. The relationship between species and body mass is different between male and female penguins.

Statistical Methods:

The statistical test that can be used to test these hypotheses is a 2-way Analysis of Variance (ANOVA). This test can be used to test if there is a significant effect of the two explanatory (independent) categorical variables of sex and species on the response (dependent) variable of penguin body mass, as well as whether these sex and species factors interact to affect the body mass. This test compares the mean differences in body mass of each species and sex, to identify if there is a significant difference, and whether there is an interaction effect between species and sex on the mean body mass.

As the sample sizes are not equal in the sex and species groups, the design for a 2-way ANOVA is unbalanced, and so a type 3 sums of squares is used. In order to do this, the "car" package needs to be installed, and the library loaded.

Run a statistical test

```
#install.packages("car")
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
penguinmodel <- aov(body_mass_g~ sex * species, data=mass_data)
Anova(penguinmodel, type = "III")
```

```
## Anova Table (Type III tests)
##
## Response: body_mass_g
##                Sum Sq  Df  F value    Pr(>F)
## (Intercept) 828480899   1 8654.649 < 2.2e-16 ***
## sex          16613442   1  173.551 < 2.2e-16 ***
## species      60350016   2  315.220 < 2.2e-16 ***
## sex:species   1676557   2    8.757 0.0001973 ***
## Residuals    31302628 327
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Results and Discussion:

This 2-way ANOVA was performed to analyse the effects of species and sex of penguins on their body mass. The results as presented in this 2-way ANOVA output table which reveals there is a statistically significant main effect of sex alone, species alone, but also a statistically significant interaction effect.

This can be seen in the "Pr(>F)" column which represents the p- value of the test. The p-value is the probability under the null hypothesis of obtaining a result equal to or more extreme than what is observed and those smaller than 0.05 are interpreted as significant, with the level of significance indicated by the stars in the "Signif. codes".

As you can see, the p values for the effects of both sex and species both have significant p-values of <2.2e-16, and this suggests that there is a main effect of both sex and species separately on the mean body mass of penguins.

However, the interaction effect is also statistically significant, with a p value of 0.000197 (3.s.f.), indicating that the relationship between species and mean body mass depends on the sex of the penguin. So there is a combined effect of species and sex on mass of the penguin. Due to this interaction effect, the main effects of sex and species cannot be interpreted on their own.

Create a Results Figure

To present these results of the interaction between sex and species, and the combined effects they have in determining the penguin mass, an interaction plot can be created.

In order to do this, first the means and standard error need to be calculated for each female and male group of each species, and this can be displayed in a table:

```
Means_SE <- mass_data %>%
  group_by(species, sex) %>%
  summarise(mass_means = mean(body_mass_g), se = sd(body_mass_g)/sqrt(length(body_mass_g)))
```

```
## `summarise()` has grouped output by 'species'. You can override using the
## `.groups` argument.
```

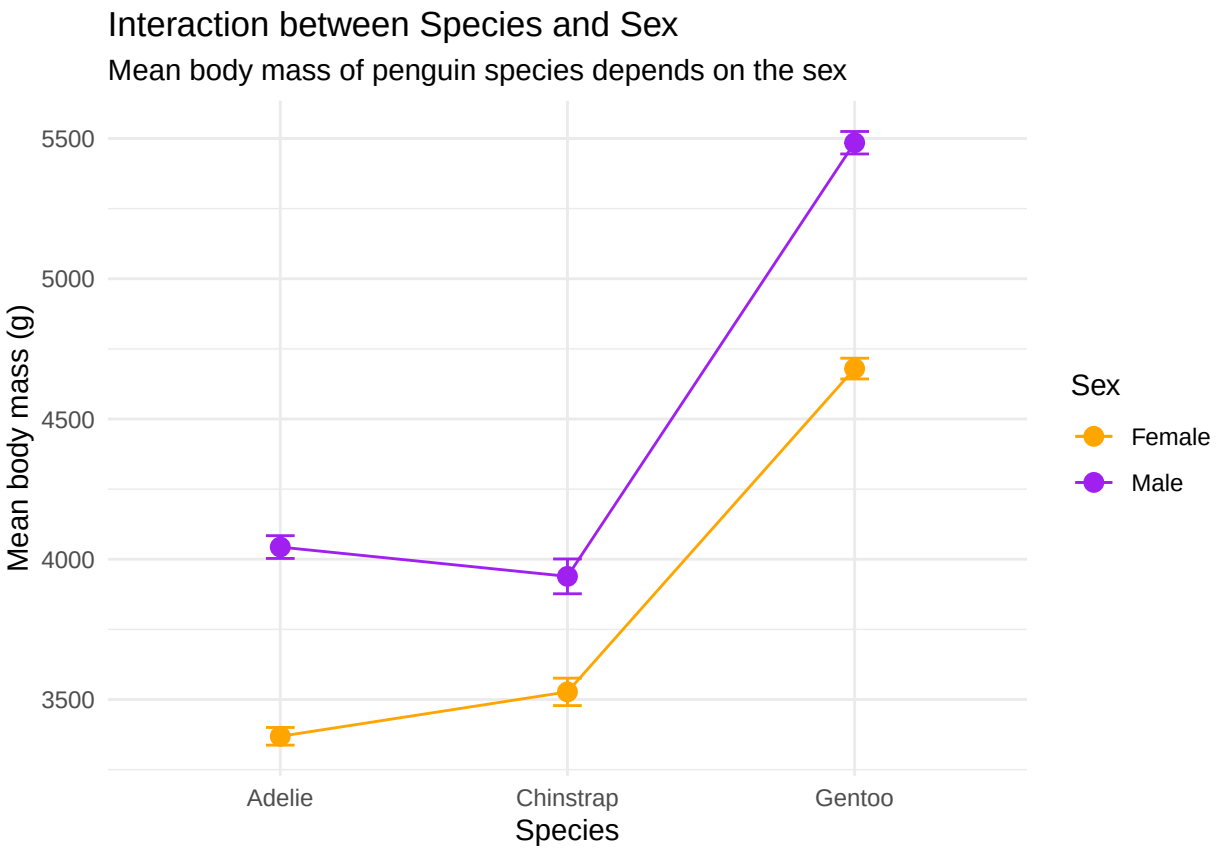
```
Means_SE
```

```
## # A tibble: 6 x 4
## # Groups:   species [3]
##   species   sex    mass_means    se
##   <chr>     <chr>       <dbl> <dbl>
## 1 Adelie    FEMALE      3369.  31.5
## 2 Adelie    MALE        4043.  40.6
## 3 Chinstrap FEMALE      3527.  48.9
## 4 Chinstrap MALE        3939.  62.1
## 5 Gentoo    FEMALE      4680.  37.0
## 6 Gentoo    MALE        5485.  40.1
```

This interaction plot can then be created. On an interaction plot, if lines remain parallel to one another, there is no interaction. Alternatively, an interaction can be visualised when the lines are not parallel to one another. The interaction effect can be seen in this figure as these lines do not run parallel to one another. The greater the interaction strength, the less parallel the lines are, and so as these lines do not overlap, it suggests there is some interaction but this is not extreme.

```
Interaction_species_sex <- ggplot(Means_SE,
                                  aes(x = species, y = mass_means, colour = sex,
                                      group = sex)) + geom_point(size = 3) +
  geom_line() + scale_fill_manual( values = c("orange", "purple")) +
  scale_colour_manual(name = "Sex", labels = c("Female", "Male"),
                      values = c("orange", "purple")) +
  labs( x = "Species", y = "Mean body mass (g)",
        title = "Interaction between Species and Sex",
        subtitle = "Mean body mass of penguin species depends on the sex ", ) +
  geom_errorbar(aes(ymax= mass_means +
                      se, ymin=mass_means-se), width=.1) +
```

```
  theme_minimal()

Interaction_species_sex
```



This interaction plot also demonstrates the main effect of sex, as the lines do not overlap one another, showing body mass of the two sexes are different for each species, with males always being of a larger mass than females for every species. As previously mentioned, interpreting the main effects when there is an interaction effect can sometimes be misleading, however as the mean mass is always higher for males, regardless of the species, it suggests that this main effect is still acting.

As well as this, the main effect of species is demonstrated as the line is not horizontal (the response mean is different for each species). In particular the line between Chinstrap and Gentoo is steep, suggesting the effect of species is strong between these two. However, male Adelie penguins are of a greater mass than male Chinstraps, but female Adelie are of smaller mass compared to female Chinstraps, and so the main effect of species here might be misleading due to the interaction.

So there is evidence to suggest there is an interaction between species and sex, suggesting sufficient evidence that this third null hypothesis can be rejected. As well as this, there is some evidence to suggest there is a main effect of sex, so the second null hypothesis that the mean body mass for each sex is the same when species is controlled has evidence for rejection. However, even though there is a significant effect of species independently on body mass, due to the interaction effect, this first null hypothesis cannot be rejected, especially when looking at Adelie and Chinstrap in particular.

This figure can also be saved as a function and placed within the plotting.R file in the functions folder:

```
plot_interaction <- function(Means_SE){
  Means_SE %>%
```

```
    ggplot(aes(x = species, y = mass_means, colour = sex, group = sex)) +
    geom_point(size = 3) +
    geom_line() + scale_fill_manual( values = c("orange", "purple")) +
    scale_colour_manual(name = "Sex", labels = c("Female", "Male"),
                        values = c("orange", "purple")) +
    labs( x = "Species", y = "Mean body mass (g)",
          title = "Interaction between Species and Sex",
          subtitle = "Mean body mass of penguin species depends on the sex ", ) +
    geom_errorbar(aes(ymax= mass_means + se, ymin=mass_means-se), width=.1) +
    theme_minimal()
}
```

Save the figure

The interaction figure can then also be saved as an svg. in the figures folder:

```
svglite("figures/interaction_plot.svg", width = 7 , height = 7 )

Interaction_species_sex

dev.off()
```

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

Conclusion:

To conclude, the body mass of penguins differs between sex and species, and there is a combined interactive effect of both of these variables on the response variable of body mass. In this project, the distribution of body mass for male and female penguins, and distribution for body mass for species has been displayed on histograms. A 2-way ANOVA was then carried out in order to identify whether there was a significant difference between the mean body mass of male and female penguins, and each species of penguin, and whether there was an interaction effect between these two categorical explanatory variables. This was then presented in an interaction plot. Results showed that there is a significant interaction effect on the mean body mass, as well as a main effect of sex, with females always having a lower body mass than males. Body mass also differs between species, but the main effect of this might be misleading due to the interaction effects.

References:

- Baker, Monya. "1,500 Scientists Lift the Lid on Reproducibility." Nature, vol. 533, no. 7604, May 2016, pp. 452–454, www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970, https://doi.org/10.1038/533452a.
- Cairo, Alberto. "Graphics Lies, Misleading Visuals." New Challenges for Data Design, by Alberto Cairo, 28 Dec. 2014, pp. 103–116.
- Chatfield, Chris. "Exploratory Data Analysis." European Journal of Operational Research, vol. 23, no. 1, Jan. 1986, pp. 5–13, https://doi.org/10.1016/0377-2217(86)90209-2.
- Tuncel, Altug, and Ali Atan. "How to Clearly Articulate Results and Construct Tables and Figures in a Scientific Paper?" Türk Üroloji Dergisi/Turkish Journal of Urology, vol. 39, no. 1, 15 Oct. 2014, pp. 16–19, www.ncbi.nlm.nih.gov/pmc/articles/PMC4548571/, https://doi.org/10.5152/tud.2013.048.

- Wagner, Clifford H. “Simpson’s Paradox in Real Life.” The American Statistician, vol. 36, no. 1, Feb. 1982, pp. 46–48, https://doi.org/10.1080/00031305.1982.10482778.
- Wainer, Howard. Visual Revelations. Link.springer.com, Copernicus New York, 1997, link.springer.com/book/97814612 Accessed 15 Nov. 2023.
- Wanzer, Dana Linnell, et al. “The Role of Titles in Enhancing Data Visualization.” Evaluation and Program Planning, vol. 84, Feb. 2021, p. 101896, https://doi.org/10.1016/j.evalprogplan.2020.101896.