Question 2: Data Pipeline

Introduction

Penguins are a group of flightless birds that inhabit almost exclusively the Southern hemisphere. There are 6 modern genera of penguins. The genus *Pygoscelis* (brushed-tailed penguins) includes 3 extant species - Adelie, Chinstrap and Gentoo. While Gentoo penguins tend to forage in deep water, Adelie and Chinstrap penguins are shallow divers. This suggests that they might have access to different food sources, which might have consequently influenced their bill morphology. Therefore, it is of interest to see how bill length is different in the 3 species. I will also consider penguin sex, to assess if there is a difference between males and females.

To do this, I will perform a two-way ANOVA, with species as one explanatory variable, and sex as the other (both categorical). The response variable is bill length (numerical). The model will allow us to test whether average bill length varies by sex and species.

But first, I will explore the data by viewing its distribution using a histogram, and calculating some descriptive statistics.

```
# LOADING THE DATA:

library(palmerpenguins)
penguins_raw
```

```
## # A tibble: 344 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
##  7 PAL0708                 7 Adelie Penguin~ Anvers Torge~ Adul~ N4A1
##  8 PAL0708                 8 Adelie Penguin~ Anvers Torge~ Adul~ N4A2
##  9 PAL0708                 9 Adelie Penguin~ Anvers Torge~ Adul~ N5A1
## 10 PAL0708                10 Adelie Penguin~ Anvers Torge~ Adul~ N5A2
## # i 334 more rows
## # 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>
```

```
library(ggplot2) #package for plotting
library(dplyr) #package for piping
```

```
##
## 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
```

```
library(janitor) #package for cleaning
```

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

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

```
library(tinytex) #package for knitting a pdf

## Saving the raw data in the project folder "data":
write.csv(penguins_raw, "data/penguins_raw.csv")

## Loading the raw data from the project folder "data":
penguins_raw <- read.csv("data/penguins_raw.csv")
```

Now that I have loaded the data, I will proceed to "clean" it - remove unnecessary columns, rename columns using suitable names, and remove NAs.

```
# CLEANING THE DATA:

## I am going to use a function that removes the species' latin names and leaves only
## their common names. This function is in the R script named "cleaning" in the
## project folder "functions", so I need to load it from there

## Loading the "cleaning" R script from the project folder "functions":
source("functions/cleaning.r")

## I will continue the data cleaning by removing the column with comments, changing
## column names to lower case and exchanging spaces for "_", and removing
## NA-containing cells of my explanatry variables:
penguins_clean <- penguins_raw %>%
  select(-Comments) %>% #removing the column with comments
  clean_names()  %>% #function in janitor to remove lower case
  species_names_short() %>% #function for shortening species names
  filter(!is.na(species), !is.na(sex)) #removing NAs
names(penguins_clean)
```

```
##  [1] "x"                  "study_name"         "sample_number"
##  [4] "species"            "region"             "island"
##  [7] "stage"              "individual_id"      "clutch_completion"
## [10] "date_egg"           "culmen_length_mm"   "culmen_depth_mm"
## [13] "flipper_length_mm"  "body_mass_g"        "sex"
## [16] "delta_15_n_o_oo"    "delta_13_c_o_oo"
## Saving the clean data in the project folder "data":
write.csv(penguins_clean, "data/penguins_clean.csv")
```

Having cleaned the data, I will now explore the dataset by calculating summary statistics, as well as observing the distribution of the data in each group using a histogram.

```
# EXPLORING THE DATA:

## A pipe for calculating summary statistics:
group_summary <- penguins_clean %>%
  group_by(species, sex) %>%
  summarize(mean_bill = mean(culmen_length_mm), median_bill = median(culmen_length_mm))
```

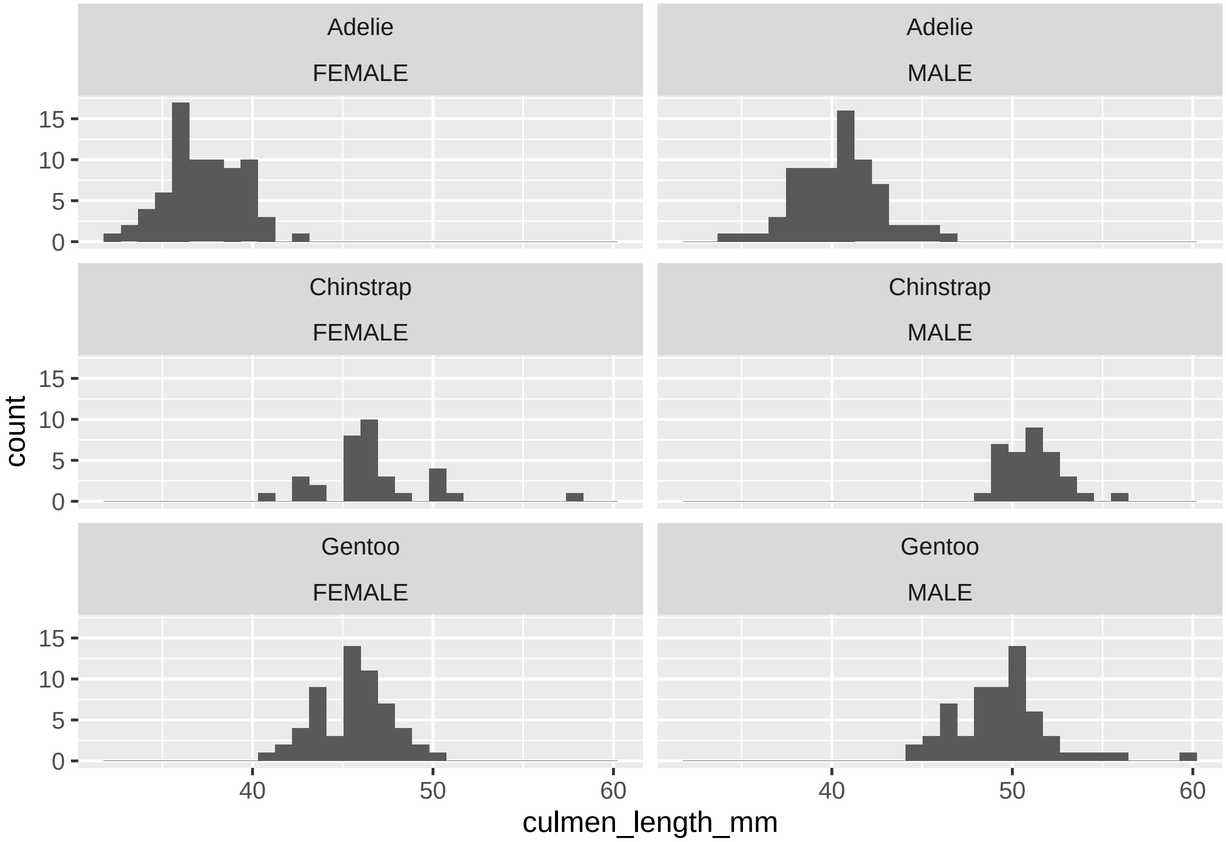
```
## ‘summarise()‘ has grouped output by ’species’. You can override using the
## ‘.groups‘ argument.
```

```
#calculating the mean and median of the data
group_summary
```

```
## # A tibble: 6 x 4
## # Groups:   species [3]
##   species   sex    mean_bill median_bill
##   <chr>     <chr>      <dbl>       <dbl>
## 1 Adelie    FEMALE      37.3        37
## 2 Adelie    MALE        40.4        40.6
## 3 Chinstrap FEMALE      46.6        46.3
## 4 Chinstrap MALE        51.1        51.0
## 5 Gentoo    FEMALE      45.6        45.5
## 6 Gentoo    MALE        49.5        49.5
```

```
## There seems to be a difference between the means and medians of the 6 groups, but a
## statistical test needs to be carried out to confirm these differences.

## Exploratory figure - distribution of the data in each group:
bill_length_hist <- ggplot(penguins_clean, aes(x=culmen_length_mm))+
  geom_histogram()+
  facet_wrap(~species*sex, ncol = 2)
bill_length_hist
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



```
## Saving the exploratory figure as a png in the project folder "plots":

library(ragg) #for producing png and other files
agg_png("plots/expfig_15x15.png", width = 15, height = 15, units = "cm", res = 600,
scaling = 1.4)
bill_length_hist
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

```
dev.off()
```

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

The data look approximately bell-shaped within each combination of levels of the explanatory variable, only with a few potential outliers. Therefore, the descriptives indicate that it can be proceeded with the analysis. In addition, the means do appear shifted with respect to each other, so it is necessary to test this properly.

Hypothesis

Hypotheses regarding main effect of the factor “species”:

- H0: The mean bill length does not differ by species;

- HA: The mean bill length of at least one of the species is different from that of at least one other species.

Hypotheses regarding main effect of the factor “sex”:

- H0: The mean bill length does not differ by sex;
- HA: The mean bill length of males and females are different from each other.

Hypotheses regarding the interaction between species and sex:

- H0: The effect of species does not depend on sex (and vice-versa);
- HA: The effect of species depends on sex (or vice-versa).

NOTE: Although the interaction hypothesis is listed last here, it will be tested first, because the result would determine if we can infer the presence of individual main effects.

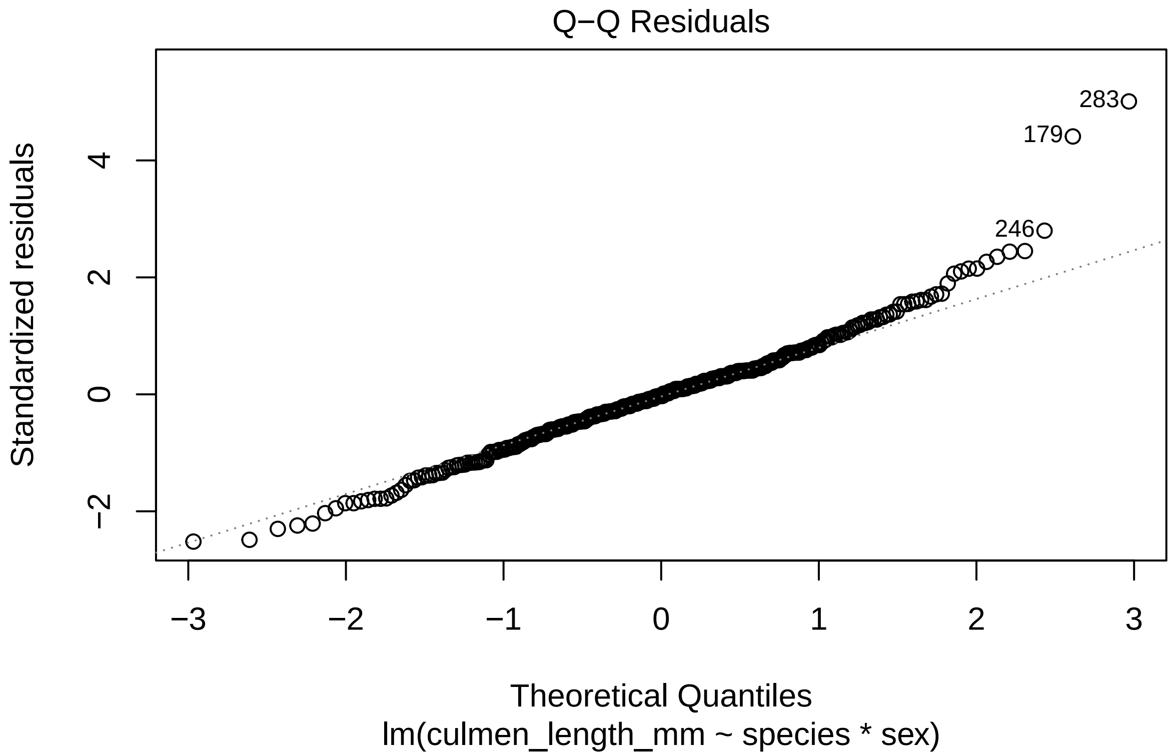
Statistical Methods

ANOVA is a parametric statistical model, so applying it involves making certain assumptions:

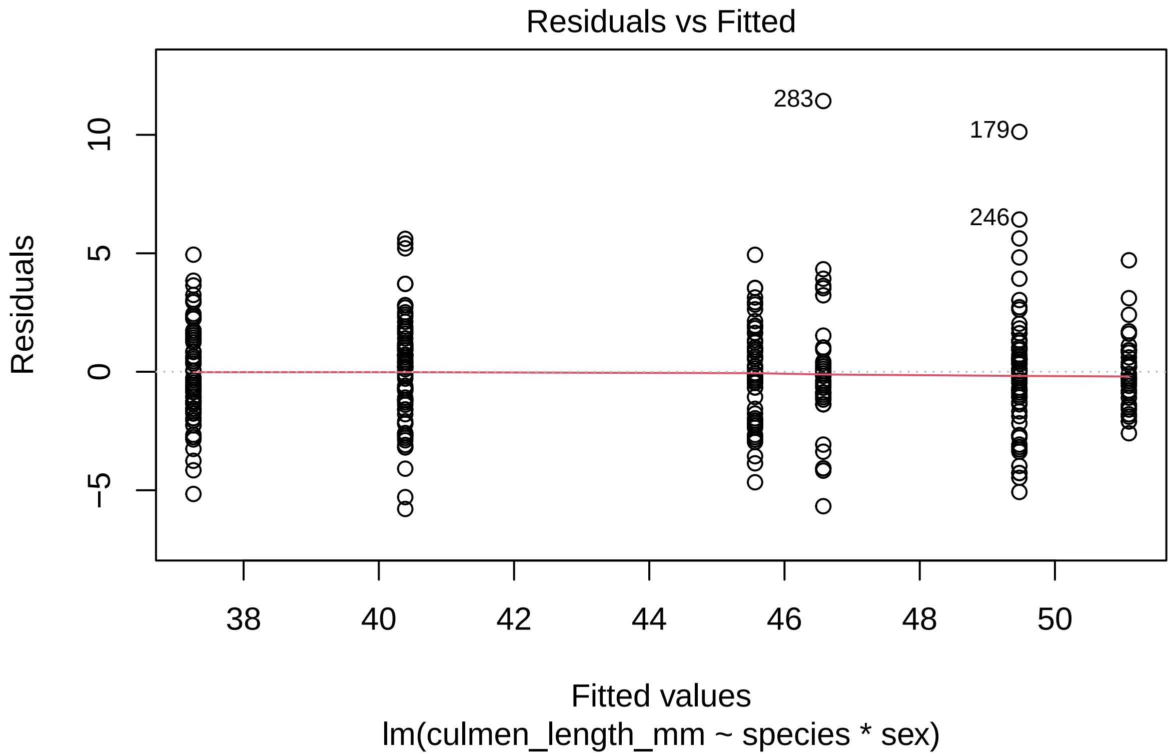
- the data are a random sample of the population;
- the observations are independent of one another;
- the response variable has a normal distribution in each combination of the categorical explanatory variables;
- the response variable has equal variance in each combination of the categorical explanatory variables.

We cannot check if the penguin data is indeed a random sample of the population/independent observations, but we could use a qqplot of residuals to check for normality, and a residuals vs fitted plot to check for equal variance.

```
# CREATING THE LINEAR MODEL - I will use the lm function to fit a linear model,
# because ANOVA is a special case of a linear model:
blength <- lm(culmen_length_mm ~ species*sex, penguins_clean)

# CHECKING THE ASSUMPTIONS (NORMALITY AND EQUAL VARIANCE)

## Checking for normality using a qqplot of residuals:
plot(blength, which = 2)
```



```
## Checking for equal variance using residuals vs fitted plot:
plot(blength, which = 1)
```



The data seem to have a roughly normal distribution (the qqplot shows approximately straight line) and a roughly equal variance (residuals are approximately symmetric around the zero line). Given the sample size, it is likely acceptable to perform an ANOVA with this data. The only concern is that there are two potential high outliers (observations 283 and 179). However, there is no additional information to believe that these are measurement errors, and since the sample size is reasonable in each cell (30 or higher) and since ANOVA has been shown to be robust to deviations from normality, I will keep these observations in the analysis.

Having checked our assumptions, I will now run the linear model and see what percentage of the variability in the response variable is explained by the model. I will then run the ANOVA to see if the interaction is significant.

- if the interaction is not significant, then I will refit the model without the interaction (to keep the model as parsimonious as possible) and see if the main effects of the 2 factors are significant;
- if the interaction is significant, I will proceed with the analysis using the original model, focusing on the main effect of one of the factors (e.g. species) within each level of the other factor (e.g. sex).

Results

```
# RUNNING THE MODEL

## Model summary - provides all the coefficients and the % of the variability in the
## response variable that is explained by the explanatory variables in the models.
## It also provides the overall F-test whether the explanatory variables explain
## significant proportion of the variability in the outcome:
summary(blength)
```

```
##
## Call:
## lm(formula = culmen_length_mm ~ species * sex, data = penguins_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7904 -1.3735 -0.0638  1.2096 11.4265
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 37.2575     0.2710 137.473  < 2e-16 ***
## speciesChinstrap             9.3160     0.4808  19.377  < 2e-16 ***
## speciesGentoo                8.3063     0.4073  20.393  < 2e-16 ***
## sexMALE                      3.1329     0.3833   8.174 6.64e-15 ***
## speciesChinstrap:sexMALE     1.3877     0.6799   2.041   0.0421 *
## speciesGentoo:sexMALE        0.7771     0.5721   1.358   0.1753
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.316 on 327 degrees of freedom
## Multiple R-squared:  0.8234, Adjusted R-squared:  0.8207
## F-statistic:   305 on 5 and 327 DF,  p-value: < 2.2e-16
```

```
## The model explains 82% of the variability in the response variable, and the
## explanatory variables are significantly associated with the outcome: F(5,327)=305,
## p<.0001.

## Performing an ANOVA - tests if the interaction between the 2 factors is significant:
anova(blength)
```

```
## Analysis of Variance Table
##
## Response: culmen_length_mm
##              Df Sum Sq Mean Sq  F value Pr(>F)
## species       2 7015.4  3507.7 654.1894 <2e-16 ***
## sex           1 1135.7  1135.7 211.8066 <2e-16 ***
## species:sex   2   24.5    12.2   2.2841 0.1035
## Residuals   327 1753.3     5.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## CONCLUSION - F(2,327)=2.28, p=0.10 - the interaction test has a p-value of 0.10, which
## is greater than our apriori chosen threshold of 0.05. Therefore, the interaction
## is not significant, so I will refit the model with just the main effects.

# REFITTING THE MODEL
blength1 <- lm(culmen_length_mm ~ species + sex, penguins_clean)

summary(blength1)
```

```
##
## Call:
## lm(formula = culmen_length_mm ~ species + sex, data = penguins_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0869 -1.3770 -0.0709  1.2254 11.0131
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        36.9770     0.2307  160.25   <2e-16 ***
## speciesChinstrap   10.0099     0.3413   29.33   <2e-16 ***
## speciesGentoo       8.6975     0.2871   30.29   <2e-16 ***
## sexMALE             3.6939     0.2548   14.50   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.325 on 329 degrees of freedom
## Multiple R-squared:  0.8209, Adjusted R-squared:  0.8193
## F-statistic: 502.8 on 3 and 329 DF,  p-value: < 2.2e-16
```

```
## The refitted model also explains ~82% of the variability in the response
## variable (bill length), and the overall F-test is statistically significant.

## Testing if the main effects of the 2 factors are significant:
anova(blength1)
```

```
## Analysis of Variance Table
##
## Response: culmen_length_mm
##            Df Sum Sq Mean Sq F value    Pr(>F)
## species     2 7015.4  3507.7  649.12 < 2.2e-16 ***
## sex         1 1135.7  1135.7  210.17 < 2.2e-16 ***
## Residuals 329 1777.8     5.4
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
## Both main effects are significant.

## Since ANOVA in base R considers the explanatory variables sequentially (Type I),
## it means that the main effect of the first variable (in this case species) is not
## adjusted for that of the second (sex). Therefore, while the resulting ANOVA table
## is useful for inferring the main effect of the second variable (sex, which has
## been adjusted for the first (species)), we need to run the ANOVA again, but changing
## the sequence of the 2 explanatory variables. This way we will also get the proper
## adjusted test for the main effect of the variable sex (i.e., Type III sums of squares):
blength2 <- lm(culmen_length_mm ~ sex + species, penguins_clean)

summary(blength2)
```

```
##
## Call:
## lm(formula = culmen_length_mm ~ sex + species, data = penguins_clean)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0869 -1.3770 -0.0709  1.2254 11.0131
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       36.9770     0.2307  160.25   <2e-16 ***
## sexMALE            3.6939     0.2548   14.50   <2e-16 ***
## speciesChinstrap  10.0099     0.3413   29.33   <2e-16 ***
## speciesGentoo      8.6975     0.2871   30.29   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.325 on 329 degrees of freedom
## Multiple R-squared:  0.8209, Adjusted R-squared:  0.8193
## F-statistic: 502.8 on 3 and 329 DF,  p-value: < 2.2e-16
```

```
## The refitted model also explains ~82% of the variability in the
## response variable

## Testing if the main effects of the 2 factors are significant:
anova(blength2)
```

```
## Analysis of Variance Table
##
## Response: culmen_length_mm
##            Df Sum Sq Mean Sq F value    Pr(>F)
## sex         1 1175.5  1175.5  217.53 < 2.2e-16 ***
## species     2 6975.6  3487.8  645.44 < 2.2e-16 ***
## Residuals 329 1777.8     5.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## CONCUSION: the adjusted main effects of sex and of species are both statistically
## significant, because the p-values associated with these tests are <0.05.
```

We did not find an interaction between the 2 factors, so we failed to reject H0 about the presence of an interaction between the 2 factors. As for the main effects of each of the 2 factors, we rejected H0 in both cases, because the main effects of both factors proved to be statistically significant. This means that:

- the mean bill length of at least one of the species is different from that of at least one other species;
- the mean bill length of males and females are different from each other.

However, since the factor "species" has 3 levels, we do not know between which levels of the factor there is a difference. To infer this, in the next section I calculate the least squares means by species, and then compute the pairwise differences in means, averaged across sex.

```
## Loading the package for calculating least squares means:
library(lsmeans)
```

```
## Loading required package: emmeans
## The ’lsmeans’ package is now basically a front end for ’emmeans’.
## Users are encouraged to switch the rest of the way.
## See help(’transition’) for more information, including how to
## convert old ’lsmeans’ objects and scripts to work with ’emmeans’.
```

```r
## Creating an object containing the least squares means, with their standard errors (and
## confidence intervals):
spec <- lsmeans(blength1, ~species)
spec
```

```
##  species   lsmean    SE  df lower.CL upper.CL
##  Adelie      38.8 0.192 329     38.4     39.2
##  Chinstrap   48.8 0.282 329     48.3     49.4
##  Gentoo      47.5 0.213 329     47.1     47.9
##
## Results are averaged over the levels of: sex
## Confidence level used: 0.95
```

```r
## Calculating pairwise differences in means, averaged across sex, in the absence of an
## interaction; adjust for multiple comparisons using Tukey:
contrast(spec, alpha=0.05, method="pairwise", adjust="Tukey")
```

```
##  contrast           estimate    SE  df t.ratio p.value
##  Adelie - Chinstrap   -10.01 0.341 329 -29.329  <.0001
##  Adelie - Gentoo       -8.70 0.287 329 -30.293  <.0001
##  Chinstrap - Gentoo     1.31 0.353 329   3.713  0.0007
##
## Results are averaged over the levels of: sex
## P value adjustment: tukey method for comparing a family of 3 estimates
```

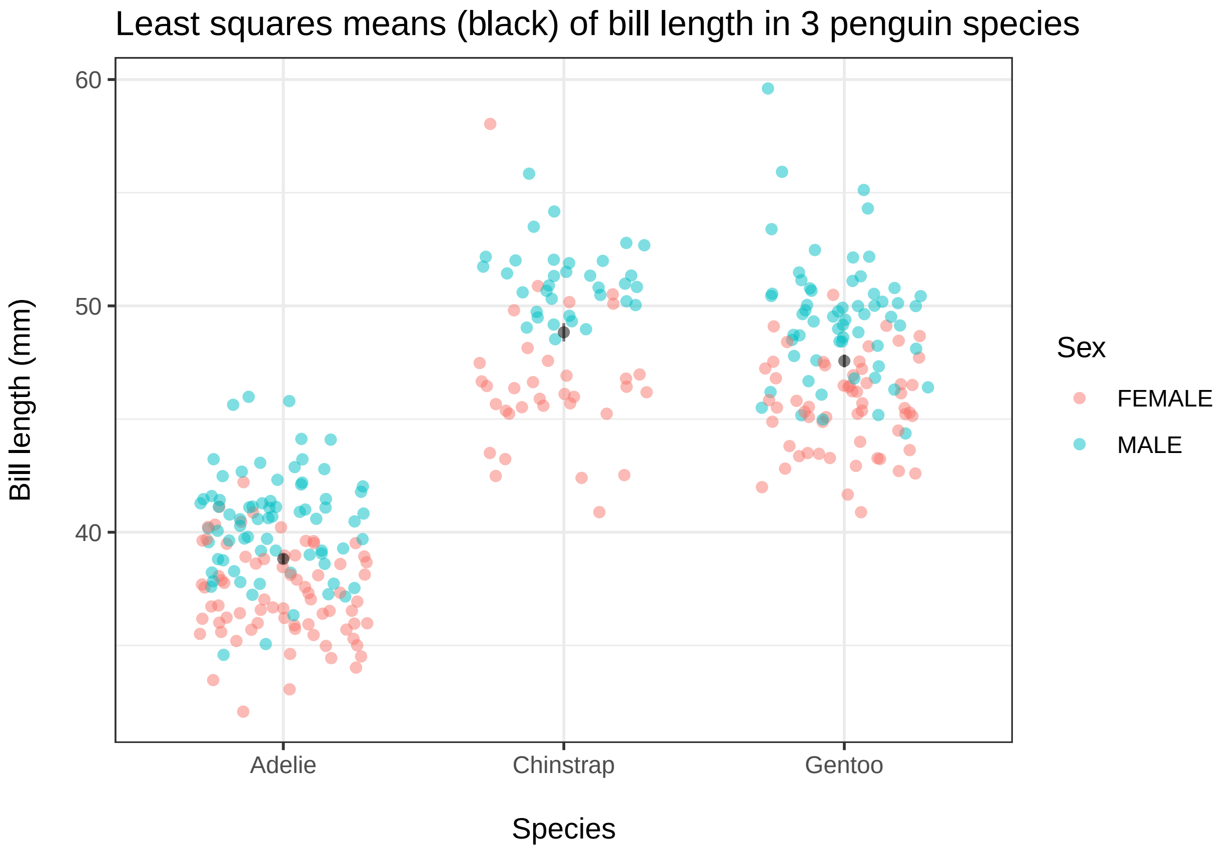
```r
## CONCLUSION: the differences in means between each pair of species is statistically
## significant.
```

Therefore, the results show that the mean bill length of each species differs from that of the remaining species. More specifically, on average Adelie penguins have shorter bills then Chinstraps, which in turn have shorter bills than Gentoos. In addition, on average females have shorter bills than males.

We can visualize the least squares means using a strip chart (dot plot), as shown in the code below.

```r
# PLOTTING A RESULTS FIGURE:

## Converting spec (the object containing the least square means) into an array:
spec1 <- summary(spec)

## Creating a function to extract the values for least squares means and confidence
## interval bounds from the spec1 array (will then be plotted on a strip chart):
mean_ci <- for(i in 1:3)
{
  m <- spec1[i,2] #mean
  ymin <- m - spec1[i,5] #lower bound
  ymax <- m + spec1[i,6] #upper bound
  return(c(y=m,ymin=ymin,ymax=ymax))
}

## I have also added this function to the "summarystats" R script of
## the "functions" folder, so I can use it in future analyses.

## Plotting the least squares means using a strip chart:
anova_means_plot <- ggplot(penguins_clean, aes(x=species, y=culmen_length_mm,
                           colour = sex))+
  geom_point(position = position_jitter(width = 0.3), alpha = 0.5)+
  #making points more visible by introducing jitter and transparency
  stat_summary(fun.data=mean_ci, color="black", alpha = 0.5, size = 0.2)+
  #adding the least squares means and confidence intervals to the plot
  labs(title = "Least squares means (black) of bill length in 3 penguin species",
       x = "\nSpecies", y = "Bill length (mm)\n", colour = "Sex")+
  theme_bw()
anova_means_plot
```

```
## No summary function supplied, defaulting to ‘mean_se()‘
```



```
## NOTE: the standard errors of the means are very close around the mean, so they
## would be difficult to see on the graph. Therefore, I made the least squares means
## point relatively transparent, and I plotted the confidence intervals instead
## of the standard errors.
```

```r
## Saving the results figure as a png in the project folder "plots":

agg_png("plots/resfig_20x15.png", width = 20, height = 15, units = "cm", res = 600,
        scaling = 1.4)
anova_means_plot
```

```
## No summary function supplied, defaulting to ‘mean_se()‘
```

```r
dev.off()
```

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

Discussion

The results show that average bill length in penguins varies both by sex and by species. More specifically, Gentoo have the longest bills, followed by Chinstrap and then Adelie. In addition, males have longer bills than females.

However, it is important to note 2 things:

- there are 2 outliers in the data (see the qqplot of residuals, and the residuals vs fitted plot);
- there could be confounding variables, such as body mass.

Future research could involve rerunning the analysis without the outliers in order to make sure that the conclusions are robust to these outliers. Furthermore, body mass could be added as an explanatory variable, in which case the ANOVA (which uses categorical explanatory variables) would be substituted with an ANCOVA (which uses both categorical and numerical explanatory variables). However, if body mass differs statistically significantly between species in the first place, using it as a covariate may yield a biased estimate of the species' differences in bill length (i.e., over-correcting the comparison). An alternative to the ANCOVA would be to run a multiple linear regression, in order to make use of other predictors in the dataset, such as flipper length and island inhabited. But in any case, more data are needed to establish what determines bill length, such as determining the diet of these 3 penguin species.

Conclusion

Bill length was shown to vary both by species (Adelie, Chinstrap, Gentoo) and by sex (male, female). However, confounding variables such as body mass, and additional information such as diet have to be considered. This could shed light on the factors that have influenced the variety of penguin bill shapes we see today.

References

- Trivelpiece, Wayne Z., et al. "Ecological Segregation of Adelie, Gentoo, and Chinstrap Penguins at King George Island, Antarctica." Ecology, vol. 68, no. 2, 1987, pp. 351–61. JSTOR, https://doi.org/10.2307/1939266. Accessed 28 Nov. 2023.
- Blanca MJ, Alarcón R, Arnau J, Bono R, Bendayan R. Non-normal data: Is ANOVA still a valid option? Psicothema. 2017 Nov;29(4):552-557. doi: 10.7334/psicothema2016.383. PMID: 29048317.

- Reference for the perils of ANCOVA: Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. Journal of Abnormal Psychology, 110(1), 40–48. https://doi.org/10.1037/0021-843X.110.1.40.
- ANOVA – type I/II/III SS explained. Univeristy of Goettingen. https://md.psych.bio.uni-goettingen.de/mv/unit/lm_cat/lm_cat_unbal_ss_explained.html.