Distinct lay dates in three species of *Pygoscelis* penguin

2024-11-30

Introduction

The genus Pygoscelis contains three different species of penguin: Adelie (*Pygoscelis adeliae*), Chinstrap (*Pygoscelis antarcticus*) and Gentoo (*Pygoscelis papua*), which occur in overlapping ranges throughout the marine ecosystem of the Western Antarctic Peninsula[1]. The penguins' breeding season begins when they return to nesting sites around October, to coincide with the start of the Antarctic summer. The partners then undergo courtship, after which they lay two eggs. After incubation and finally hatching, the parents will take turns between guarding the chick and foraging for food[2]. While the general life history pattern between species is similar, there are also some major differences between them in terms of migration patterns, fasting behaviour and diet[2]. These differences in life history likely impact the optimal lay date for each species.

Adelie penguins migrate long distances to common overwintering grounds on the ice. When they return to summer breeding grounds, they fast for three weeks during courtship and egg laying[3]. On the other hand, Gentoo penguins overwinter on the near shore and forage in the sea all year round. The female fasts for just one week before egg laying[3].

All three species forage in ice-free waters during the summer, but previous research has shown that they have different diets[4]. For example, Chinstrap penguins have a higher consumption of krill, and the fish they eat tend to come from one pelagic species and two mesopelagic species[4]. In constrast, Gentoo birds also eat krill but have a higher proportion of fish in their diet. These fish tend to be a combination of pelagic and benthic species[4].

For all birds, it is important to align the food demand of the chick with food availability in the environment. Species that are better at making this alignment should have greater reproductive success[3]. This is particularly important for penguins, not just because they must factor in a fasting period pre-laying, but also because the Antarctic is a highly seasonal environment where food availability and weather varies greatly throughout the year[5]. It has previously been shown that penguins that delay breeding from the optimum have reduced reproductive success[6] because of the limited time available to provision food to offspring before food resources are depleted[3].

Hypothesis

Given the differences between the species in migration patterns, fasting behaviour, and diet, all of which are likely to influence trade-offs for the optimal timing of egg laying, I hypothesize that the three species will have distinct egg-laying dates.

Methods

Gorman et al (2014) studied Pygoscelis penguins on the Palmer Archipelago, located west of the Antarctic Peninsula, near Anvers island[1]. This study collected morphological data on nesting penguins to study sexual dimorphism, but at the same time they recorded the clutch initiation date (CID) of each nesting pair. This data can now be freely accessed in the 'palmerpenguins' package in R. The study nests were located on three different islands - Biscoe (64° 489S, 63° 469W), Torgersen (64° 469S, 64° 049W), and Dream (64° 439S, 64°

139W). The study ran for three years from 2007-2009 and in total recorded 76 Adelie nests, 34 chinstrap nests, and 62 Gentoo nests[1].

I used this dataset to examine the CID of the different species, and look at how this varies across the three years of the study. First, I converted the calendar dates into Julian dates so that 'lay date' became a continuous variable. Julian dates give each day of the year a number from 1 to 365, starting from the 1st January. I created some exploratory figures, and then used ANOVA to test for differences in means between species. A statistically significant result from an ANOVA does not tell us which group means are different from each other, and therefore I used a Tukey-Kramer test to find out which groups have significantly different means.

```
knitr::opts_chunk$set(echo = TRUE)

#load all the required packages
library(tidyverse)#for cleaning the data
library(janitor)#for cleaning the data
library(lubridate)#for manipulating dates
library(palmerpenguins)#contains the dataset
library(here)#to specify the directory
library(ggplot2)#for making plots
library(lme4)#for making linear models
library(grid)#for making multi-panel figures
library(gridExtra)#for making multi-panel figures
library(dplyr)#for manipulating the data
library(ggsignif) #for adding significance stars to plots
```

```
#Load the raw data and save it
penguins_raw <- read_csv(here("data", "penguins_raw.csv"), show_col_types = FALSE)
write_csv(penguins_raw, here("data", "penguins_raw.csv"))
```

```
#Clean the data and save it
#The cleaning function cleans up column names, removes the 'comments' column, shortens the species name
source(here("functions", "cleaning.R"))
penguins_clean <- cleaning_penguins(penguins_raw)
write_csv(penguins_clean, here("data", "PenguinsClean.csv"))
```

```
#convert calendar dates into Julian dates

#save date in a date format
penguins_clean$date_egg <- as.Date(penguins_clean$date_egg, format = "%Y-%m-%d")

#convert to julian date
penguins_clean$julian_date <- as.integer(format(penguins_clean$date_egg, "%j"))

#Add a column called 'year_egg_laid'
penguins_clean$year_egg_laid <- format(penguins_clean$date_egg, "%Y")
```

Results

QUESTION 1: Creating a bad exploratory figure

```r
#calculate mean lay date for the three different penguin species
mean_lay_dates <- penguins_clean %>%
  group_by(species) %>%
  summarize(mean_julian_date = mean(julian_date, na.rm = TRUE))

#assign colours to the three different species
species_colours <- c("Adelie" = "darkorange",
                     "Chinstrap" = "purple",
                     "Gentoo" = "cyan4")

#make a really bad exploratory boxplot
bad_plot <- ggplot(
        data = mean_lay_dates,
        aes(
        x = factor(species),
        y = mean_julian_date,
        fill = species)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = species_colours) +
  labs(
       x = "Species",
       y = "Mean Julian Lay Date") +
  theme_bw()
bad_plot
```

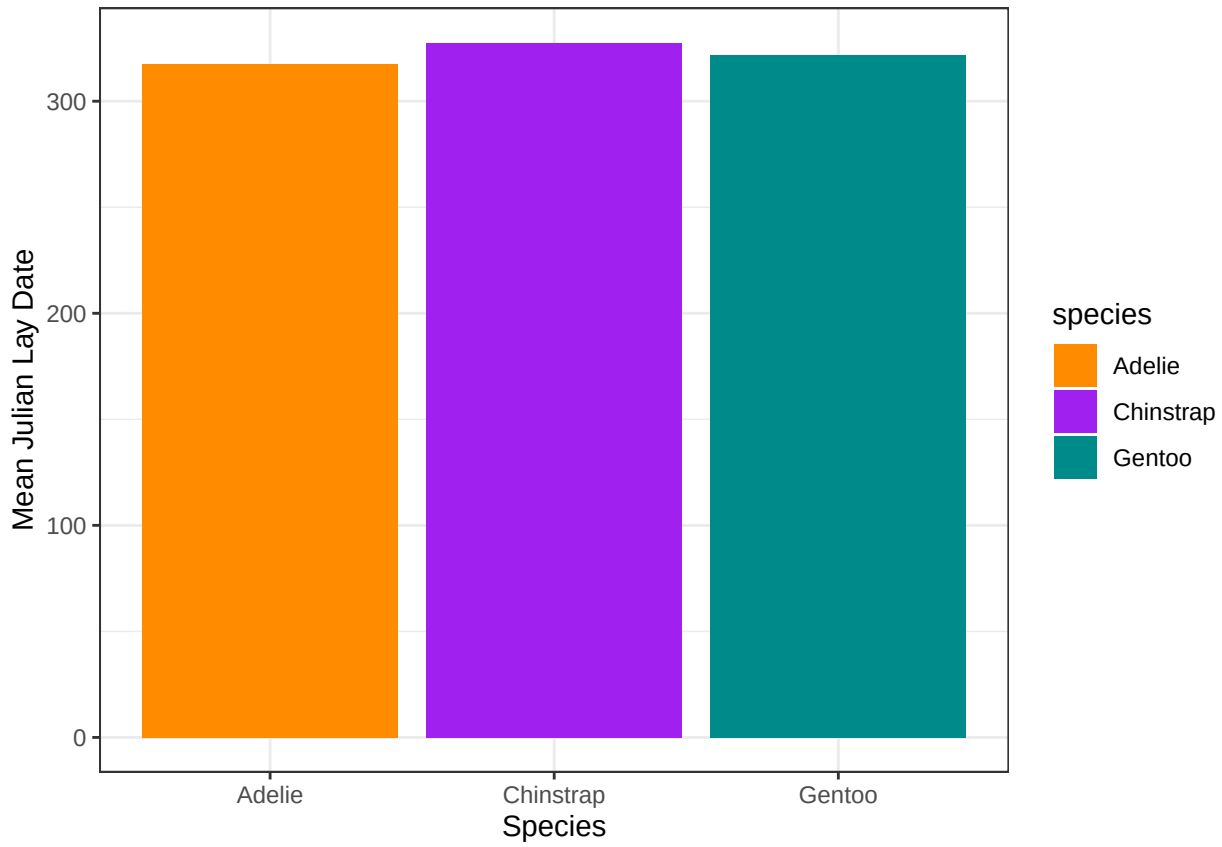

Figure 1: A bar chart of the mean lay date of the three species of Pygoscelis penguin

Figure 1 shows the mean late day for all three different penguin species. While it is technically correct, there are number of issues with it that make it misleading:

- It does not show the raw datapoints. This means that we have no idea of the sample size in each group, and we cannot see the distribution of the data. A lot of statistical analyses assume normally distributed data, and therefore without seeing the distribution we do not know whether any subsequent statistical tests are valid.
- The scale on the y-axis is misleading, suggesting that there is very little difference in mean lay date between species. However, if we were simply to re-scale the y-axis so that it ran from 300 to 330 we would see that there is in fact a 10-day difference in the mean lay date of Adelie penguins (x $= 317$) compared to Chinstrap penguins (x $= 327$).
- The bar chart does not include any error bars, and so we do not know how confident we are in these estimates of mean lay date for each species.
- The lay date of different species would be much better displayed using a boxplot and jittered data points to show the raw data, or boxplot and violin plot to show the distribution of the data. A label including the sample size for each group would also be helpful.

QUESTION 2: Creating a good exploratory figure

```
#lay date vs species
lay_date_vs_species <- ggplot(
    data = penguins_clean,
    aes(
        x = factor(species),
        y = julian_date)) +
  geom_violin(aes(fill = species),
        width = 0.4,
        alpha = 0.8,
        show.legend = FALSE) +
  geom_boxplot(aes(),
        width = 0.05,
        alpha = 1.0,
        show.legend = FALSE) +
  scale_fill_manual(values = species_colours) +
  theme_bw() +
  labs(
    x = "Species",
    y = " Julian Lay Date"
  )

#lay date in different years
lay_date_vs_year <- ggplot(
    data = penguins_clean,
    aes(
        x = year_egg_laid,
        y = julian_date)) +
   geom_violin(aes(fill = species),
        alpha = 0.8,
        position = position_dodge(0.8),
        show.legend = FALSE) +
```

```
  scale_fill_manual(values = species_colours) +
  labs(
    x = "Year",
    y = "Julian Lay Date") +
  theme_bw()

#make a multi-panel figure
grid.arrange(lay_date_vs_species,  lay_date_vs_year,  ncol = 2)

grid.text("A", x = 0.07, y = 0.95, gp = gpar(fontsize = 12)) #Adds label 'A' to first graph
grid.text("B", x = 0.57, y = 0.95, gp = gpar(fontsize = 12)) #Adds label 'B' to second graph
```

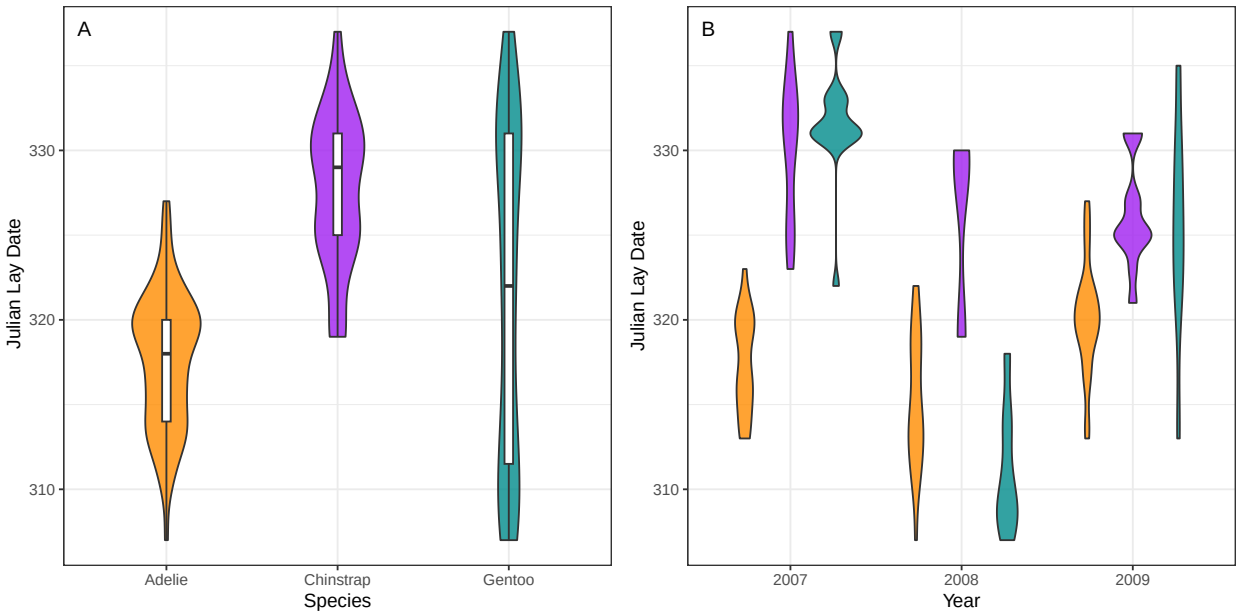

Figure 2: The distribution of lay dates for three species of Pygoscelis penguin. A) violin plot and boxplots of the lay date for the three species. B) Violin plot of the lay dates across the three years of the study. Boxplots show median, interquartile range, and maximum and minimum values. Violin plots show the density distribution of the data.

We can see from figure 2.A that lay date seems to differ by species. Adelie penguins have the earliest CID, while the Chinstrap penguins have a much later CID. Gentoos have a much broader distribution of CID. The data for Adelie and Chinstrap penguins appears to be slightly bi-modal, while the Gentoo data clearly deviates from the normal distribution. Various data transformations (natural log, $\log_{10}$, square root, and arcsine) were tested but none made the distributions more normal. Therefore, I still went ahead and ran ANOVA on the data because the test is reasonably robust to violations of normality, however I acknowledge that this is a limitation of my analysis.

From figure 2.B we can see that across the three years of the study, the CID remained roughly consistent for Adelies and Chinstraps, but Gentoo birds showed an earlier CID in 2008 compared to the other years of the study.

Following these observations from my exploratory analysis, I ran ANOVA and Tukey-Kramer tests for two hypotheses:

1) That the three species have distinct lay dates from each other

2) That Gentoos laid significantly earlier in 2008

Statistical analysis for lay date across species

```
#ANOVA for lay date across species
date_vs_species_model <- lm(julian_date ~ species, data = penguins_clean)
anova(date_vs_species_model)
```

```
## Analysis of Variance Table
##
## Response: julian_date
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## species     2  4570.7 2285.35  50.015 < 2.2e-16 ***
## Residuals 321 14667.6   45.69
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#pairwise comparisons using Tukey-Kramer test
TukeyHSD(aov(date_vs_species_model))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = date_vs_species_model)
##
## $species
##                        diff       lwr       upr   p adj
## Chinstrap-Adelie 10.000966  7.633738 12.368194 0.0e+00
## Gentoo-Adelie     4.060541  2.068174  6.052909 7.3e-06
## Gentoo-Chinstrap -5.940425 -8.375199 -3.505651 1.0e-07
```

```
#Boxplot showing the results of the Tukey-Kramer test
results_plot <- ggplot(penguins_clean,
  aes(x = species, y = julian_date)) +
  geom_boxplot(
    aes(fill = species),
    width = .25,
    alpha = .3,
    outlier.shape = NA) +
  geom_point(
    aes(color = species),
    size = 1.5,
    position = position_jitter(seed = 0, width = .1)
  ) +
  scale_color_manual(values = species_colours, guide = "none") +
  scale_fill_manual(values = species_colours, guide = "none")  +
  labs(
    x = "Species",
    y = " Julian Lay Date") +
  geom_signif(comparisons = list(c("Adelie", "Chinstrap"),
                                 c("Chinstrap", "Gentoo"),
                                 c("Adelie", "Gentoo")),
```

```
    annotations = c("***", "***", "***"), #adding the sigificance stars
    test = "aov",
    y_position = c(338, 339, 342)) +
  scale_y_continuous(limits = c(307, 345)) + #changing the scale of the y-axis
  theme_bw()

results_plot
```

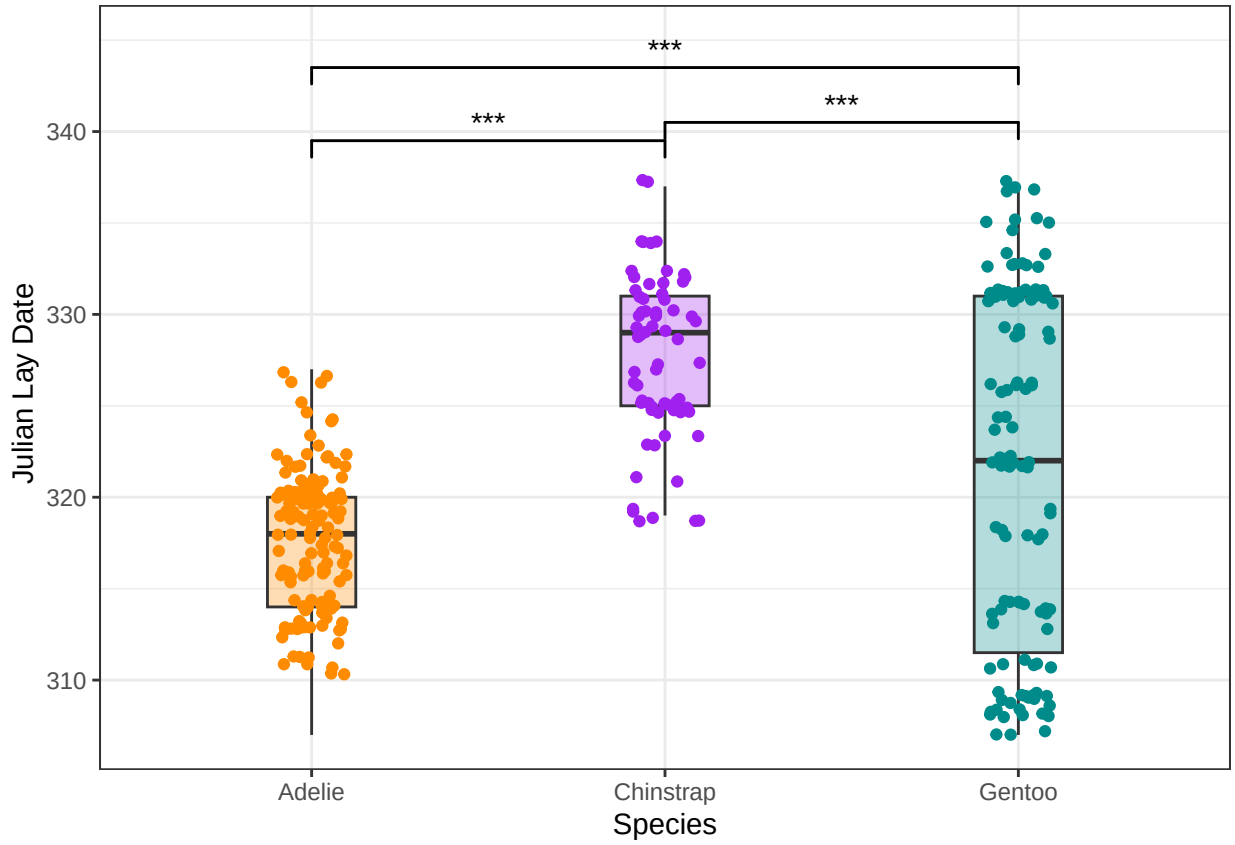

*Figure 3. A boxplot showing the results of the Tukey-Kramer test. A p-value of <0.001 is shown with '***', showing that all three species had significantly distinct CID from all other species. Boxplots show median, interquartile range and maximum and minimum values.*

Statistical analysis for lay date vs year for Gentoo penguins

```
#filter the data to just look at Gentoo penguins
gentoo_birds <- penguins_clean %>%
  filter(species == "Gentoo")

#ANOVA for lay date across different years for Gentoo penguins
date_vs_year_model <- lm(julian_date ~ year_egg_laid, data = gentoo_birds)
anova(date_vs_year_model)
```

```
## Analysis of Variance Table
##
## Response: julian_date
##                Df Sum Sq Mean Sq F value    Pr(>F)
```

```
## year_egg_laid   2 8825.8  4412.9  220.88 < 2.2e-16 ***
## Residuals     115 2297.6    20.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Pairwise comparisons using Tukey-Kramer test
TukeyHSD(aov(date_vs_year_model))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = date_vs_year_model)
##
## $year_egg_laid
##                  diff        lwr        upr p adj
## 2008-2007 -20.576389 -23.030593 -18.122185 0e+00
## 2009-2007  -6.199695  -8.703157  -3.696233 1e-07
## 2009-2008  14.376694  12.085312  16.668076 0e+00
```

Discussion

Lay date vs species

We can see from Figure 3 that the three species have distinct lay dates from each other. Indeed, the ANOVA gave a p-value of 2.2 x 10^{-16}, and the Tukey-Kramer test showed that all pairwise comparisons were significant. We already know that the timing of CID is important for penguins, because birds that delay breeding have lower reproductive success[6]. Therefore, we can conclude that the different species of Pygoscelid penguin have different optimal lay dates.

Food available to feed the chick might influence the optimal lay date, and since the species have different summer diets, perhaps these differences in lay date reflect the availability of different food sources throughout the Antarctic summer. Future research could examine the chick-provisioning diet of the three species, and see whether CID is timed to correlate with maximum abundance of those food sources when the chick requires most food.

These three species have overlapping ranges and often nest together. Another possible explanation for why they have distinct lay dates could be that it is a mechanism of niche separation, allowing them to feed on different diets and therefore promote coexistence.

The different migration patterns of the birds might also influence lay date. Adelie and Chinstrap birds migrate, while Gentoos do not[2]. This might explain why Gentoos are able to be more flexible with their CID, because they do not have to factor in time for migration, and are able to assess the local conditions and adjust their CID accordingly. Perhaps by the time Adelie and Chinstrap birds arrive at their nesting sites, they must lay eggs whether or not the local conditions are exactly ideal, because they must also factor in time for migration to and from summer breeding grounds.

Lay date vs year in Gentoo birds

The second ANOVA, looking at Gentoo lay date across the three years, gave a significant result with a p-value of 2.2 x 10^{-16}. The Tukey-Kramer test confirmed that all three years had distinct lay dates, but the most significant comparisons were those involving the year 2008, while the comparison 2009-2007 was less significant.

In 2014, Gorman et al reported that 2008 was the lowest sea ice season since 1979, whereas 2007 had intermediate sea ice, and 2009 had high sea ice[1]. Indeed, Hinke et al reported in 2012 that Gentoo birds appear to be able to plastically adjust their lay date in response to changing temperature[3]. Therefore, it

seems likely that Gentoos have a wider distribution of lay dates and can therefore plastically adjust their CID. 2008 was a particularly warm year, with lower sea ice, and therefore these birds laid earlier.

Antarctica is one of the most rapidly warming parts of the planet[7]. Under these changing environmental conditions, species which can plastically adjust their lay date are more likely to have greater reproductive success and therefore higher fitness. The results of this analysis suggest that Gentoos might be more resilient to climate change than Adelie or Chinstrap penguins.

Conclusion

Overall, this analysis has produced two conclusions:

1. Different species of Pygoscelid have distinct lay dates
2. Gentoo penguins laid earlier in 2008, suggesting they might have more plasticity in their CID.

We have discussed how factors such as migration, fasting, diet, and interspecific competition are likely to influence the optimum lay date for each species. Many abotic factors might also be important: we have discussed how sea ice coverage and temperature might explain the earlier lay date of Gentoos in 2008, but many other abotic factors such as rainfall and snowfall are also likely to be important in determining the optimal CID.

Future research should focus on understanding exactly which of these factors (both life-history traits and abiotic factors) influence lay date, and this might explain exactly why lay date differs between species. In this study, we have only been able to hypothesise that these factors might be important. A longer-term dataset over 10 years or more, combined with temperature recordings would allow a more detailed analysis of how temperature affects lay date. This could be done using a mixed-effects model, with 'year' as a random effect.

It is important to understand what factors affect lay date, particularly in the context of warming temperatures in the Antarctic. The limited analysis done here suggests that Gentoo birds might be more resilient to climate change than Adelie and Chinstrap penguins.

References

1. Gorman, K. B., Williams, T. D. & Fraser, W. R. Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (Genus Pygoscelis). PLoS One 9, (2014).
2. Black, C. E. A comprehensive review of the phenology of Pygoscelis penguins. Polar Biology vol. 39 405–432 Preprint at https://doi.org/10.1007/s00300-015-1807-8 (2016).
3. Hinke, J. T., Polito, M. J., Reiss, C. S., Trivelpiece, S. G. & Trivelpiece, W. Z. Flexible reproductive timing can buffer reproductive success of Pygoscelis spp. penguins in the Antarctic Peninsula region. Mar Ecol Prog Ser 454, 91–104 (2012).
4. Polito, M. J. et al. Contrasting specialist and generalist patterns facilitate foraging niche partitioning in sympatric populations of Pygoscelis penguins. Mar Ecol Prog Ser 519, 221–237 (2015).
5. Biochem, C. & Clarke, A. SEASONALITY IN THE ANTARCTIC MARINE ENVIRONMENT. Physiol vol. 90 (1988).
6. Juáres, M. A. et al. Better late than never? Interannual and seasonal variability in breeding chronology of gentoo penguins at stranger point, Antarctica. Polar Research vol. 32 Preprint at https://doi.org/10.3402/polar.v32i0.18448 (2013).

7. Bromwich, D. H. et al. Central West Antarctica among the most rapidly warming regions on Earth. Nat Geosci 6, 139–145 (2013).

QUESTION 3

a) Upload to GitHub. My repo link:

https://github.com/Biology5785/PenguinAssessmentProject

b) Share your repo with a partner and try to run their data pipeline. Partner's GitHub link: https://github.com/Tsavoboi/Homework_For_Palmers_Penguins

c) Reflect on running their code (300-500 words)

Overall, my partner's code was good – it was comprehensible and it ran. They added comments to many parts of the code, and this helped me to understand what they were doing at each step.

Almost all of the code ran.

- I had to install one package, but that was not a problem.
- Because the clean dataset was saved in a folder called 'data', I had to download the R Project file, the .Rmd file, and also create a folder called 'data' so that the clean dataset would actually save. This was not much of an issue, but it created a bit more work than having everything you need in just one self-contained .Rmd file.
- They wrote some code to save the figure as .png, but this was specific to their own file path. Hence, this code did not run on my computer, but again, this would not be too difficult to change the file path to fit my own documents.

To make the code more understandable and reproducible, I would suggest using more indentation when coding figures. Having the code all on one long line can make it difficult to read at times, so I would suggest splitting the code over several lines. This would also allow them to add comments to particular sections of the long line of code.

In figure 1, they plotted body mass and flipper length, but they chose to connect the datapoints from the same species together. This is a slightly odd choice given that the body mass (on the x-axis) does not change over time, so I am not sure why they connected the points together like a line graph. It might have been better to simply have the datapoints of different species in different colours, but not joined together.

Their choice of linear regression made sense given that they were correlating two continuous variables. However, the choice of an ANOVA does not seem quite right given that the explanatory variable (body mass) is a continuous variable, and an ANOVA is usually done with a categorial explanatory variable.

Overall, however, it was easy enough to follow what they were doing because of the combination of written text and commented code. I think it would be relatively easy to change their code if I needed to.

d) Reflect on your own code based on your experience with your partner's code and their review of yours (300-500 words)

My partner suggested that it might be better to not include the cleaning function in a separate file because this would allow the whole code to be run by simply downloading the .rRmd file, rather than having to separately also download the cleaning.R script. I agree with my partner that there are pros and cons to saving the cleaning function separately, such as allowing the cleaning of other datasets, but it also makes downloading and running the data pipeline a bit more complicated.

They also suggested saving the figures separately as a .png file to allow the figures to be extracted for evaluation and publication. I think this would be a good idea, and so upon their feedback I have included the code to do this below:

```
library(ragg)#package for saving figures as .png

agg_png("figures/bad_plot.png",
        width = 20, #width of figure in cm
        height = 20, #height of figure in cm
        units = "cm",
        res = 300, #resolution
        scaling = 1)
print(bad_plot)


agg_png("figures/lay_date_vs_year.png",
        width = 20,
        height = 20,
        units = "cm",
        res = 300,
        scaling = 1)
print(lay_date_vs_year)

agg_png("figures/lay_date_vs_species.png",
        width = 20,
        height = 20,
        units = "cm",
        res = 300,
        scaling = 1)
print(lay_date_vs_species)

agg_png("figures/results_plot.png",
        width = 20,
        height = 20,
        units = "cm",
        res = 300,
        scaling = 1)
print(results_plot)
```

Overall, I have learnt how it is important to add comments to code to make it clear what is being done at each step, but not so many comments that it becomes difficult to read the code. Using indentation and splitting up the code over several lines makes it much easier to follow. Organising the R.project into different folders such as 'data', 'functions' and 'figures' breaks the whole analysis down into manageable chunks.