# Group contract

**Group number:** L93G01

**Name:** Garth Tarr

**GitHub link:** https://github.sydney.edu.au/gtar4178/L93G01

I agree to:

* Abide by the terms of this contract in relation to the group assessment for DATA2002/2902.
* Store all my contributions to the assessment in the GitHub repository.
* Keep an accurate record of my contribution to the assessment. A copy of this may be requested by the coordinator.
* Work cooperatively, treat each other with respect, act honestly and ethically and not engage in any activities that could be perceived as bullying or harassment, as detailed in the [Student Charter](https://www.sydney.edu.au/policies/showdoc.aspx?recnum=PDOC2011/215&RendNum=0)
* Communicate in two main ways: informal discussions on Slack and using the [“Issues” functionality on GitHub](https://docs.github.com/en/issues/tracking-your-work-with-issues/about-issues) to provide updates on specific tasks, including tagging responsibility to specific group members.
* Check Slack daily and check in with GitHub at least once a week and more regularly as we get closer to the deadline. If something on GitHub is urgent, it will be highlighted in Slack.
* Attend labs in the weeks before the tasks are due and meet for lunch on the day of the lab to give us time to informally discuss any issues we’re facing. Other meetings will be held via Zoom and arranged on an ad hoc basis.

I understand that:

* My agreement to these terms is indicated through the act of submitting this in Canvas.
* If I fail to meet my obligations as detailed in this group contract, then I have failed to meet the assessment requirements for DATA2002/2902 and may be awarded a mark of zero for some or all of the project components.

# Exploratory data analysis

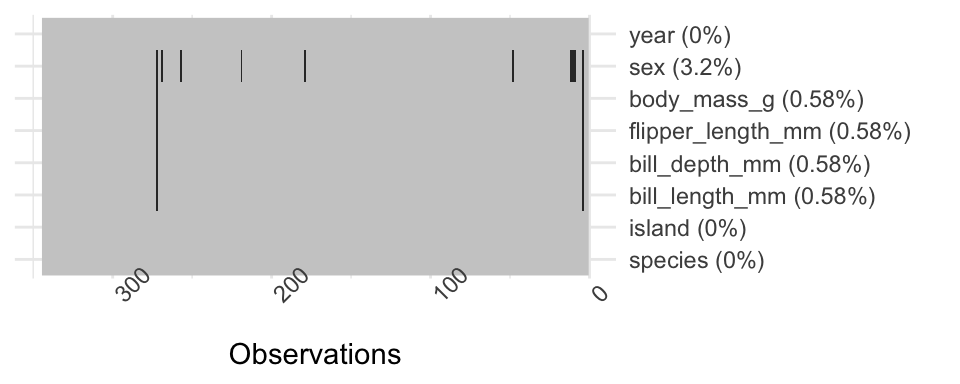
**Data set:** Palmer penguins

**Dependent variable:** Body mass

library(tidyverse)  
theme\_set(theme\_bw())  
data("penguins", package = "palmerpenguins")

Looking at the pattern of missing data:

visdat::vis\_miss(penguins) + coord\_flip() + theme(legend.position = "none")



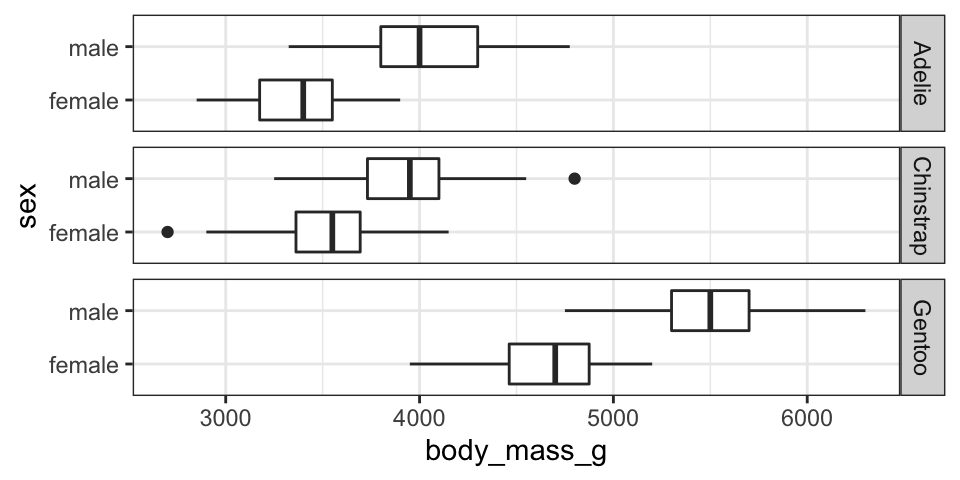
Descriptive statistics:

table1::table1(~ bill\_length\_mm + bill\_depth\_mm + flipper\_length\_mm + body\_mass\_g + island | species + sex, data = penguins, overall = FALSE)

|  | Adelie | | | Chinstrap | | Gentoo | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | female (N=73) | male (N=73) | female (N=34) | male (N=34) | female (N=58) | male (N=61) |
| **bill\_length\_mm** | |  |  |  |  |  |  |
| Mean (SD) | | 37.3 (2.03) | 40.4 (2.28) | 46.6 (3.11) | 51.1 (1.56) | 45.6 (2.05) | 49.5 (2.72) |
| Median [Min, Max] | | 37.0 [32.1, 42.2] | 40.6 [34.6, 46.0] | 46.3 [40.9, 58.0] | 51.0 [48.5, 55.8] | 45.5 [40.9, 50.5] | 49.5 [44.4, 59.6] |
| **bill\_depth\_mm** | |  |  |  |  |  |  |
| Mean (SD) | | 17.6 (0.943) | 19.1 (1.02) | 17.6 (0.781) | 19.3 (0.761) | 14.2 (0.540) | 15.7 (0.741) |
| Median [Min, Max] | | 17.6 [15.5, 20.7] | 18.9 [17.0, 21.5] | 17.7 [16.4, 19.4] | 19.3 [17.5, 20.8] | 14.3 [13.1, 15.5] | 15.7 [14.1, 17.3] |
| **flipper\_length\_mm** | | | | | | | |
| Mean (SD) | | 188 (5.60) | 192 (6.60) | 192 (5.75) | 200 (5.98) | 213 (3.90) | 222 (5.67) |
| Median [Min, Max] | | 188 [172, 202] | 193 [178, 210] | 192 [178, 202] | 201 [187, 212] | 212 [203, 222] | 221 [208, 231] |
| **body\_mass\_g** | |  |  |  |  |  |  |
| Mean (SD) | | 3370 (269) | 4040 (347) | 3530 (285) | 3940 (362) | 4680 (282) | 5480 (313) |
| Median [Min, Max] | | 3400 [2850, 3900] | 4000 [3330, 4780] | 3550 [2700, 4150] | 3950 [3250, 4800] | 4700 [3950, 5200] | 5500 [4750, 6300] |
| **island** | |  |  |  |  |  |  |
| Biscoe | | 22 (30.1%) | 22 (30.1%) | 0 (0%) | 0 (0%) | 58 (100%) | 61 (100%) |
| Dream | | 27 (37.0%) | 28 (38.4%) | 34 (100%) | 34 (100%) | 0 (0%) | 0 (0%) |
| Torgersen | | 24 (32.9%) | 23 (31.5%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) |

Initial visualisations:

penguins |> drop\_na() |> ggplot() + aes(y = sex, x = body\_mass\_g) +   
 geom\_boxplot() + facet\_grid( species ~ .)



penguins |> select(-island, - year) |>   
 pivot\_longer(cols = c(bill\_length\_mm, bill\_depth\_mm, flipper\_length\_mm),  
 names\_to = "variable", values\_to = "values") |>   
 drop\_na() |> ggplot() +   
 aes(x = values, y = body\_mass\_g, colour = sex) +   
 geom\_point(alpha = 0.3) +   
 facet\_grid(species ~ variable, scales = "free\_x") +   
 scale\_colour\_brewer(palette = "Set1") +   
 geom\_smooth(method = "lm", se = FALSE) +   
 theme(legend.position = "top")



Summary: From my initial EDA it looks like species and sex are both important factors for predicting body mass. There also appears to be relationships between body mass and the bill and flipper measurements.

*Note: your EDA doesn’t necessarily need to include the R code, I included mine just to show how I generated the outputs. I still used a .qmd file but then copied the output into this Word document. If you’re interested, the start of the .qmd file was:*

---

title: 'Project EDA'

format:

docx:

fig-format: retina

---

```{r, message=FALSE}

library(tidyverse)

theme\_set(theme\_bw())

data("penguins", package = "palmerpenguins")

```