Project EDA

## Group Contract

**Group number:** 003E06

**Name:** Anabel Geraldine

**GitHub link:** https://github.sydney.edu.au/ager5893/003E06

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 Contract](https://www.sydney.edu.au/policies/showdoc.aspx?recnum=PDOC2011/215&RendNum=0)
* Communicate in two main ways: informal discussions on Messenger 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 Messenger frequently 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 Messenger.
* Meet after labs to talk about any updates or issues. 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:** Abalone

**Dependent variable:** Whole Weight

### Reading Data

library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.3 ✔ readr 2.1.4  
✔ forcats 1.0.0 ✔ stringr 1.5.0  
✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
✔ purrr 1.0.1   
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(dplyr)  
x = readr::read\_csv("abalone.data.csv")

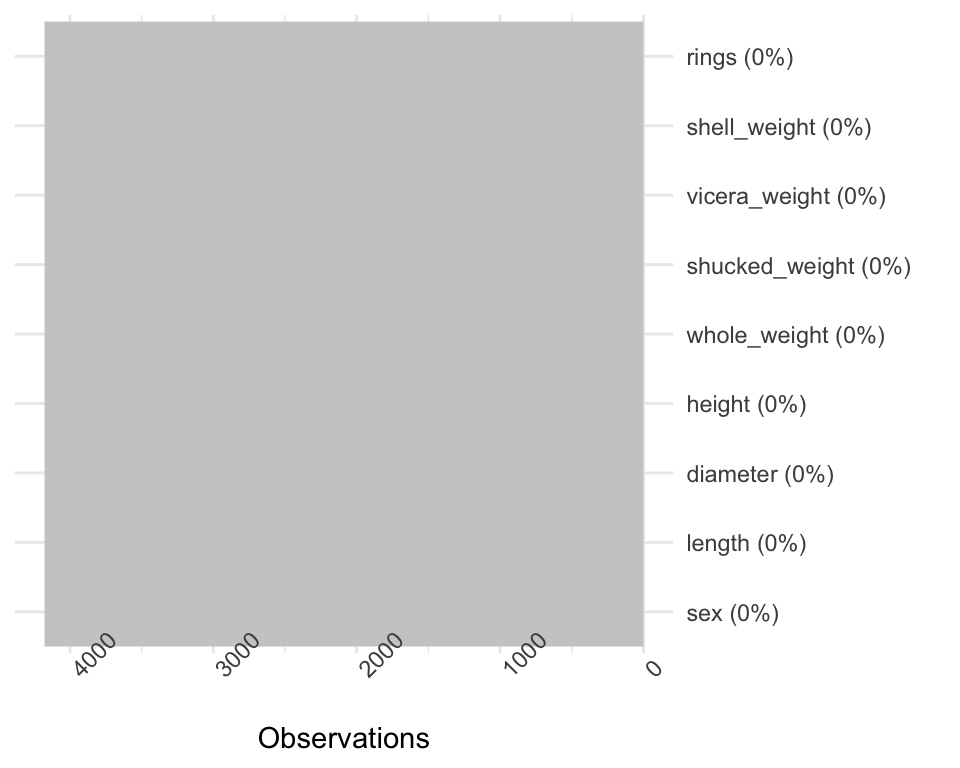
Rows: 4176 Columns: 9  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (1): M  
dbl (8): 0.455, 0.365, 0.095, 0.514, 0.2245, 0.101, 0.15, 15  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

colnames(x) <- c("sex","length","diameter","height","whole\_weight","shucked\_weight","vicera\_weight","shell\_weight","rings")

### Checking for missing Data

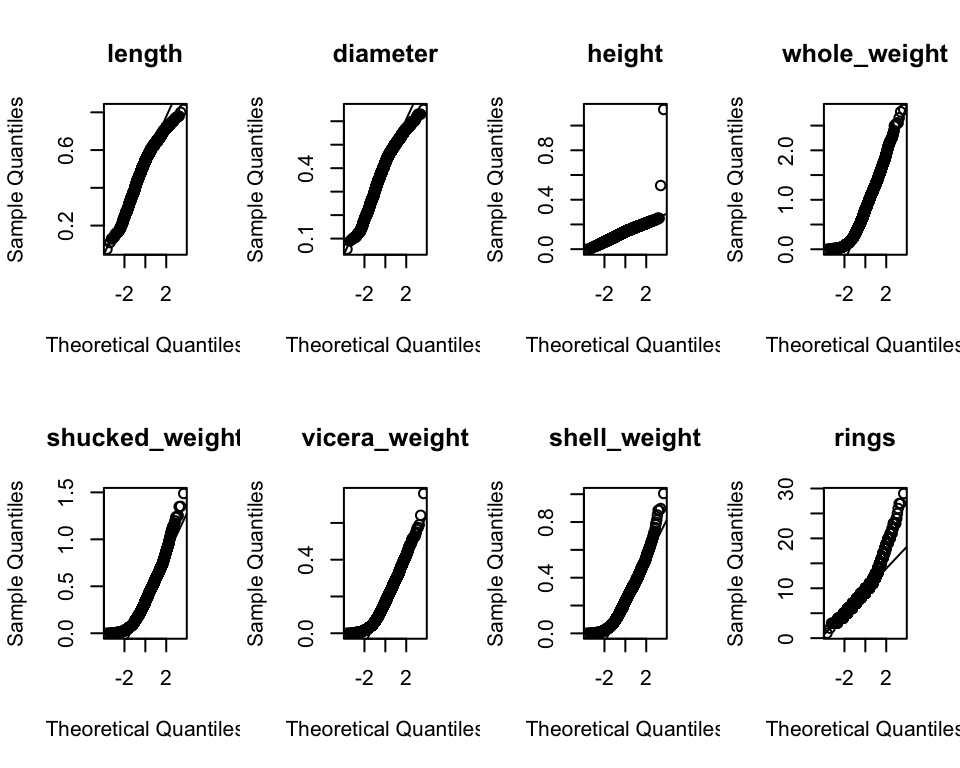
According to the website where the data set is taken from, in the data set description (variable table), there is supposed to be no missing value in all their variables.

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



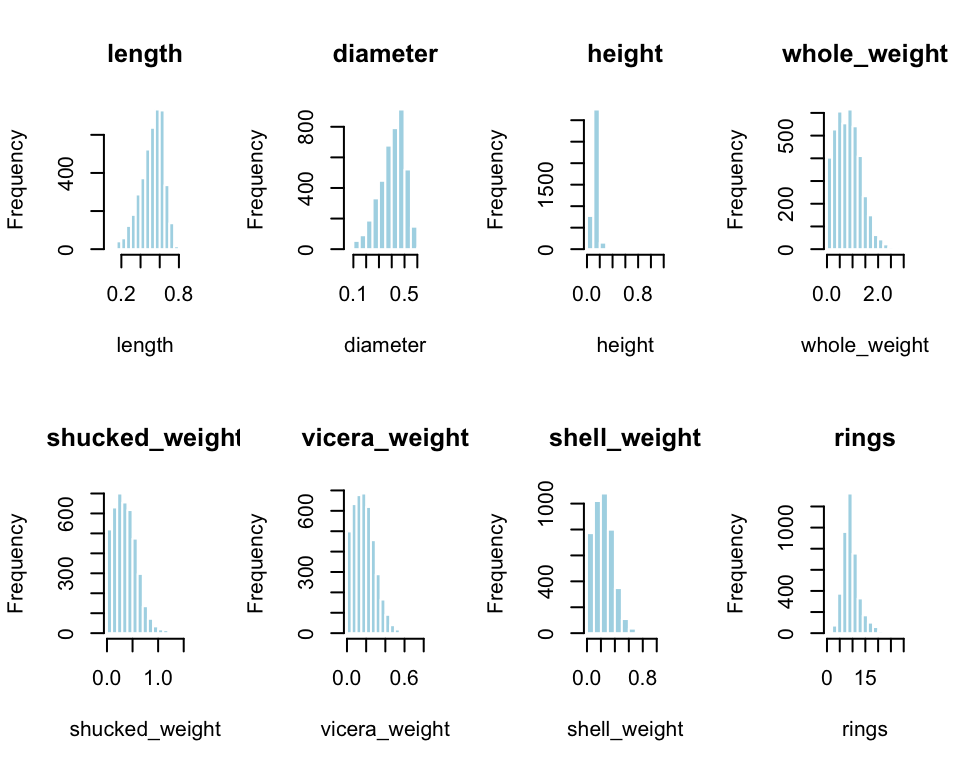
### Checking for normality

numeric <- c("length", "diameter", "height", "whole\_weight", "shucked\_weight", "vicera\_weight", "shell\_weight", "rings")  
  
par(mfrow = c(2, 4))  
  
for (var in numeric) {  
 qqnorm(x[[var]], main = var)  
 qqline(x[[var]])  
}



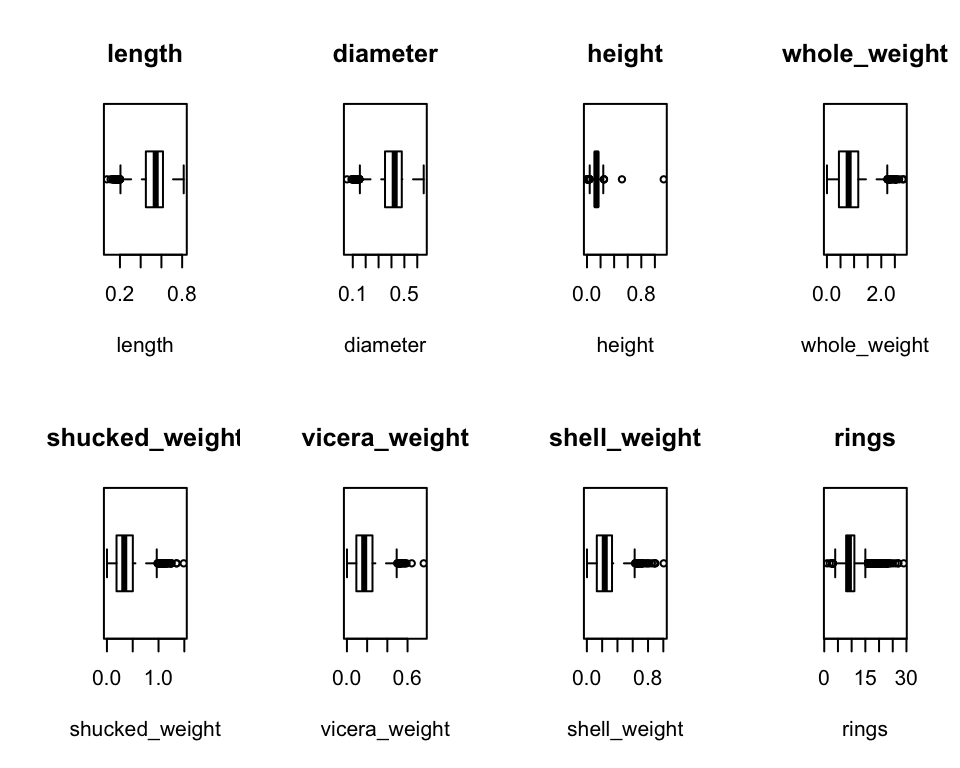
par(mfrow = c(1, 1))

numvar <- c("length", "diameter", "height", "whole\_weight", "shucked\_weight", "vicera\_weight", "shell\_weight", "rings")  
  
par(mfrow = c(2, 4))  
  
for (var in numvar) {  
 hist(x[[var]], main = var, xlab = var, col = "light blue", border = "white")  
}



par(mfrow = c(1, 1))

par(mfrow = c(2, 4))  
  
for (var in numvar) {  
 boxplot(x[[var]], main = var, col = "white", border = "black", xlab = var, horizontal = TRUE)  
}



par(mfrow = c(1, 1))

### Descriptive Statistics

library(table1)

Attaching package: 'table1'

The following objects are masked from 'package:base':  
  
 units, units<-

table1::table1(~ length + diameter + height + whole\_weight + shucked\_weight + vicera\_weight + shell\_weight + rings | sex, data = x)

Get nicer `table1` .docx output by simply installing the `flextable` package

F I  
1 (N=1307) (N=1342)  
2 length   
3 Mean (SD) 0.579 (0.0862) 0.428 (0.109)  
4 Median [Min, Max] 0.590 [0.275, 0.815] 0.435 [0.0750, 0.725]  
5 diameter   
6 Mean (SD) 0.455 (0.0710) 0.326 (0.0881)  
7 Median [Min, Max] 0.465 [0.195, 0.650] 0.335 [0.0550, 0.550]  
8 height   
9 Mean (SD) 0.158 (0.0400) 0.108 (0.0320)  
10 Median [Min, Max] 0.160 [0.0150, 1.13] 0.110 [0, 0.220]  
11 whole\_weight   
12 Mean (SD) 1.05 (0.430) 0.431 (0.286)  
13 Median [Min, Max] 1.04 [0.0800, 2.66] 0.384 [0.00200, 2.05]  
14 shucked\_weight   
15 Mean (SD) 0.446 (0.199) 0.191 (0.128)  
16 Median [Min, Max] 0.441 [0.0310, 1.49] 0.170 [0.00100, 0.774]  
17 vicera\_weight   
18 Mean (SD) 0.231 (0.0976) 0.0920 (0.0625)  
19 Median [Min, Max] 0.224 [0.0210, 0.590] 0.0805 [0.000500, 0.441]  
20 shell\_weight   
21 Mean (SD) 0.302 (0.126) 0.128 (0.0849)  
22 Median [Min, Max] 0.295 [0.0250, 1.01] 0.113 [0.00150, 0.655]  
23 rings   
24 Mean (SD) 11.1 (3.10) 7.89 (2.51)  
25 Median [Min, Max] 10.0 [5.00, 29.0] 8.00 [1.00, 21.0]  
 M Overall  
1 (N=1527) (N=4176)  
2   
3 0.561 (0.103) 0.524 (0.120)  
4 0.580 [0.155, 0.780] 0.545 [0.0750, 0.815]  
5   
6 0.439 (0.0844) 0.408 (0.0992)  
7 0.455 [0.110, 0.630] 0.425 [0.0550, 0.650]  
8   
9 0.151 (0.0348) 0.140 (0.0418)  
10 0.155 [0.0250, 0.515] 0.140 [0, 1.13]  
11   
12 0.992 (0.471) 0.829 (0.490)  
13 0.976 [0.0155, 2.83] 0.800 [0.00200, 2.83]  
14   
15 0.433 (0.223) 0.359 (0.222)  
16 0.422 [0.00650, 1.35] 0.336 [0.00100, 1.49]  
17   
18 0.216 (0.105) 0.181 (0.110)  
19 0.210 [0.00300, 0.760] 0.171 [0.000500, 0.760]  
20   
21 0.282 (0.131) 0.239 (0.139)  
22 0.276 [0.00500, 0.897] 0.234 [0.00150, 1.01]  
23   
24 10.7 (3.03) 9.93 (3.22)  
25 10.0 [3.00, 27.0] 9.00 [1.00, 29.0]

### Initial Visualization

#### 1. Length vs whole\_weight, shucked\_weight, vicera\_weight, shell\_weight (and sex)

library(ggplot2)  
library(gridExtra)

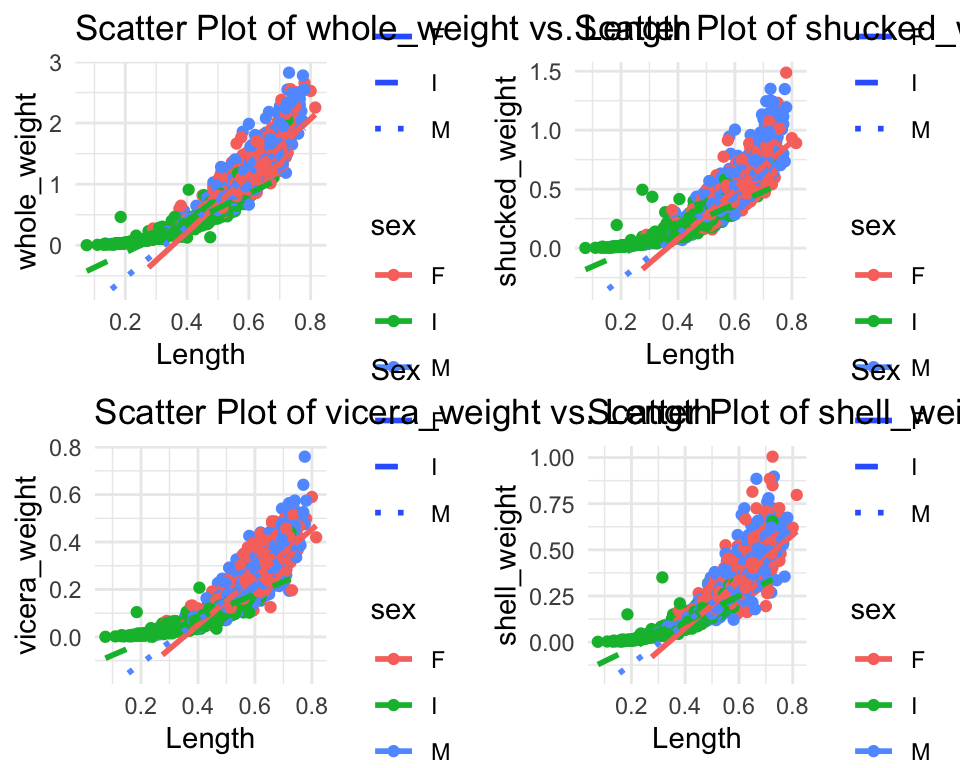
Attaching package: 'gridExtra'

The following object is masked from 'package:dplyr':  
  
 combine

variables <- c("whole\_weight", "shucked\_weight", "vicera\_weight", "shell\_weight")  
  
scatter\_plots <- lapply(variables, function(var) {  
 ggplot(x, aes(x = length, y = get(var), color = sex)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE, aes(group = sex, linetype = sex)) +  
 labs(title = paste("Scatter Plot of", var, "vs. Length"),  
 x = "Length", y = var,  
 linetype = "Sex") +  
 scale\_linetype\_manual(values = c("solid", "dashed", "dotted")) +  
 theme\_minimal()  
})  
  
grid.arrange(  
 scatter\_plots[[1]], scatter\_plots[[2]], scatter\_plots[[3]], scatter\_plots[[4]],  
 ncol = 2  
)

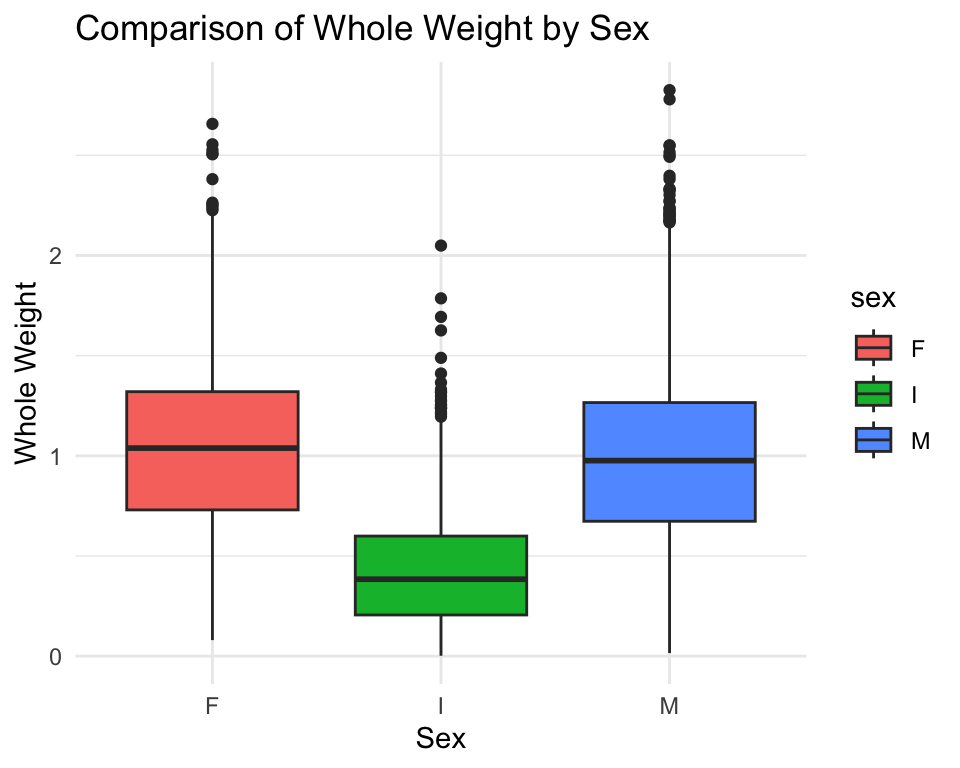
`geom\_smooth()` using formula = 'y ~ x'

`geom\_smooth()` using formula = 'y ~ x'  
`geom\_smooth()` using formula = 'y ~ x'  
`geom\_smooth()` using formula = 'y ~ x'



#### 2. Sex vs Whole\_weight

library(ggplot2)  
  
boxplot <- ggplot(x, aes(x = sex, y = whole\_weight, fill = sex)) +  
 geom\_boxplot() +   
 labs(title = "Comparison of Whole Weight by Sex",  
 x = "Sex", y = "Whole Weight") +   
 theme\_minimal()   
  
print(boxplot)



library(ggplot2)  
  
x$sex <- factor(x$sex, levels = c("M", "F", "I"))  
  
columns\_to\_plot <- c("whole\_weight", "shucked\_weight", "vicera\_weight", "shell\_weight")  
  
x\_long <- tidyr::gather(x, key = "Variable", value = "Value", columns\_to\_plot)

Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.  
ℹ Please use `all\_of()` or `any\_of()` instead.  
 # Was:  
 data %>% select(columns\_to\_plot)  
  
 # Now:  
 data %>% select(all\_of(columns\_to\_plot))  
  
See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.

ggplot(x\_long, aes(x = Variable, y = Value, fill = sex)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(title = "Bar Chart of Variables by Sex",  
 x = "Variables",  
 y = "Value") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, vjust = 0.9, hjust = 1))

