The first project task

2022-11-21

```
library(ggplot2)
library(tidyr)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

1 User function

```
tablel_merge <- function(path_to_data) {
  names <- list.files(path=path_to_data, full.names=TRUE)
  tables <- lapply(names, function (x) read.csv(file=x, sep=','))
  Reduce(function(x,y) merge(x,y, all = TRUE), tables)}
```

2 A brief EDA

Downloading data:

```
work_table <- tablel_merge('/Users/asabukreeva/Downloads/Data')
```

EDA

```
summary(work_table)
```

```
##     Rings            Sex..1...male..2...female..3...uvenil.    Length
##  Length:4177        Length:4177                               Length:4177
##  Class :character   Class :character                          Class :character
##  Mode  :character   Mode  :character                          Mode  :character
##
##
##
##
##     Diameter         Height         Whole_weight    Shucked_weight
##  Min.   :0.055   Min.   :0.0000   Min.   :0.0020   Min.   :0.0010
##  1st Qu.:0.350   1st Qu.:0.1150   1st Qu.:0.4415   1st Qu.:0.1865
##  Median :0.425   Median :0.1400   Median :0.7995   Median :0.3360
##  Mean   :0.408   Mean   :0.1395   Mean   :0.8285   Mean   :0.3595
##  3rd Qu.:0.480   3rd Qu.:0.1650   3rd Qu.:1.1530   3rd Qu.:0.5020
```

```
##  Max.   :0.650   Max.   :1.1300   Max.   :2.8255   Max.   :1.4880
##  NA's   :5       NA's   :2        NA's   :1        NA's   :3
##  Viscera_weight    Shell_weight
##  Min.   :0.0005   Min.   :0.0015
##  1st Qu.:0.0935   1st Qu.:0.1300
##  Median :0.1710   Median :0.2335
##  Mean   :0.1806   Mean   :0.2388
##  3rd Qu.:0.2530   3rd Qu.:0.3285
##  Max.   :0.7600   Max.   :1.0050
##                   NA's   :2
```

Table check Working with NA in numeric columns

```
work_table$Length <- as.numeric(work_table$Length)
```

```
## Warning: NAs introduced by coercion
```

```
work_table$Length[is.na(work_table$Length)]<-mean(work_table$Length, na.rm = T)
work_table$Diameter[is.na(work_table$Diameter)]<-mean(work_table$Diameter, na.rm = T)
work_table$Height[is.na(work_table$Height)]<-mean(work_table$Height, na.rm = T)
work_table$Whole_weight[is.na(work_table$Whole_weight)]<-mean(work_table$Whole_weight, na.rm = T)
work_table$Shucked_weight[is.na(work_table$Shucked_weight)]<-mean(work_table$Shucked_weight, na.rm = T)
work_table$Shell_weight[is.na(work_table$Shell_weight)]<-mean(work_table$Shell_weight, na.rm = T)
work_table <- work_table %>% drop_na()

summary(work_table)
```

```
##     Rings           Sex..1...male..2...female..3...uvenil.     Length
##  Length:4176        Length:4176                            Min.   :0.0750
##  Class :character   Class :character                       1st Qu.:0.4500
##  Mode  :character   Mode  :character                       Median :0.5450
##                                                            Mean   :0.5241
##                                                            3rd Qu.:0.6150
##                                                            Max.   :0.8150
##     Diameter          Height        Whole_weight    Shucked_weight
##  Min.   :0.055   Min.   :0.0000   Min.   :0.0020   Min.   :0.0010
##  1st Qu.:0.350   1st Qu.:0.1150   1st Qu.:0.4415   1st Qu.:0.1865
##  Median :0.425   Median :0.1400   Median :0.7995   Median :0.3360
##  Mean   :0.408   Mean   :0.1395   Mean   :0.8285   Mean   :0.3595
##  3rd Qu.:0.480   3rd Qu.:0.1650   3rd Qu.:1.1530   3rd Qu.:0.5020
##  Max.   :0.650   Max.   :1.1300   Max.   :2.8255   Max.   :1.4880
##  Viscera_weight     Shell_weight
##  Min.   :0.00050   Min.   :0.0015
##  1st Qu.:0.09337   1st Qu.:0.1300
##  Median :0.17075   Median :0.2338
##  Mean   :0.18058   Mean   :0.2388
##  3rd Qu.:0.25300   3rd Qu.:0.3285
##  Max.   :0.76000   Max.   :1.0050
```

Now I have the missing values in all numeric columns replaced by the column average. Next step is correcting the incorrect values in the columns.

```
work_table$Rings[(work_table$Rings) == 'nine']<- 9
work_table$Sex..1...male..2...female..3...uvenil.[(work_table$Sex..1...male..2...female..3...uvenil.) ==
work_table$Sex..1...male..2...female..3...uvenil.[(work_table$Sex..1...male..2...female..3...uvenil.) ==
work_table$Sex..1...male..2...female..3...uvenil.[(work_table$Sex..1...male..2...female..3...uvenil.) ==
```

```
work_table <- work_table %>% drop_na()
```

I decided to change names and values of sex column (original name is long and really annoying)

```
work_table$Sex..1...male..2...female..3...uvenil.[(work_table$Sex..1...male..2...female..3...uvenil.) ==
work_table$Sex..1...male..2...female..3...uvenil.[(work_table$Sex..1...male..2...female..3...uvenil.) ==
work_table$Sex..1...male..2...female..3...uvenil.[(work_table$Sex..1...male..2...female..3...uvenil.) ==
names(work_table)[names(work_table) == 'Sex..1...male..2...female..3...uvenil.'] <- 'Sex'
names(work_table)
```

```
## [1] "Rings"          "Sex"            "Length"         "Diameter"
## [5] "Height"         "Whole_weight"   "Shucked_weight" "Viscera_weight"
## [9] "Shell_weight"
```

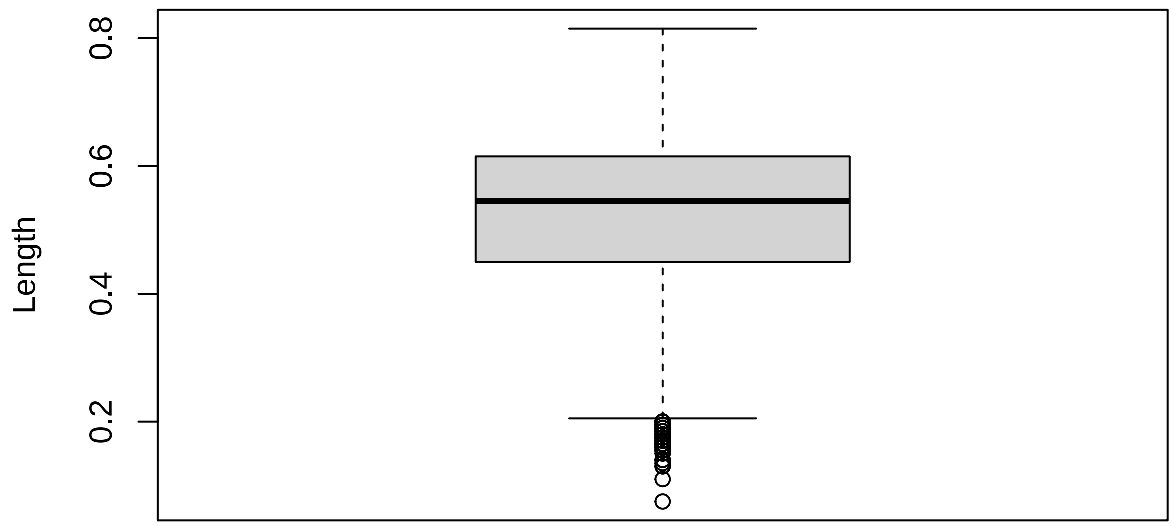
I'm not sure that it is correct, but I decided to change Rings collumn as numeric.

```
work_table$Rings <- as.numeric(work_table$Rings)
```

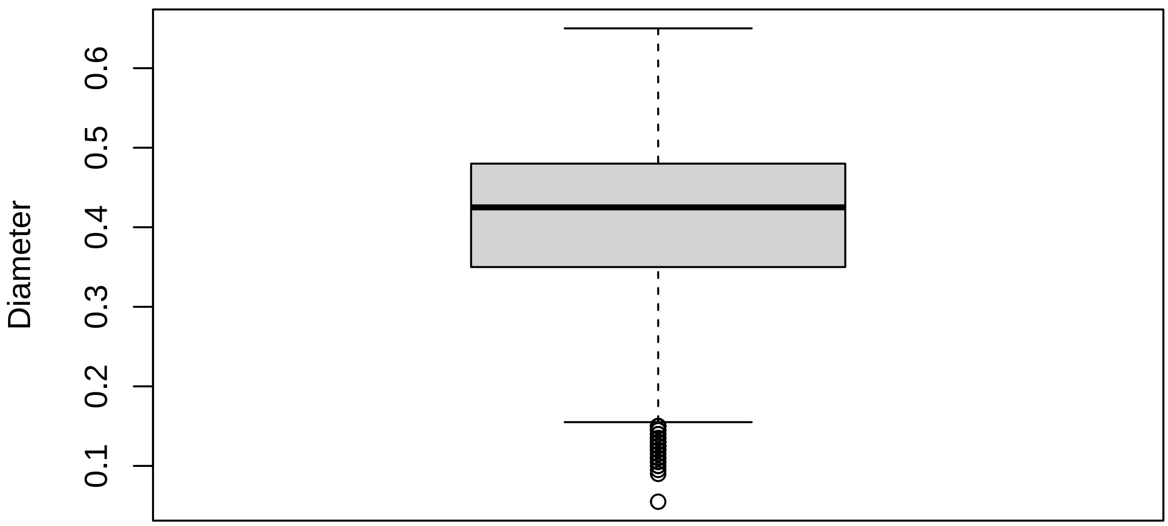
Fine, it looks like the table is now clean and I can build boxplots to check for outliers.

Checking Emissions

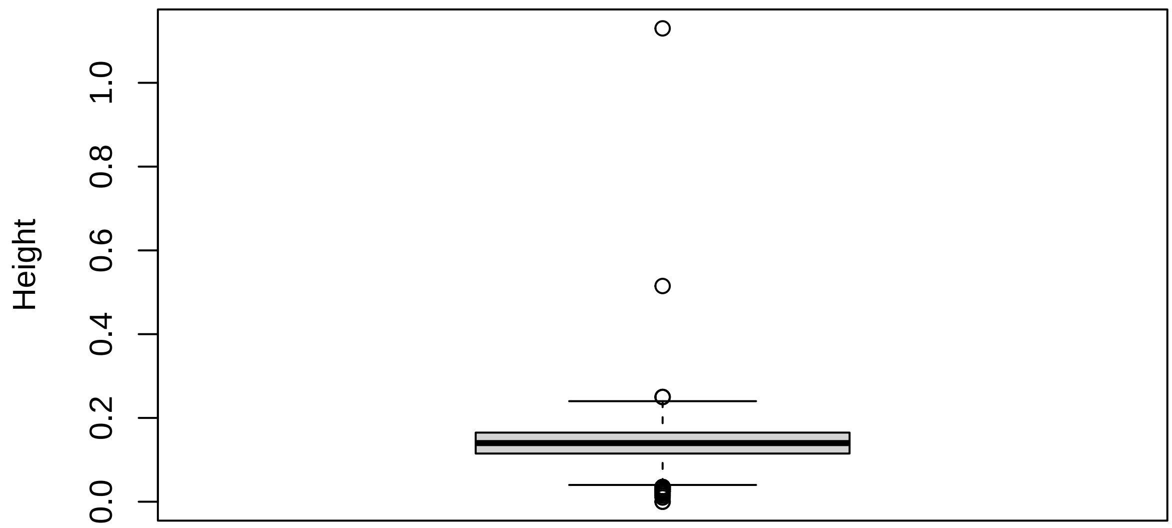
```
boxplot(work_table$Length,
        ylab = "Length")
```



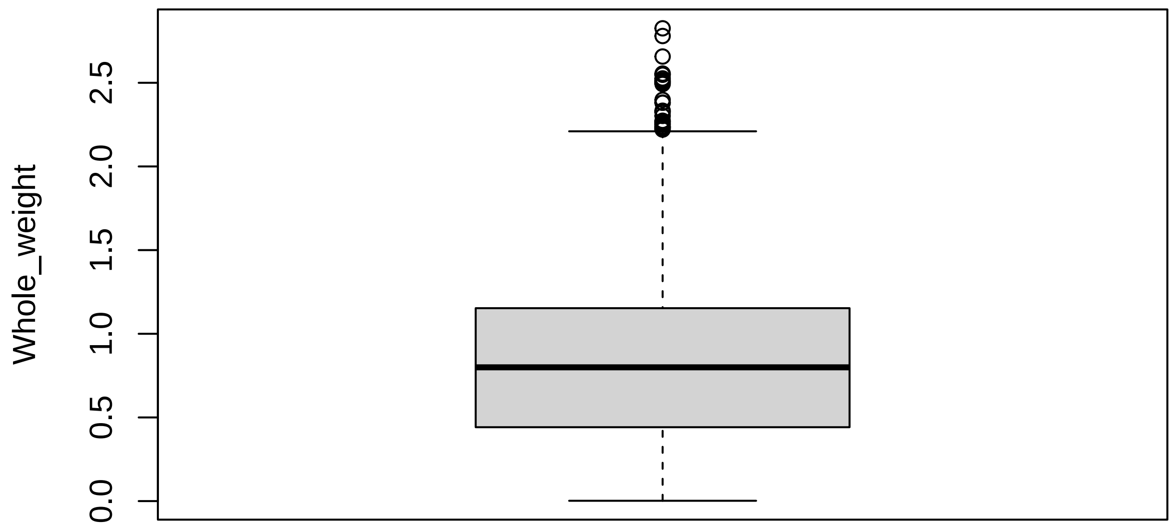

```
boxplot(work_table$Diameter,
        ylab = "Diameter")
```



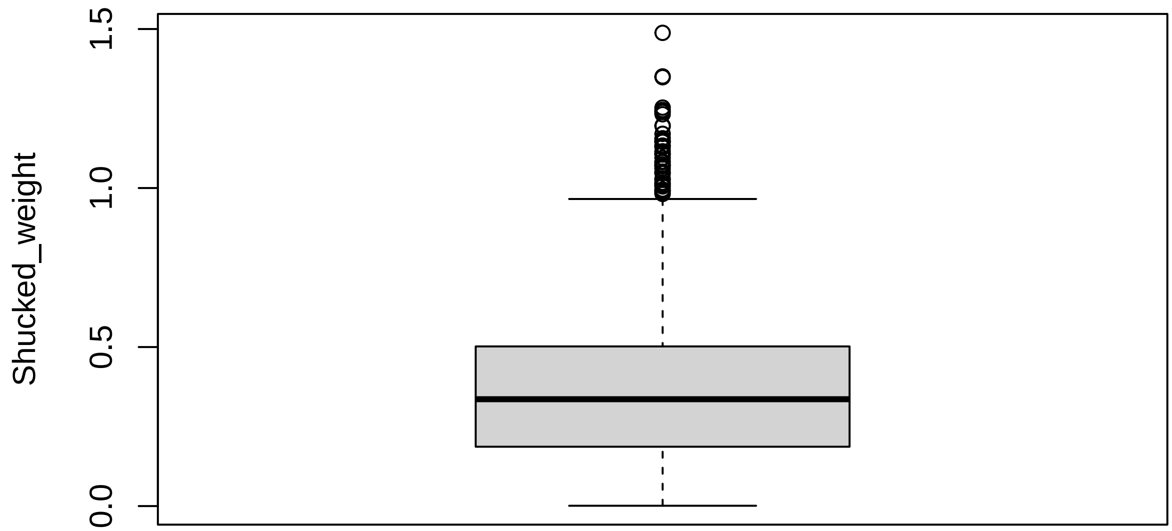

```
boxplot(work_table$Height,
        ylab = "Height")
```



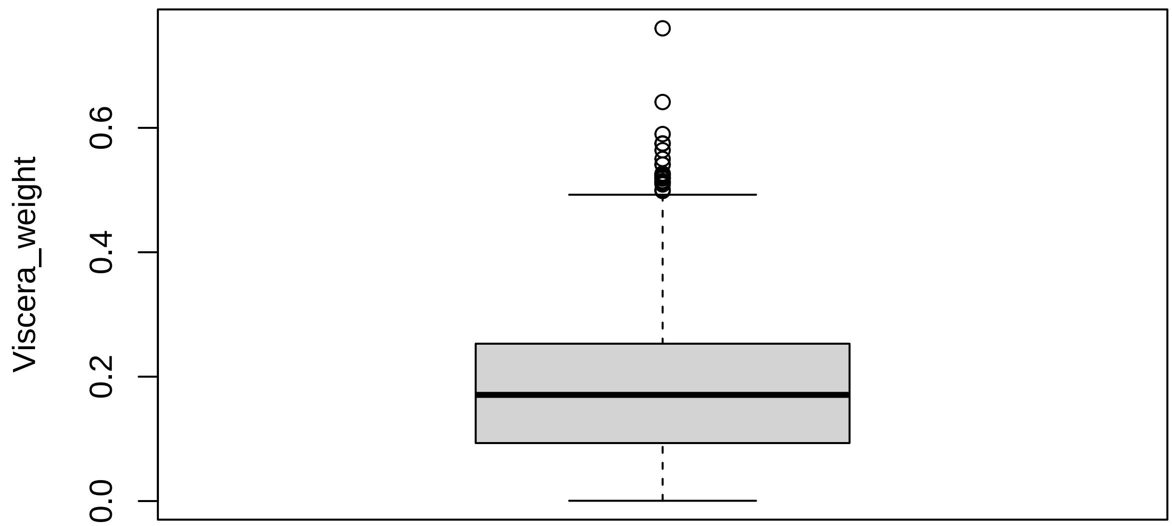

```
boxplot(work_table$Whole_weight,
        ylab = "Whole_weight")
```



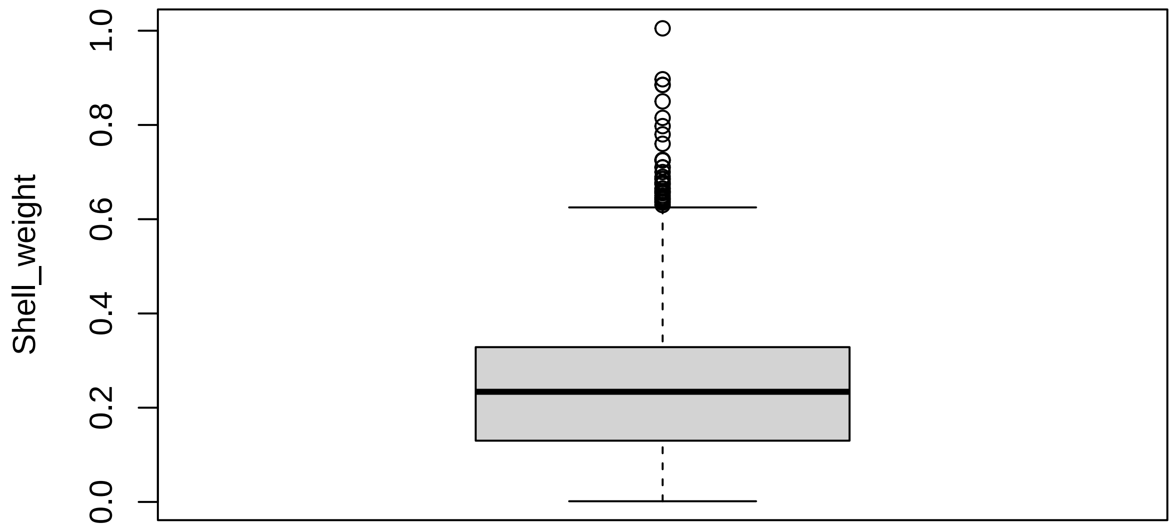

```
boxplot(work_table$Shucked_weight,
        ylab = "Shucked_weight")
```



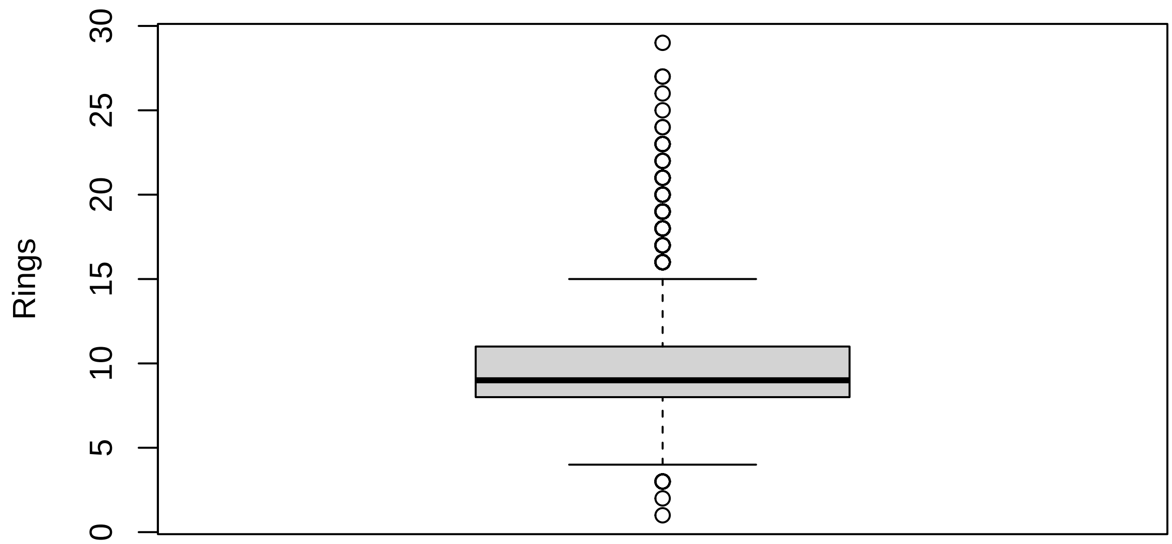

```
boxplot(work_table$Viscera_weight,
        ylab = "Viscera_weight")
```




```
boxplot(work_table$Shell_weight,
        ylab = "Shell_weight")
```

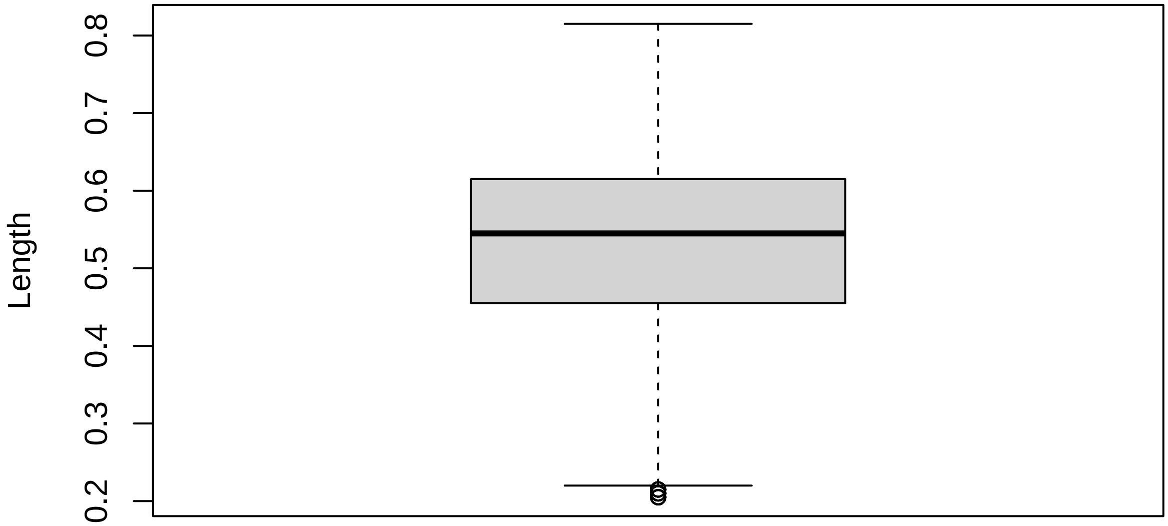



```
boxplot(work_table$Rings,
        ylab = "Rings")
```

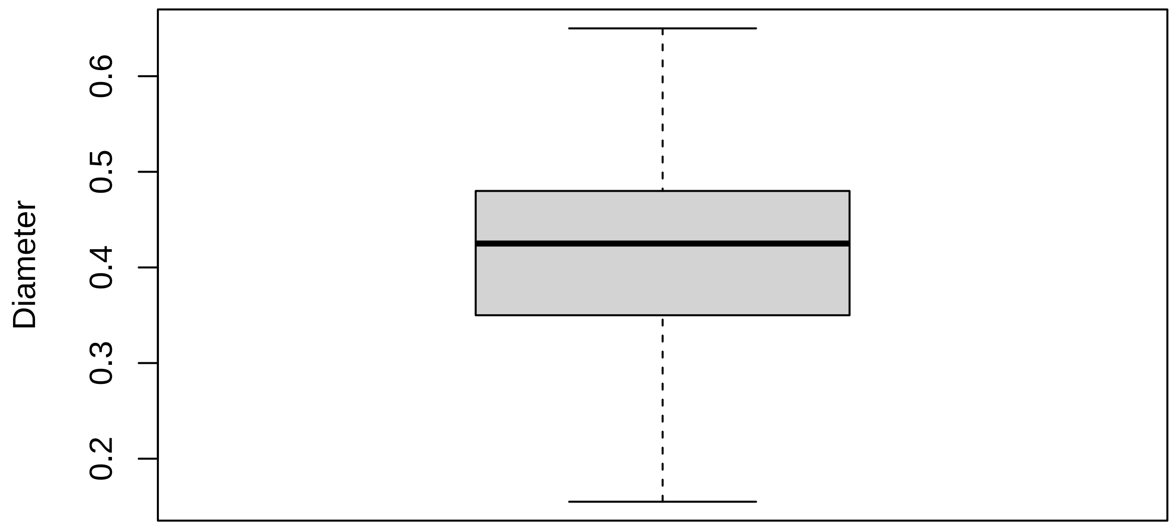


Oh, the outliers are everywhere. I try to get rid of them using quantile restrictions.

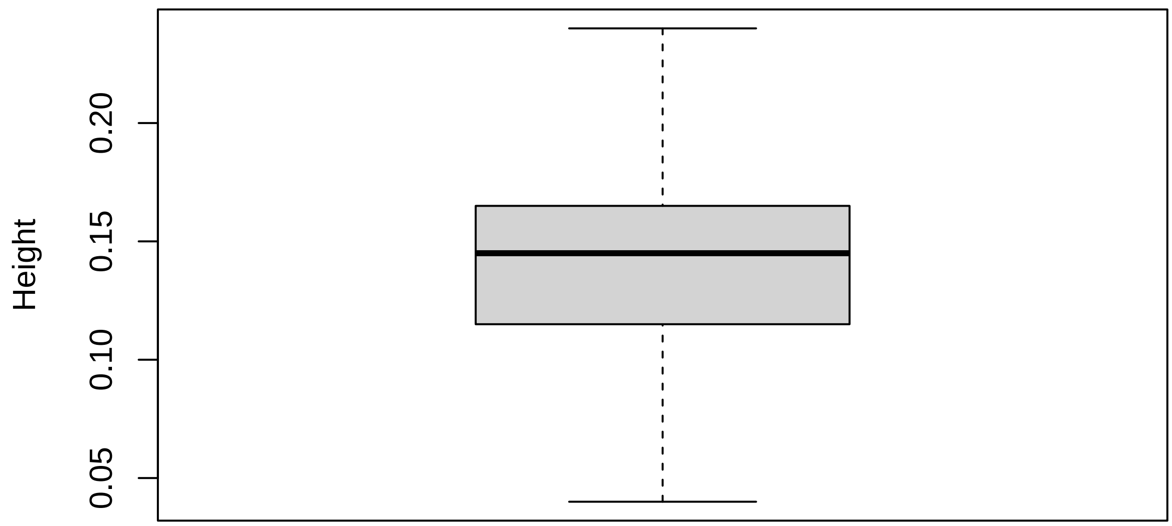
```
quartiles <- quantile(work_table$Length, probs=c(.25, .75), na.rm = FALSE)
IQR <- IQR(work_table$Length)

Lower <- quartiles[1] - 1.5*IQR
Upper <- quartiles[2] + 1.5*IQR

table_no_outlier <- subset(work_table, work_table$Length > Lower & work_table$Length < Upper)
boxplot(table_no_outlier$Length,
        ylab = "Length")
```



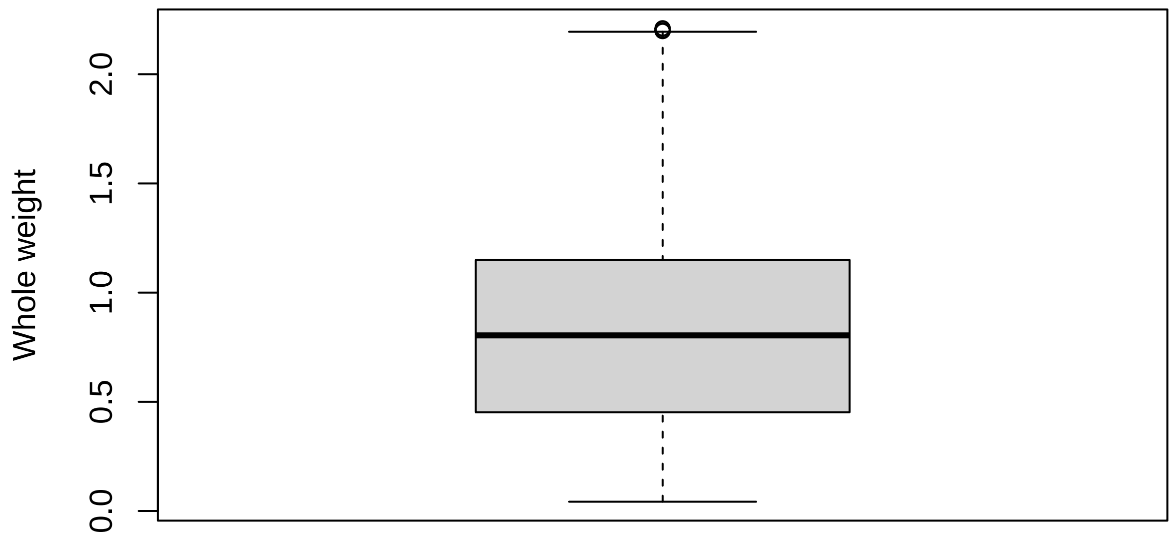

```
quartiles <- quantile(table_no_outlier$Diameter, probs=c(.25, .75), na.rm = FALSE)
IQR <- IQR(table_no_outlier$Diameter)

Lower <- quartiles[1] - 1.5*IQR
Upper <- quartiles[2] + 1.5*IQR

table_no_outlier <- subset(table_no_outlier, table_no_outlier$Diameter> Lower & table_no_outlier$Diamet
boxplot(table_no_outlier$Diameter,
        ylab = "Diameter")
```



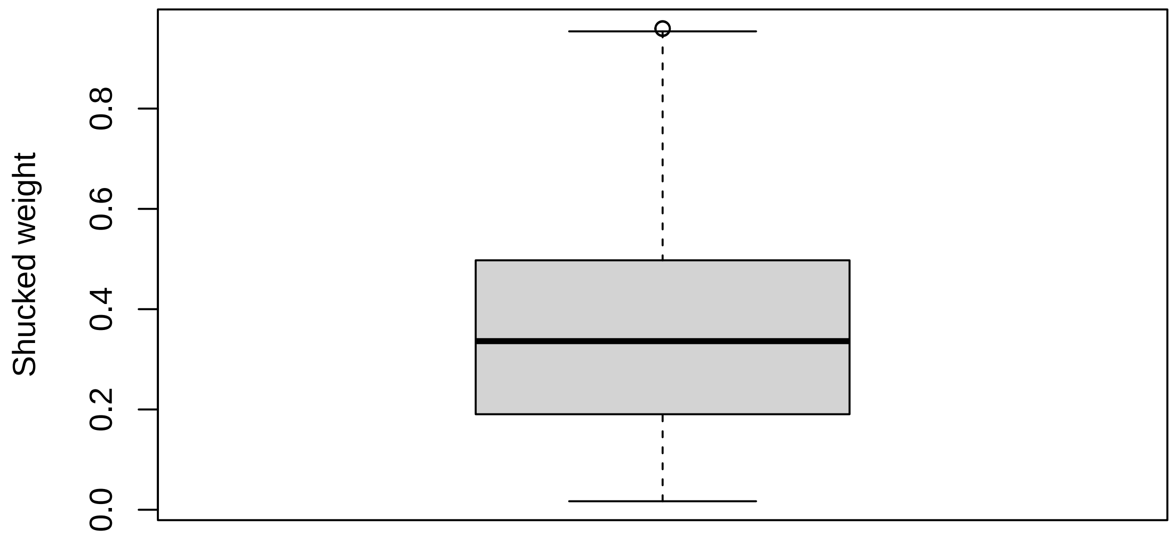

```
quartiles <- quantile(table_no_outlier$Height, probs=c(.25, .75), na.rm = FALSE)
IQR <- IQR(table_no_outlier$Height)

Lower <- quartiles[1] - 1.5*IQR
Upper <- quartiles[2] + 1.5*IQR

table_no_outlier <- subset(table_no_outlier, table_no_outlier$Height> Lower & table_no_outlier$Height <
boxplot(table_no_outlier$Height,
ylab = "Height")
```



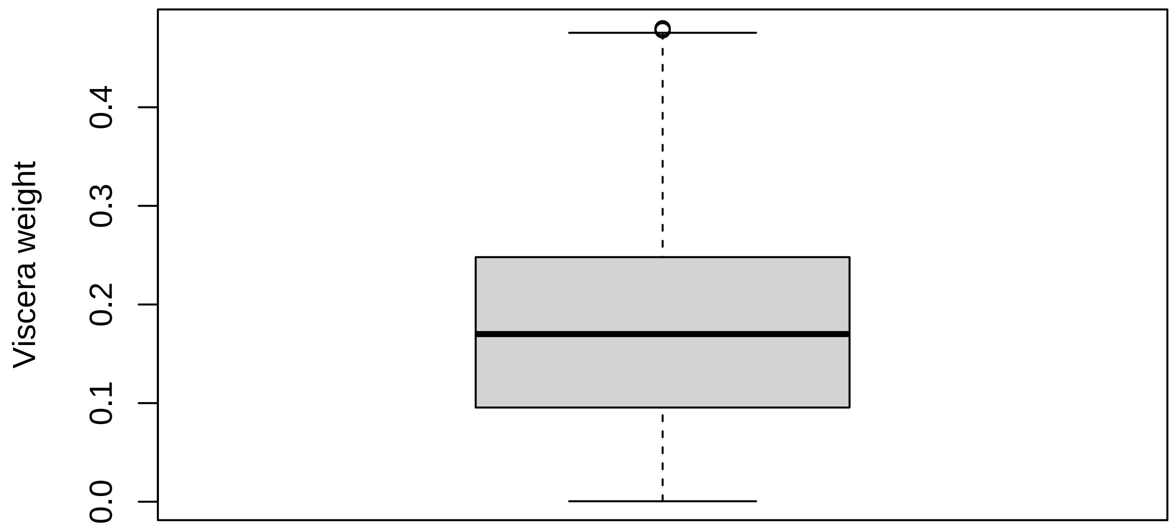

```
quartiles <- quantile(table_no_outlier$Whole_weight, probs=c(.25, .75), na.rm = FALSE)
IQR <- IQR(table_no_outlier$Whole_weight)

Lower <- quartiles[1] - 1.5*IQR
Upper <- quartiles[2] + 1.5*IQR

table_no_outlier <- subset(table_no_outlier, table_no_outlier$Whole_weight> Lower & table_no_outlier$Who
boxplot(table_no_outlier$Whole_weight,
        ylab = "Whole weight")
```



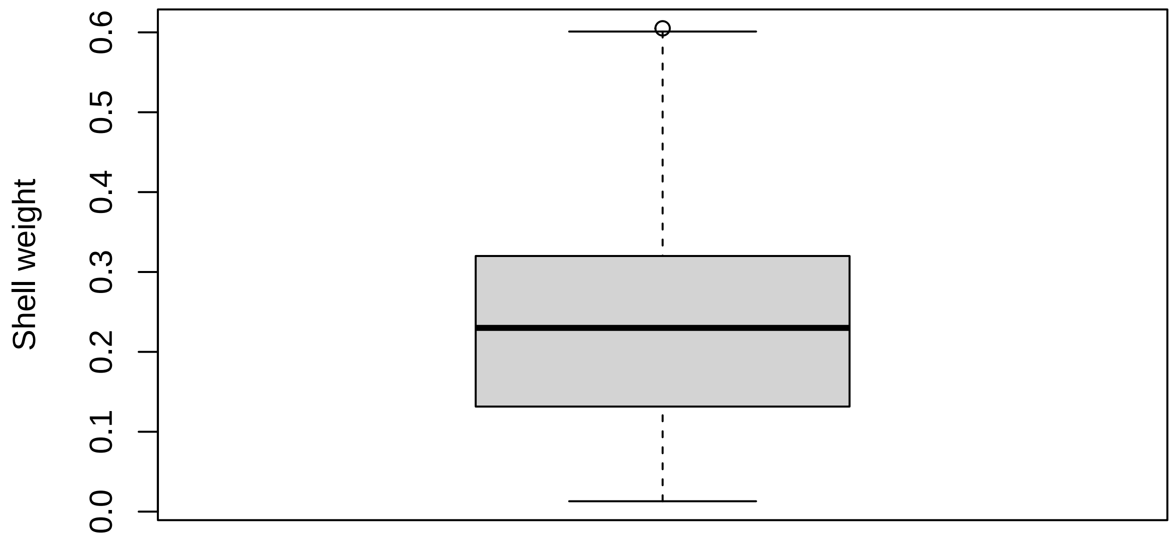

```
quartiles <- quantile(table_no_outlier$Shucked_weight, probs=c(.25, .75), na.rm = FALSE)
IQR <- IQR(table_no_outlier$Shucked_weight)

Lower <- quartiles[1] - 1.5*IQR
Upper <- quartiles[2] + 1.5*IQR

table_no_outlier <- subset(table_no_outlier, table_no_outlier$Shucked_weight> Lower & table_no_outlier$
boxplot(table_no_outlier$Shucked_weight,
        ylab = "Shucked weight")
```



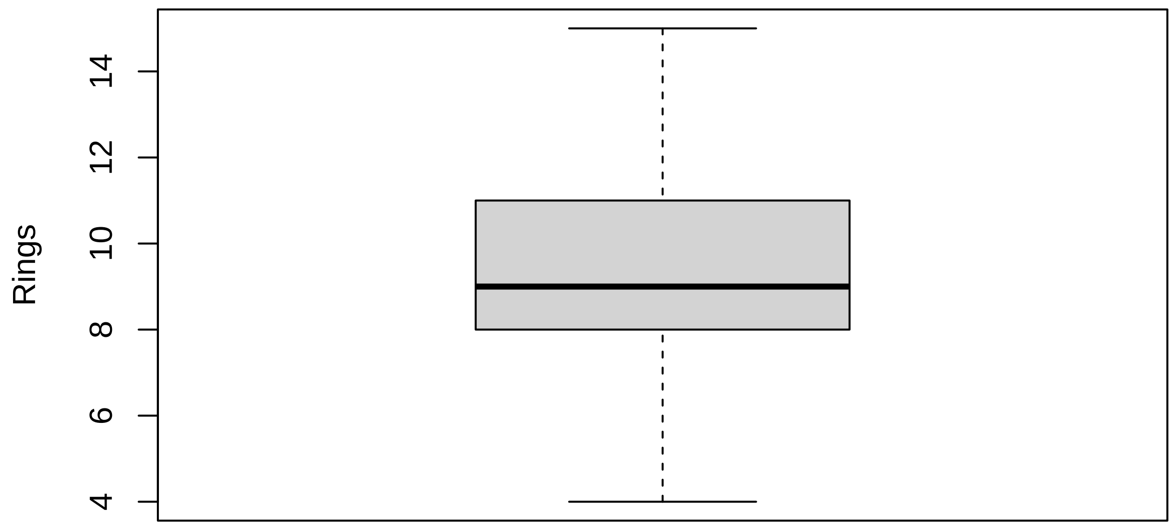

```
quartiles <- quantile(table_no_outlier$Viscera_weight, probs=c(.25, .75), na.rm = FALSE)
IQR <- IQR(table_no_outlier$Viscera_weight)

Lower <- quartiles[1] - 1.5*IQR
Upper <- quartiles[2] + 1.5*IQR

table_no_outlier <- subset(table_no_outlier, table_no_outlier$Viscera_weight> Lower & table_no_outlier$V
boxplot(table_no_outlier$Viscera_weight,
         ylab = "Viscera weight")
```




```
quartiles <- quantile(table_no_outlier$Shell_weight, probs=c(.25, .75), na.rm = FALSE)
IQR <- IQR(table_no_outlier$Shell_weight)

Lower <- quartiles[1] - 1.5*IQR
Upper <- quartiles[2] + 1.5*IQR

table_no_outlier <- subset(table_no_outlier, table_no_outlier$Shell_weight> Lower & table_no_outlier$She
boxplot(table_no_outlier$Shell_weight,
        ylab = "Shell weight")
```



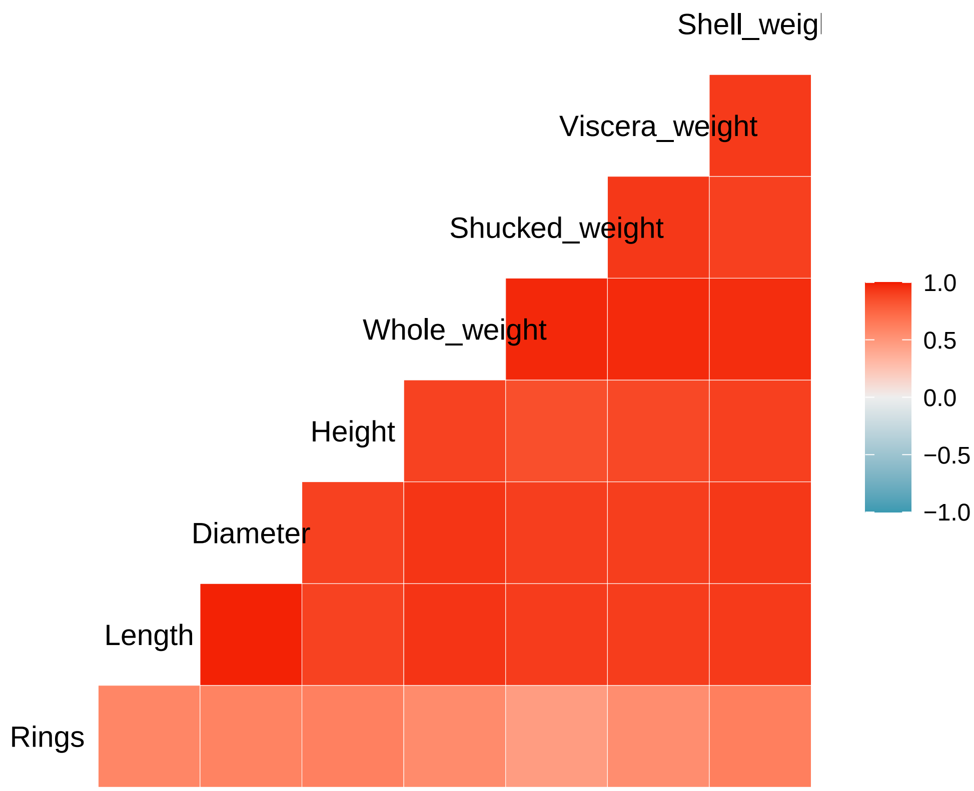

```
quartiles <- quantile(table_no_outlier$Rings, probs=c(.25, .75), na.rm = FALSE)
IQR <- IQR(table_no_outlier$Rings)

Lower <- quartiles[1] - 1.5*IQR
Upper <- quartiles[2] + 1.5*IQR

table_no_outlier <- subset(table_no_outlier, table_no_outlier$Rings> Lower & table_no_outlier$Rings < Up
boxplot(table_no_outlier$Rings,
        ylab = "Rings")
```



Seems like got rid of emissions, but it's not certain. Now let's look at the correlations. *Correlations*

```
ggcorr(table_no_outlier)
```

```
## Warning in ggcorr(table_no_outlier): data in column(s) 'Sex' are not numeric and
## were ignored
```



Cool, a strongly positive correlation is observed almost everywhere.

3 The standard deviation of the Length variable for molluscs of different sexes.

```
tapply(X = table_no_outlier$Length, INDEX = table_no_outlier$Sex, FUN = mean)
```

```
##    female      male    uvenil
## 0.5728231 0.5552010 0.4360096
```

```
tapply(X = table_no_outlier$Length, INDEX = table_no_outlier$Sex, FUN = sd)
```

```
##     female       male     uvenil
## 0.08447206 0.09693074 0.09721442
```

4 The percentage of molluscs that have a value of the Height variable less than 0,165.

```
x <- table_no_outlier[table_no_outlier$Height >= 0.165, ]
short <- dim(x)
all <- dim(table_no_outlier)
pr <- 100*short[1]/all[1]
pr
```

```
## [1] 25.21209
```

5 Values of the variable Length that are higher than 92% of all observations.

```r
Length <- c(table_no_outlier$Length)
q <- quantile(table_no_outlier$Length, probs= c(0.92, 0.92), na.rm = FALSE)
q
```

```
##  92%  92%
## 0.66 0.66
```

6 Z-transformation of Length

```r
m <- mean(Length)
s <- sd(Length)
Lenght_z_scores <- (Length-m)/s
```

Perhaps there is a standard feature for it, but I could not find it.

7 Comparison between the diameter of clams with the number of rings 5 and 15.

```r
five_rings <- subset(table_no_outlier, table_no_outlier$Rings == 5)
summary(five_rings$Diameter)
```

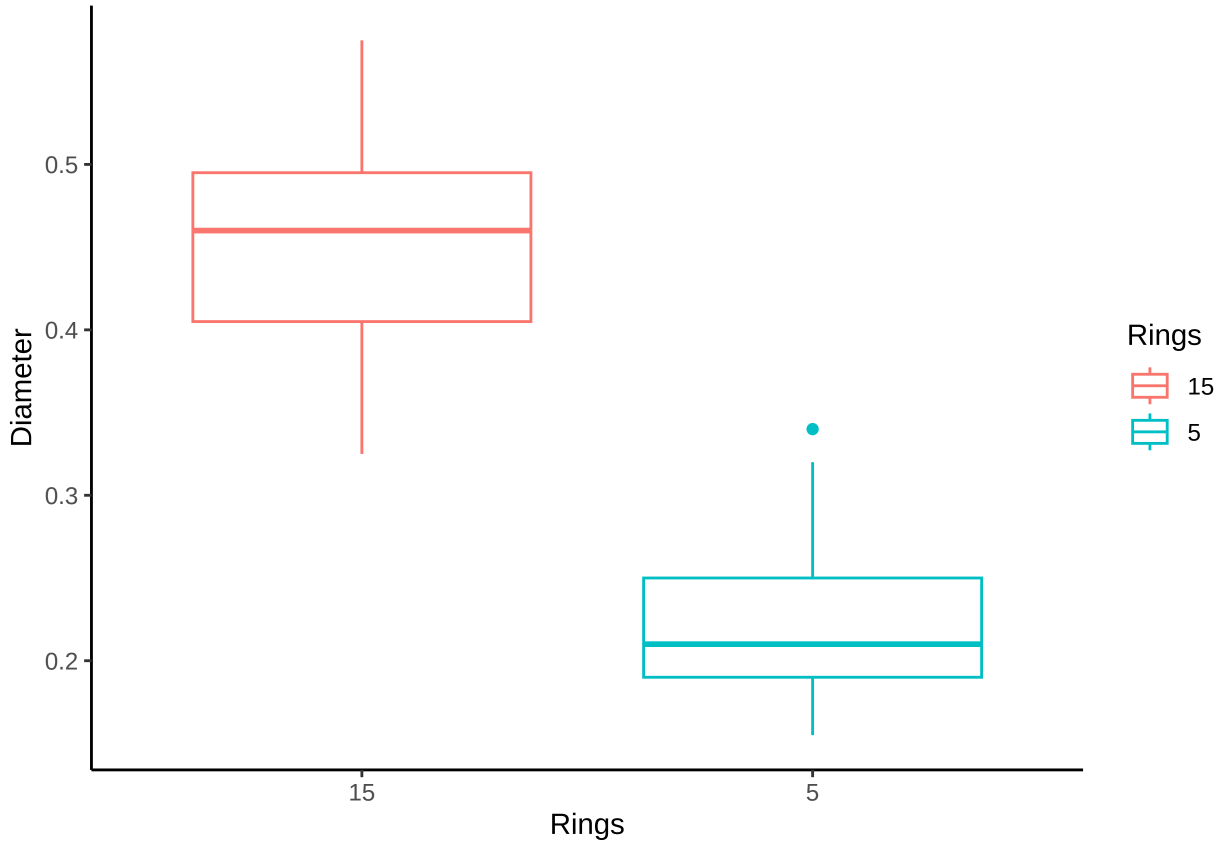
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1550  0.1900  0.2100  0.2212  0.2500  0.3400
```

```r
fiveteen_rings <- subset(table_no_outlier, table_no_outlier$Rings == 15)
summary(fiveteen_rings$Diameter)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3250  0.4050  0.4600  0.4551  0.4950  0.5750
```

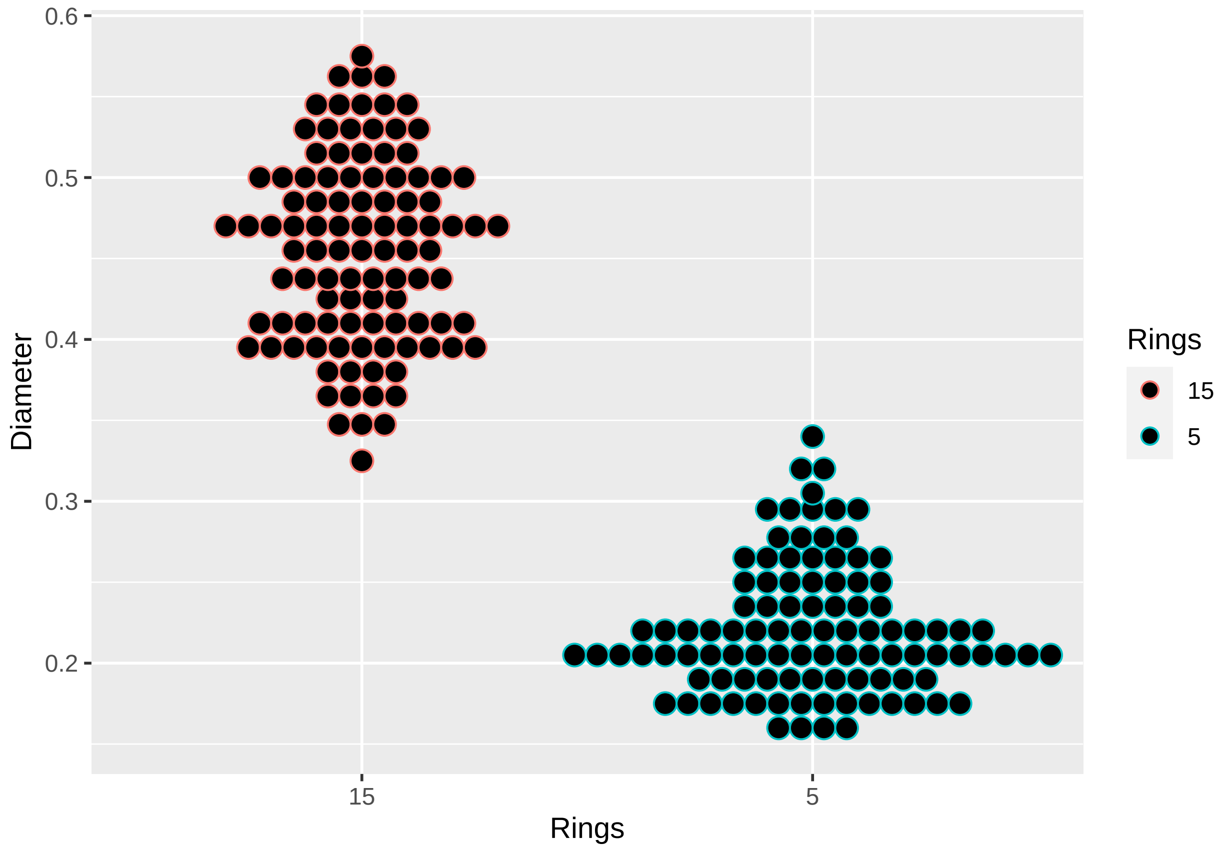
```r
f_rings <- subset(table_no_outlier, table_no_outlier$Rings == 5 | table_no_outlier$Rings == 15)
f_rings$Rings <- as.character(f_rings$Rings)
```

```r
ggplot(f_rings, aes(x = Rings, y = Diameter, color = Rings)) +
  geom_boxplot() +
  theme_classic()
```




```
ggplot(f_rings, aes(x = Rings, y = Diameter, color = Rings)) +
  geom_dotplot(binaxis='y', stackdir='center')
```

```
## Bin width defaults to 1/30 of the range of the data. Pick better value with
## `binwidth`.
```



You can choose which one to use in your publication. Conclusions: Diameter of mollusks has a positive correlation with the number of rings. The median for molluscs with 5 rings is 0.2100, and with 15 rings 0.4600. The difference is more than twofold.

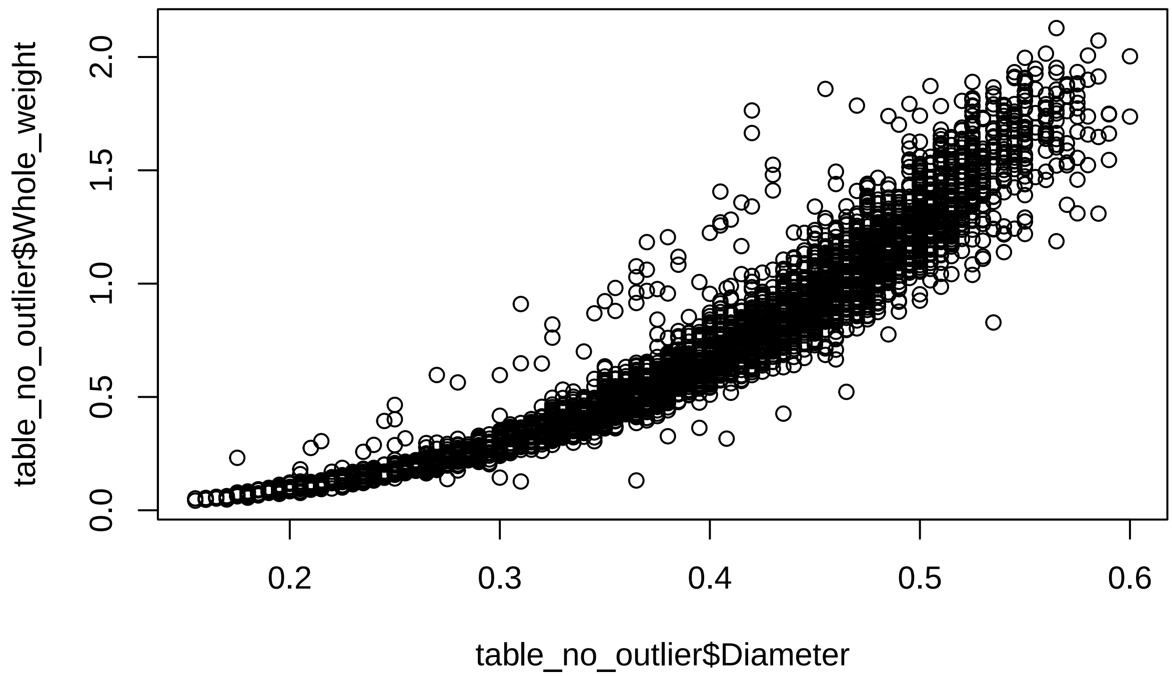
8 Diameter and Whole_weight

```
cor.test(table_no_outlier$Diameter, table_no_outlier$Whole_weight)
```

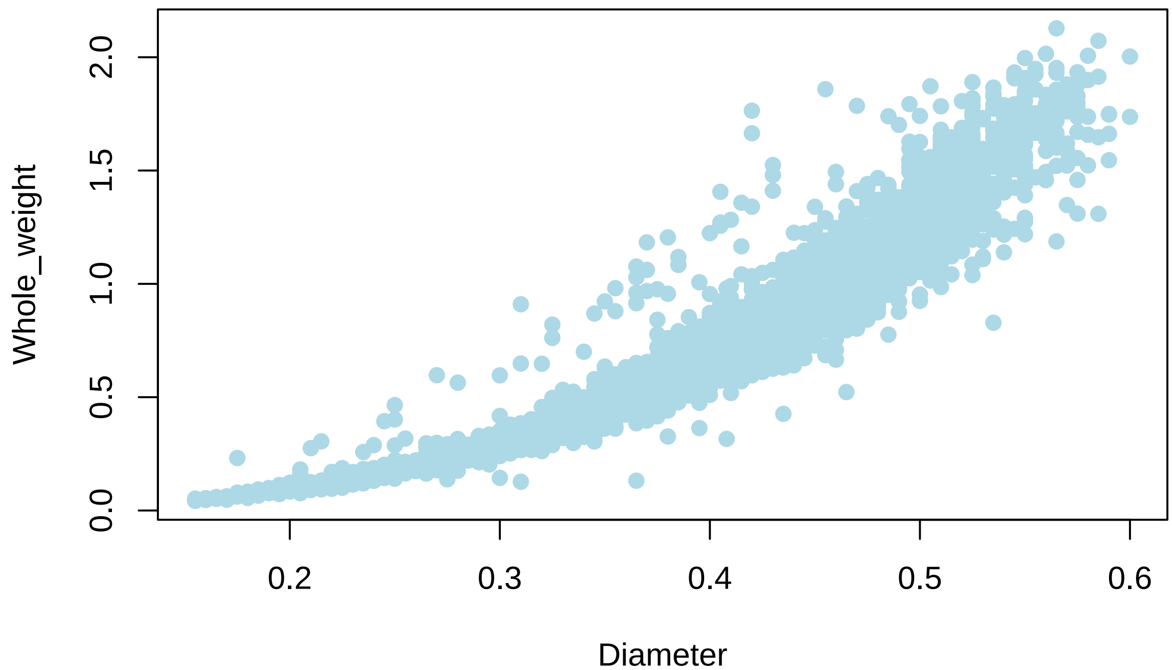
```
##
##  Pearson's product-moment correlation
##
## data:  table_no_outlier$Diameter and table_no_outlier$Whole_weight
## t = 167.29, df = 3770, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9348626 0.9424468
## sample estimates:
##       cor
## 0.9387684
```

There is a strong positive correlation.

```
plot(table_no_outlier$Diameter, table_no_outlier$Whole_weight)
```

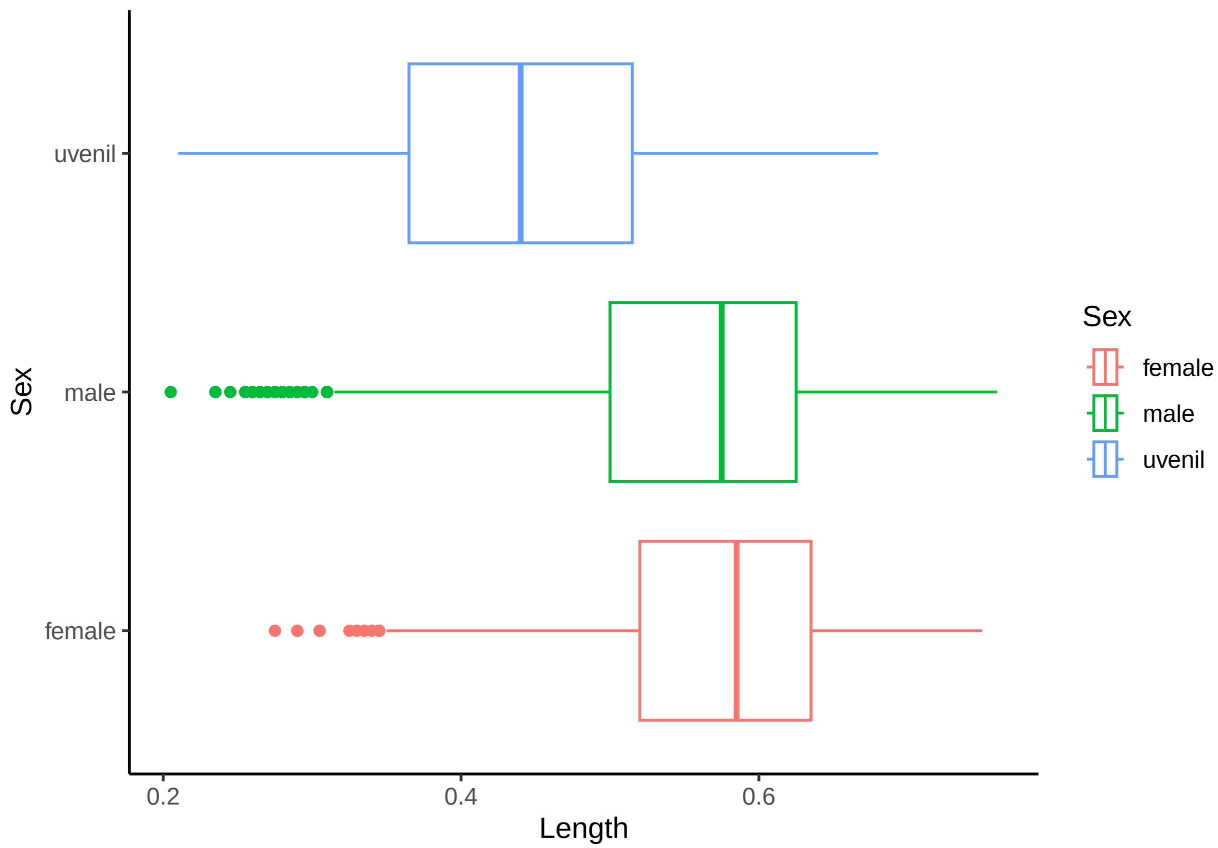



```
Diameter <- table_no_outlier$Diameter
Whole_weight <- table_no_outlier$Whole_weight
plot(Diameter, Whole_weight, pch = 19, col = "lightblue")
```

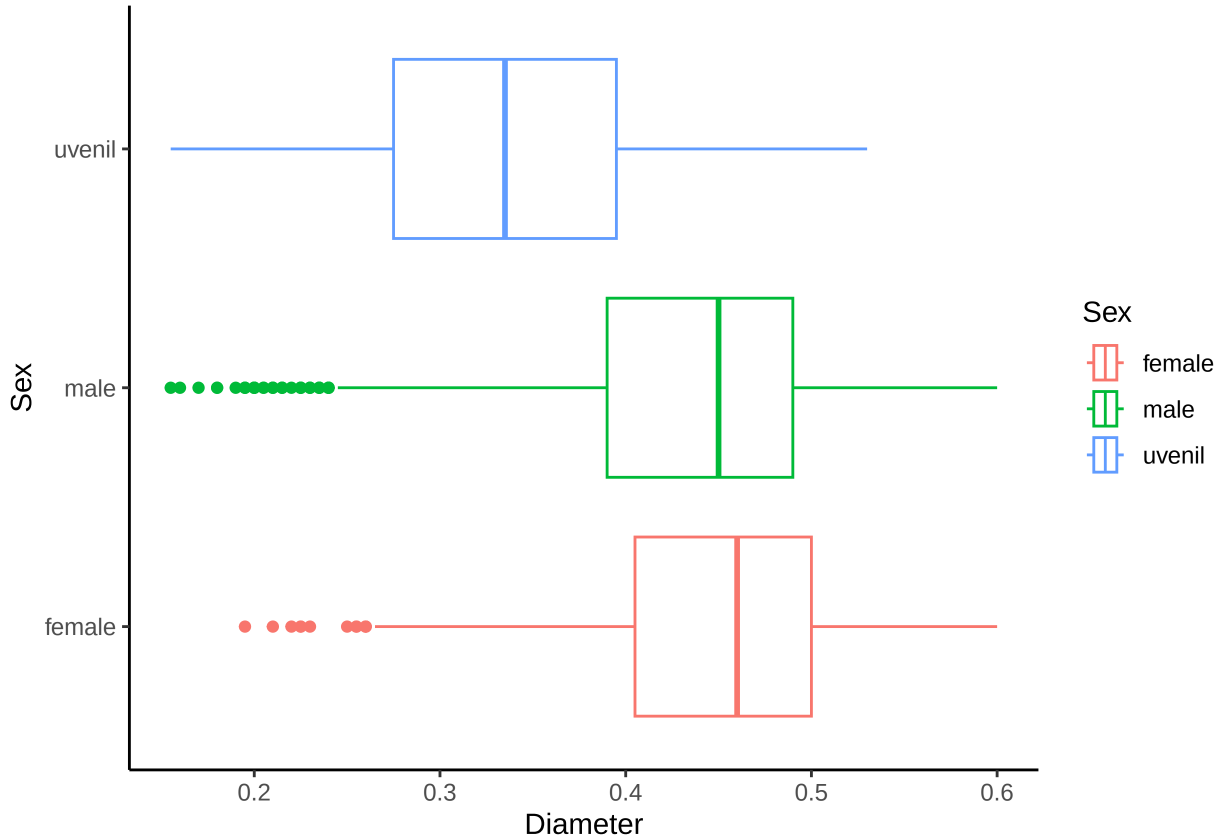


9 My suggestions Correlation with sex I was curious to see if there was a correlation with gender, so I decided to check it out.

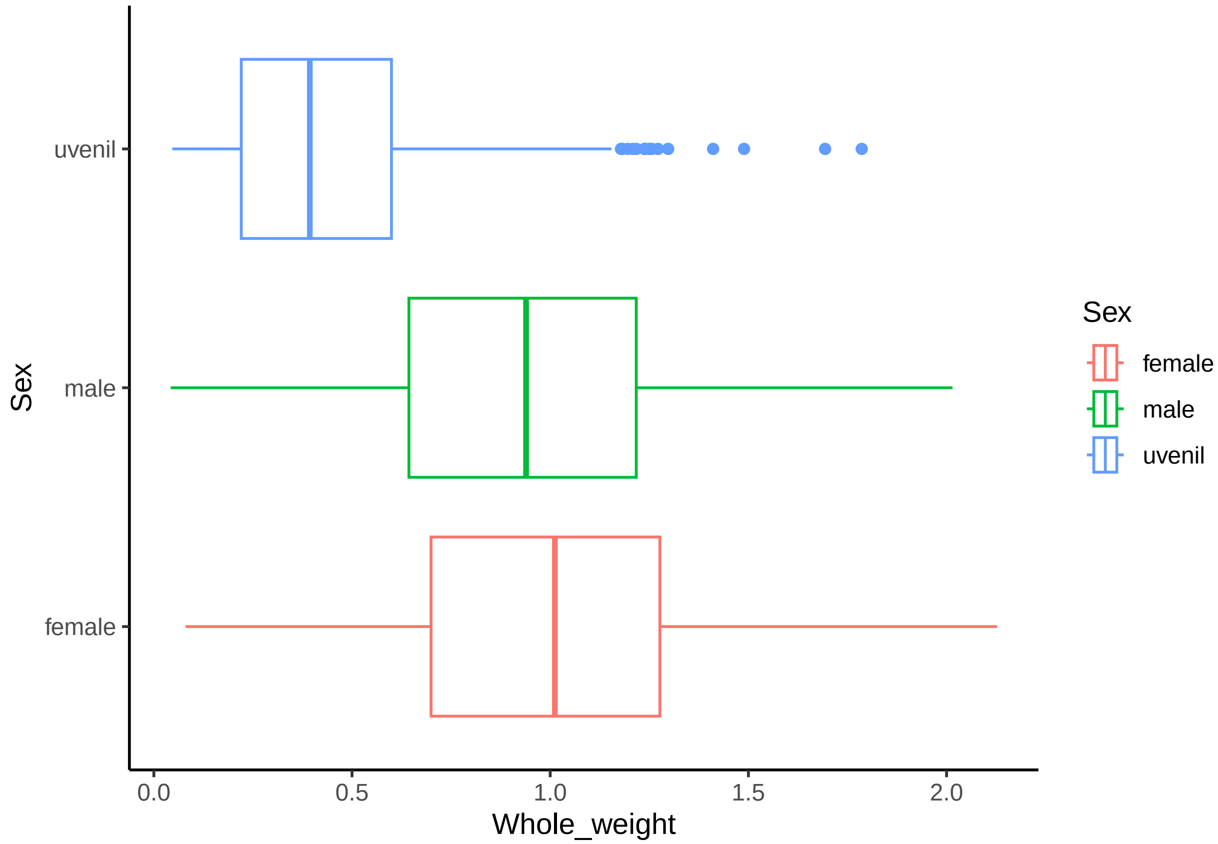
```
ggplot(table_no_outlier, aes(x = Length, y = Sex, color = Sex)) +
  geom_boxplot() +
  theme_classic()
```



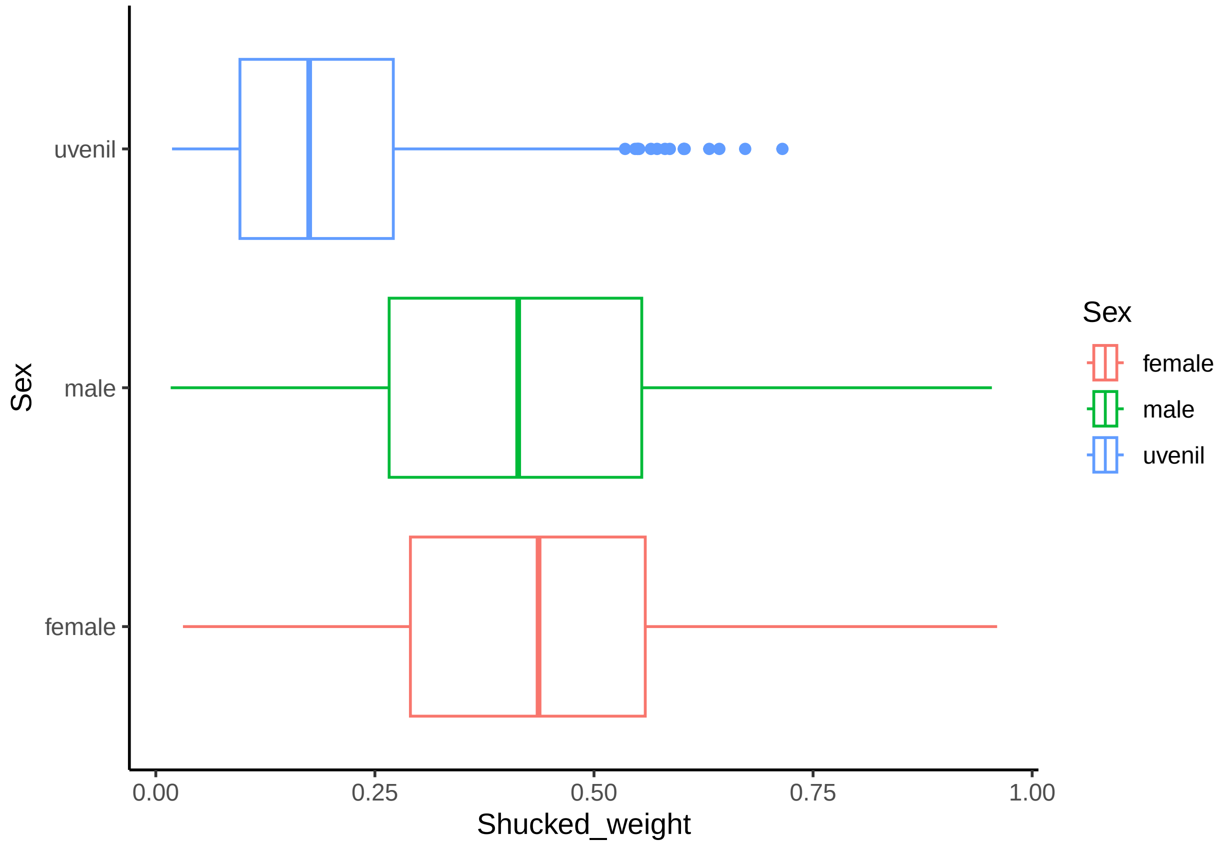

```r
ggplot(table_no_outlier, aes(x = Diameter, y = Sex, color = Sex)) +
  geom_boxplot() +
  theme_classic()
```



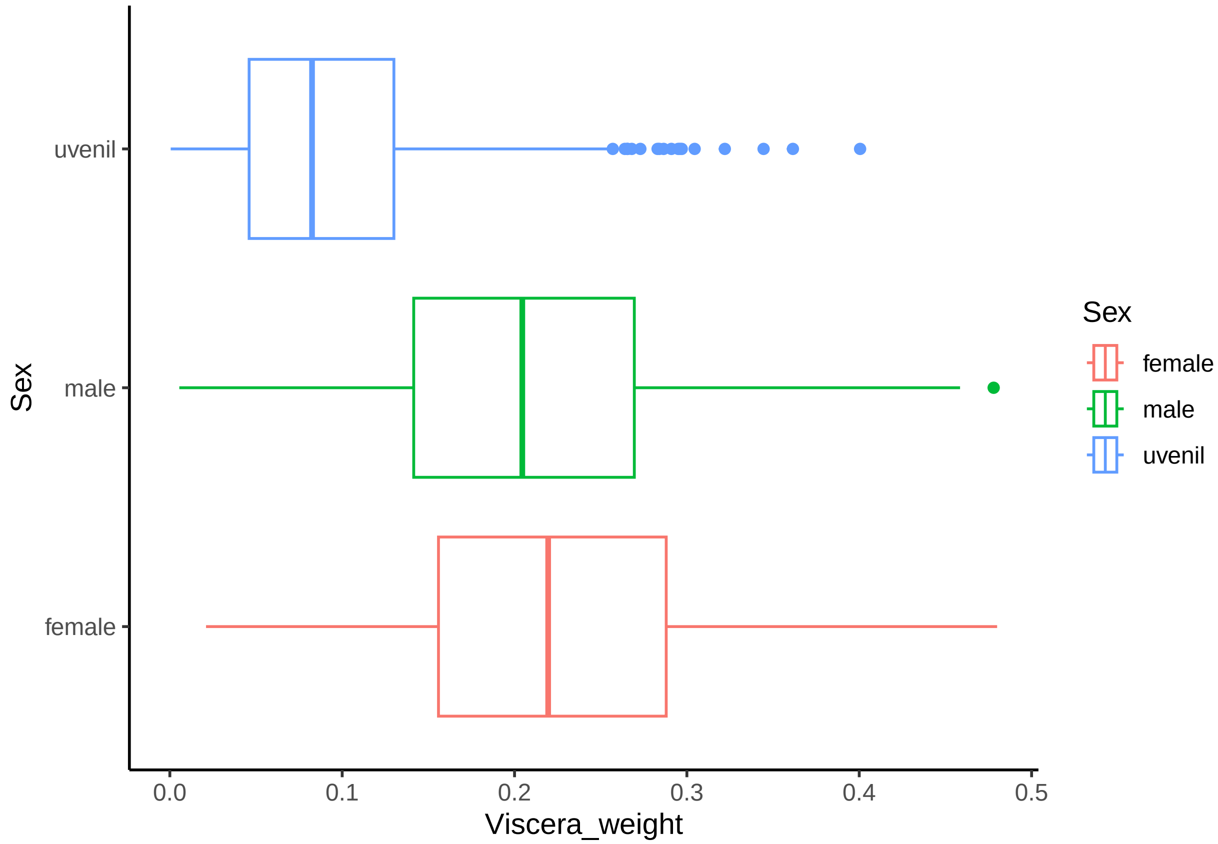

```
ggplot(table_no_outlier, aes(x = Whole_weight, y = Sex, color = Sex)) +
  geom_boxplot() +
  theme_classic()
```



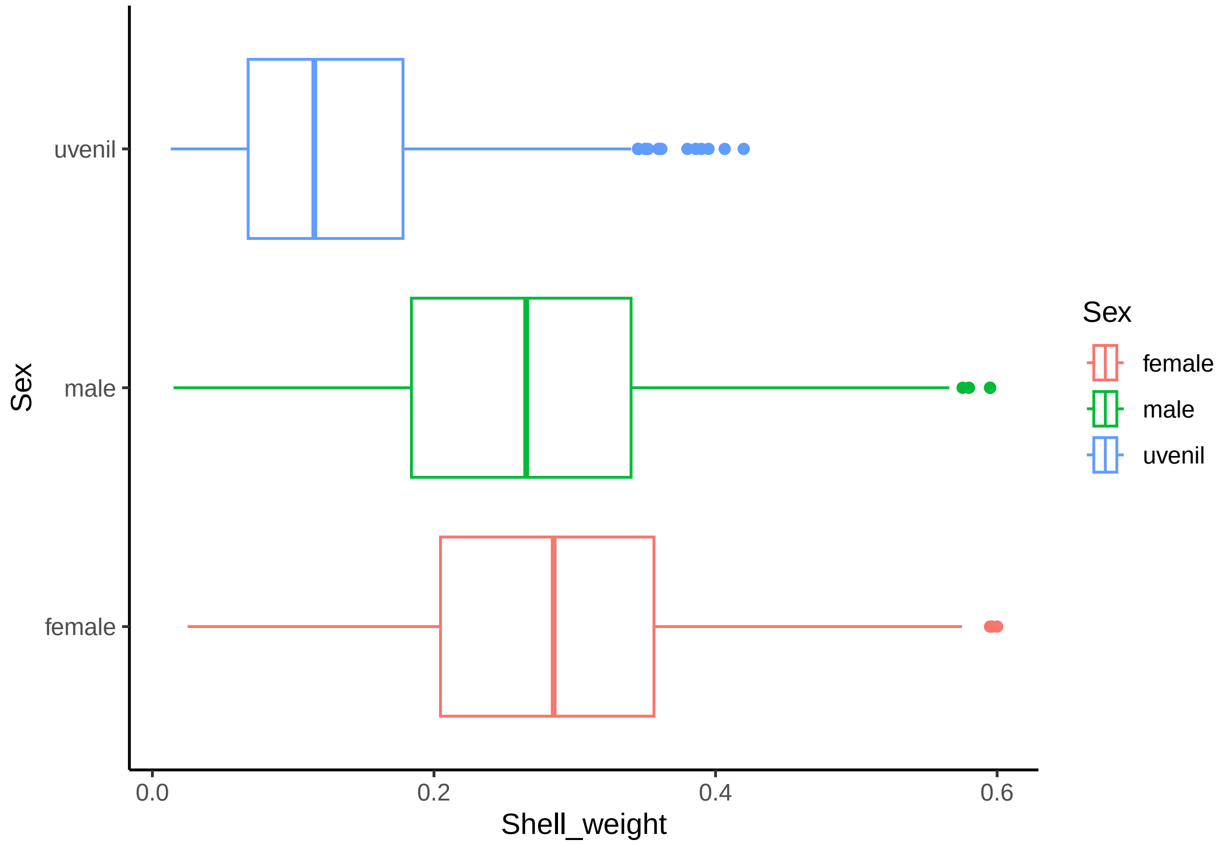

```
ggplot(table_no_outlier, aes(x = Shucked_weight, y = Sex, color = Sex)) +
  geom_boxplot() +
  theme_classic()
```




```
ggplot(table_no_outlier, aes(x = Viscera_weight, y = Sex, color = Sex)) +
  geom_boxplot() +
  theme_classic()
```




```
ggplot(table_no_outlier, aes(x = Shell_weight, y = Sex, color = Sex)) +
  geom_boxplot() +
  theme_classic()
```



Conclusions: There is no correlation between females and males, but there is a correlation between sexually mature and nonsexually mature individuals.