Midia_B

A***_B****

2022-11-21

Collecting the data

Since every year students collect the same data, its quite convenient to have a function for merging all little datasets in one:

```
merge_csv <- function(folder_path){
  return(list.files(path = folder_path, pattern = "*.csv", full.names = TRUE)%>% read_csv %>% bind_rows
}
```

and use it like this:

```
df <- merge_csv("/home/bananna/Downloads/Rproject1/Data")
```

```
## Rows: 4177 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (9): Rings, Sex (1 - male, 2 - female, 3 - uvenil), Length, Diameter, He...
##
## 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.
```

```
df
```

```
## # A tibble: 4,177 x 9
##    Rings Sex (1 - male, ~1 Length Diame~2 Height Whole~3 Shuck~4 Visce~5 Shell~6
##    <dbl>             <dbl>  <dbl>   <dbl>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1    18                 2  0.575   0.45   0.17    1.05   0.378   0.170   0.385
##  2    14                 3  0.385   0.305  0.095   0.252  0.0915  0.055   0.09
##  3     8                 1  0.475   0.37   0.125   0.649  0.347   0.136   0.142
##  4    12                 1  0.665   0.525  0.18    1.43   0.672   0.29    0.4
##  5     4                 3  0.28    0.12   0.075   0.117  0.0455  0.029   0.0345
##  6     4                 3  0.22    0.16   0.05    0.049  0.0215  0.01    0.015
##  7    12                 1  0.72    0.565  0.2     2.11   1.02    0.363   0.494
##  8     8                 2  0.55    0.43   0.15    0.655  0.264   0.122   0.221
##  9     6                 3  0.235   0.175  0.055   0.067  0.027   0.0125  0.018
## 10     8                 2  0.4     0.3    0.115   0.302  0.134   0.0465  0.0935
## # ... with 4,167 more rows, and abbreviated variable names
## #   1: `Sex (1 - male, 2 - female, 3 - uvenil)`, 2: Diameter, 3: Whole_weight,
## #   4: Shucked_weight, 5: Viscera_weight, 6: Shell_weight
```

we can make it just in one row like this:

```
#df <- list.files(path = "/home/bananna/Downloads/Rproject1/Data", pattern = "*.csv", full.names = TRUE
```

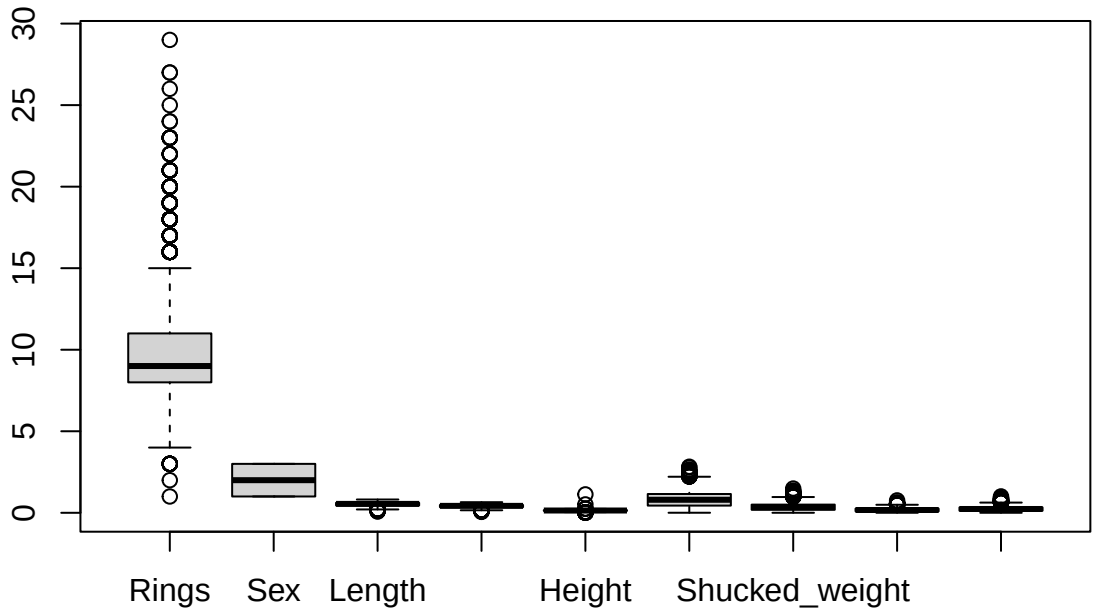
datasets were not quite the same, but bind_rows() automatically changes them to an appropriate type of data. However, 'Sex' column stays numeric, while it should be factor. Also the name is too long to my taste, and it also has spaces, which is not cool. So we will change that column:

```
df <- rename(df, Sex = 'Sex (1 - male, 2 - female, 3 - uvenil)')
df <- mutate(df,Sex = factor(Sex))
levels(df$Sex) <- c('Male', 'Female', 'Uvenile')
```

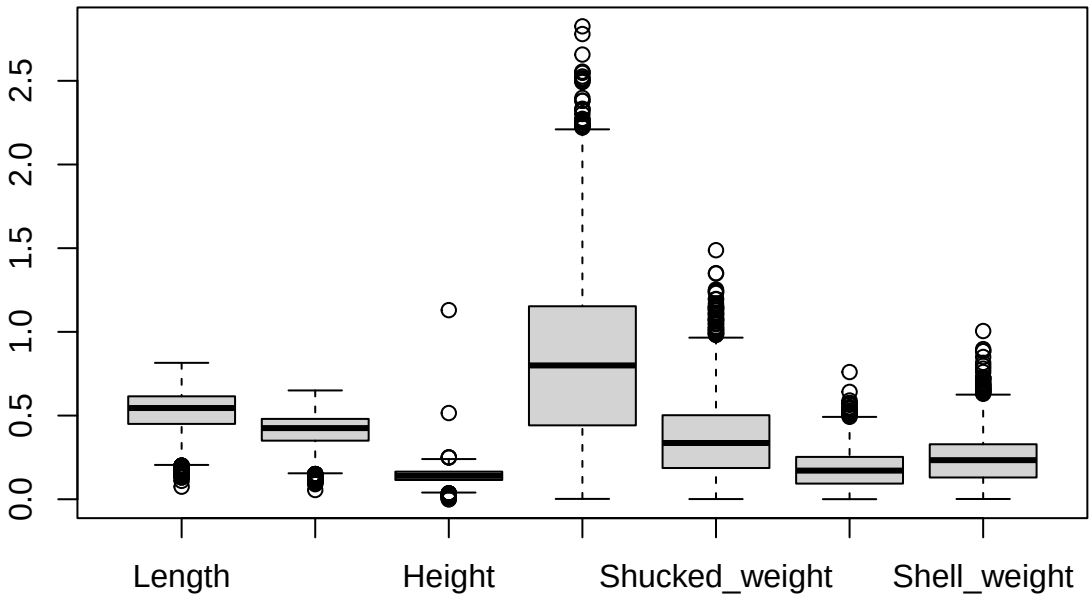
EDA

Outliers can be visualized with boxplots:

```
boxplot(df)
```

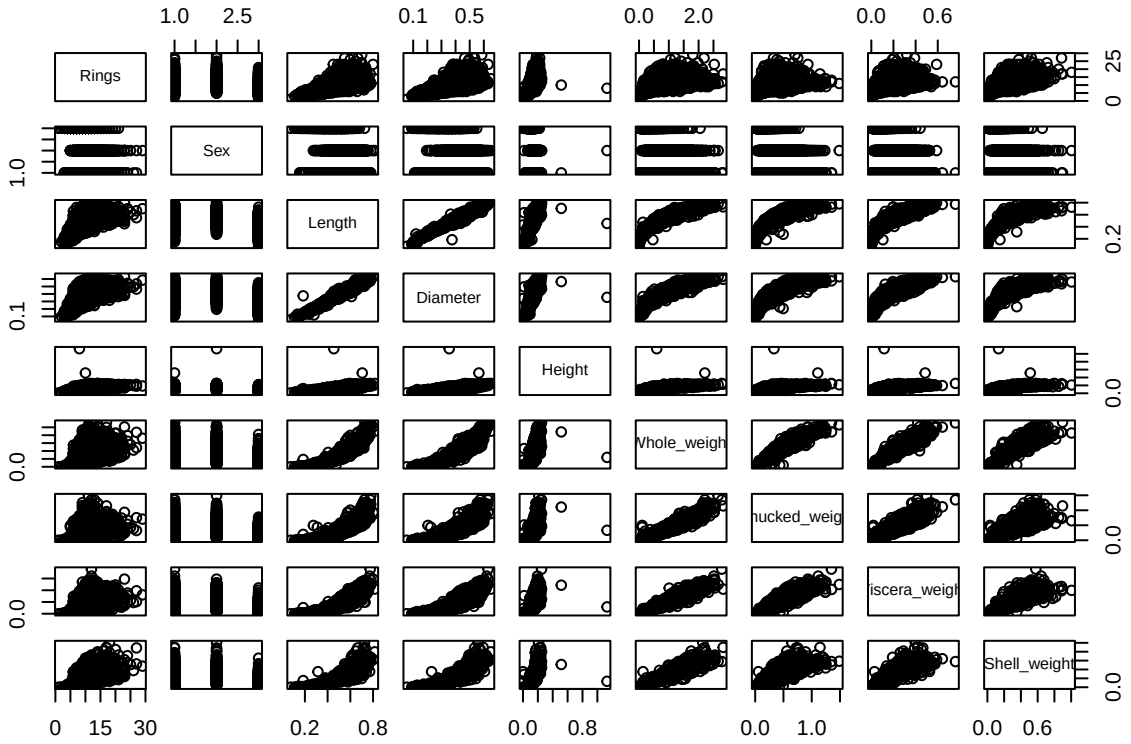



```
boxplot(df[,-(1:2)])
```



it seems like we have outliers everywhere. Maybe it's because that's not normal distribution what we have as a data. To notice the relationship of the data to each other let's look at all the possible plots:

```
pairs(df)
```



We can see that length and diameter are linearly correlated, and this seems logical. Seems like all weights are linearly correlated too. But the age, or the rings amount, behaves strangely. Sex doesnt seem to be meaningful for now.

Hypothesis:

1. Age doesnt really influence the size and the weight of midia afters some point. While the uvenile age it grows gradually, but then it can stop growing, for example, because of lack of food or high competition or smth else...
2. Length and Diameters are highly correlated, it might be good to include only one factor in some model.
3. Midias grow evenly: if the shell weights more, it means the visceral part weights more as well.
4. Sex (except of uvenile stage) doesnt mean much (we need some more tests about this).
5. Midia stops growing in diameter after some point of gained weight. The only explanation of this - it grows in height. Also maybe the weight grows and sizes don't because of some parasites on the shell of old midias or pearls inside? :D

Correlations We can perform a fast analises of correlations between variables. We have variable Rings which can be considered both as numeric and categorial one. also we have categorial Sex. We can perform Chi-test on Rings and Sex later, and ANOVA with sex and numeric variables.

```
df2 <- df[,-2]
corr.test(df2)
```

```
## Call:corr.test(x = df2)
## Correlation matrix
##                Rings Length Diameter Height Whole_weight Shucked_weight
## Rings           1.00   0.56     0.57   0.56         0.54           0.42
## Length          0.56   1.00     0.99   0.83         0.93           0.90
## Diameter        0.57   0.99     1.00   0.83         0.93           0.89
## Height          0.56   0.83     0.83   1.00         0.82           0.77
## Whole_weight    0.54   0.93     0.93   0.82         1.00           0.97
## Shucked_weight  0.42   0.90     0.89   0.77         0.97           1.00
## Viscera_weight  0.50   0.90     0.90   0.80         0.97           0.93
## Shell_weight    0.63   0.90     0.91   0.82         0.96           0.88
##                Viscera_weight Shell_weight
## Rings                    0.50         0.63
## Length                   0.90         0.90
## Diameter                 0.90         0.91
## Height                   0.80         0.82
## Whole_weight             0.97         0.96
## Shucked_weight           0.93         0.88
## Viscera_weight           1.00         0.91
## Shell_weight             0.91         1.00
## Sample Size
##                Rings Length Diameter Height Whole_weight Shucked_weight
## Rings           4176   4169     4171   4174         4175           4173
## Length          4169   4170     4165   4168         4169           4167
## Diameter        4171   4165     4172   4170         4171           4169
## Height          4174   4168     4170   4175         4174           4172
## Whole_weight    4175   4169     4171   4174         4176           4173
## Shucked_weight  4173   4167     4169   4172         4173           4174
## Viscera_weight  4176   4170     4172   4175         4176           4174
## Shell_weight    4174   4168     4170   4173         4174           4172
##                Viscera_weight Shell_weight
## Rings                    4176         4174
## Length                   4170         4168
## Diameter                 4172         4170
## Height                   4175         4173
## Whole_weight             4176         4174
## Shucked_weight           4174         4172
## Viscera_weight           4177         4175
## Shell_weight             4175         4175
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##                Rings Length Diameter Height Whole_weight Shucked_weight
## Rings              0      0        0      0            0              0
## Length             0      0        0      0            0              0
## Diameter           0      0        0      0            0              0
## Height             0      0        0      0            0              0
## Whole_weight       0      0        0      0            0              0
## Shucked_weight     0      0        0      0            0              0
## Viscera_weight     0      0        0      0            0              0
## Shell_weight       0      0        0      0            0              0
##                Viscera_weight Shell_weight
## Rings                       0            0
## Length                      0            0
## Diameter                    0            0
## Height                      0            0
```

```
## Whole_weight                0           0
## Shucked_weight              0           0
## Viscera_weight              0           0
## Shell_weight                0           0
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

Chi-test on Rings and Sex shows a significant correlation:

```
chisq.test(df$Sex, as.factor(df$Rings))
```

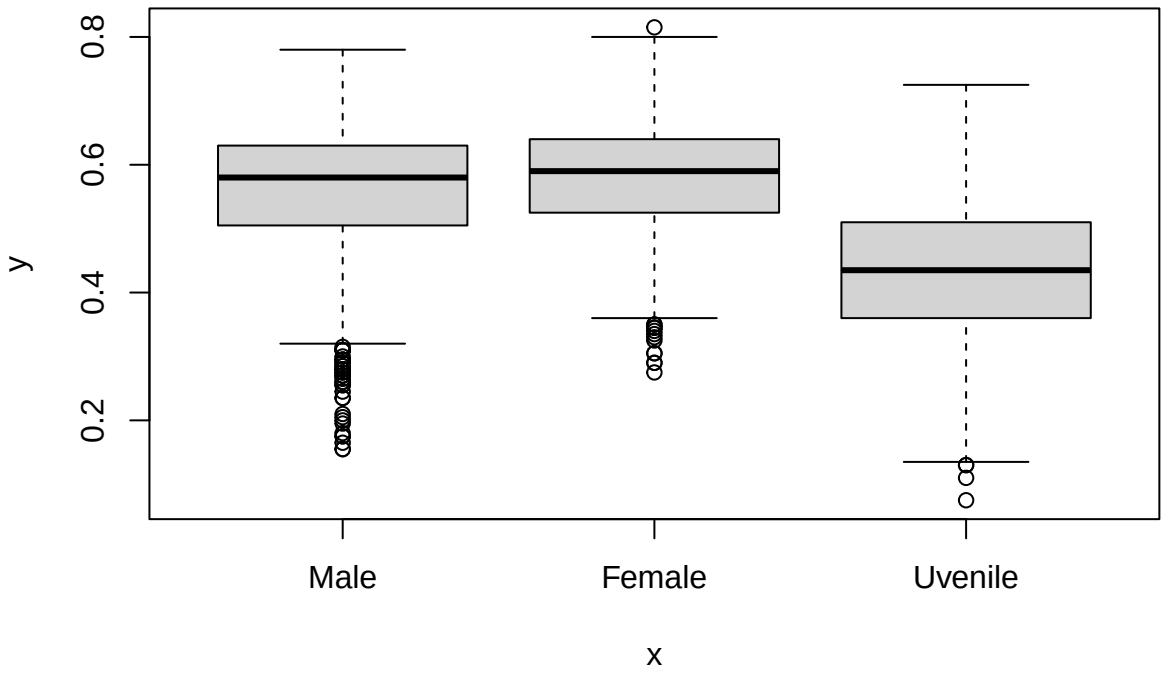
```
##
##  Pearson's Chi-squared test
##
## data:  df$Sex and as.factor(df$Rings)
## X-squared = 1314.9, df = 54, p-value < 2.2e-16
```

It means female midias live longer

Sex and Length

Let's chech if Sex influences Length:

```
plot(df$Sex,df$Length)
```



```
lngth_sex_means <- c(mean(df[as.numeric(df$Sex)==1,]$Length, na.rm = T), mean(df[as.numeric(df$Sex)==2,]
lngth_sex_sds <- c(sd(df[as.numeric(df$Sex)==1,]$Length, na.rm = T), sd(df[as.numeric(df$Sex)==2,]$Leng
as.data.frame(lngth_sex_means, row.names = c('Male', 'Female', 'Uvenile'))
```

```
##         lngth_sex_means
## Male          0.5616306
## Female        0.5791750
## Uvenile       0.4277782
```

```
as.data.frame(lngth_sex_sds, row.names = c('Male', 'Female', 'Uvenile'))
```

```
##         lngth_sex_sds
## Male       0.10262709
## Female     0.08625859
## Uvenile    0.10897681
```

```
summary(aov(df$Length ~ df$Sex)) # ANOVA analyses
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## df$Sex         2  18.52   9.261   926.4 <2e-16 ***
## Residuals   4163  41.62   0.010
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 11 observations deleted due to missingness
```

So, as we can see, length, or size is different in m, f and u groups.

Height < 0.165

```
length(which(df$Height<0.165))/length(df$Height)*100
```

```
## [1] 71.22337
```

So, approximately 71% of midias are smaller then 0.165. It means that (maybe) the most of them dyes before reaching the height maximum, and it can also mean that the most of them are uvenile. lets check:

```
nrow(df[df$Sex=="Uvenile",])/nrow(df)*100 # percent(%) of Uvenile midias
```

```
## [1] 32.20014
```

```
nrow(df[df$Sex=="Male",])/nrow(df)*100 # percent(%) of Male midias
```

```
## [1] 36.62916
```

```
nrow(df[df$Sex=="Female",])/nrow(df)*100 # percent(%) of Female midias
```

```
## [1] 31.36222
```

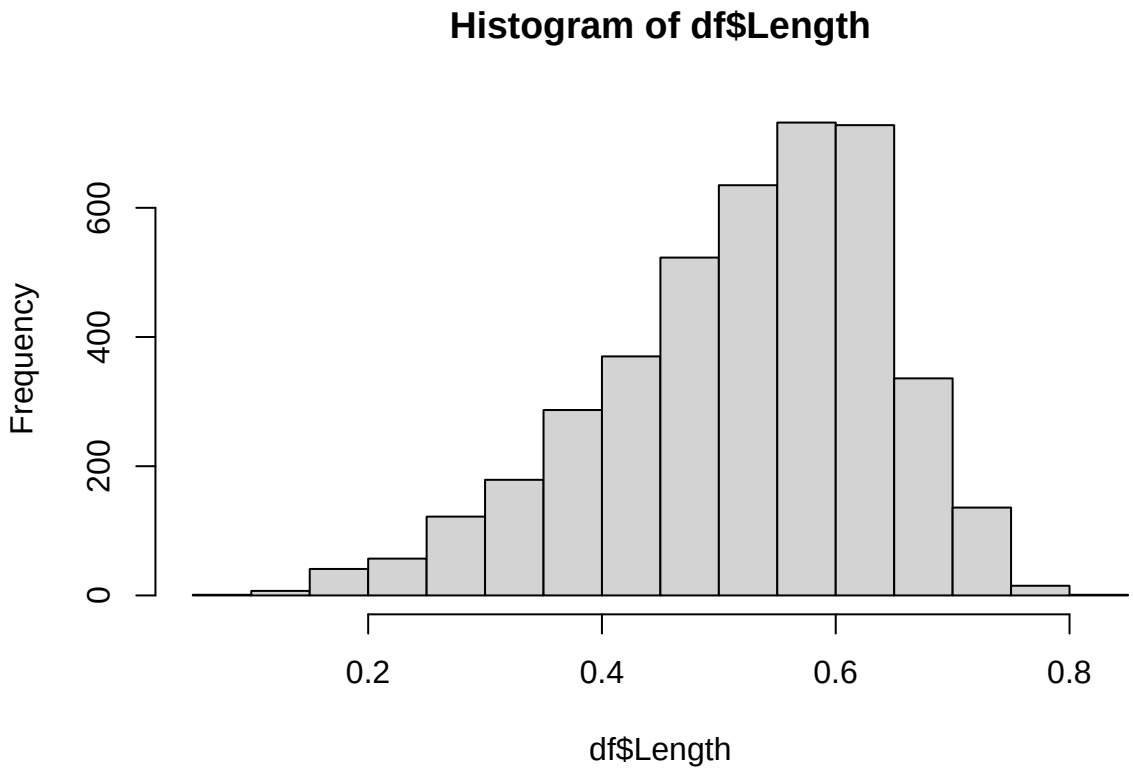
One third of all midaias are uvenile, not 71%, so most of midias are just small (in height), but not necessarily uvenile.

Length > 92%

```
sort(df$Length)[round(92/100*length(df$Length))]
```

```
## [1] 0.67
```

```
hist(df$Length)
```



We can asuume that most midias reach approximately 0.62 in Length and dont grow bigger. But thats not the maximum size. The distribution is not normal, that corfirms the previous acsessment:

```
shapiro.test(df$Length)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$Length
## W = 0.96952, p-value < 2.2e-16
```

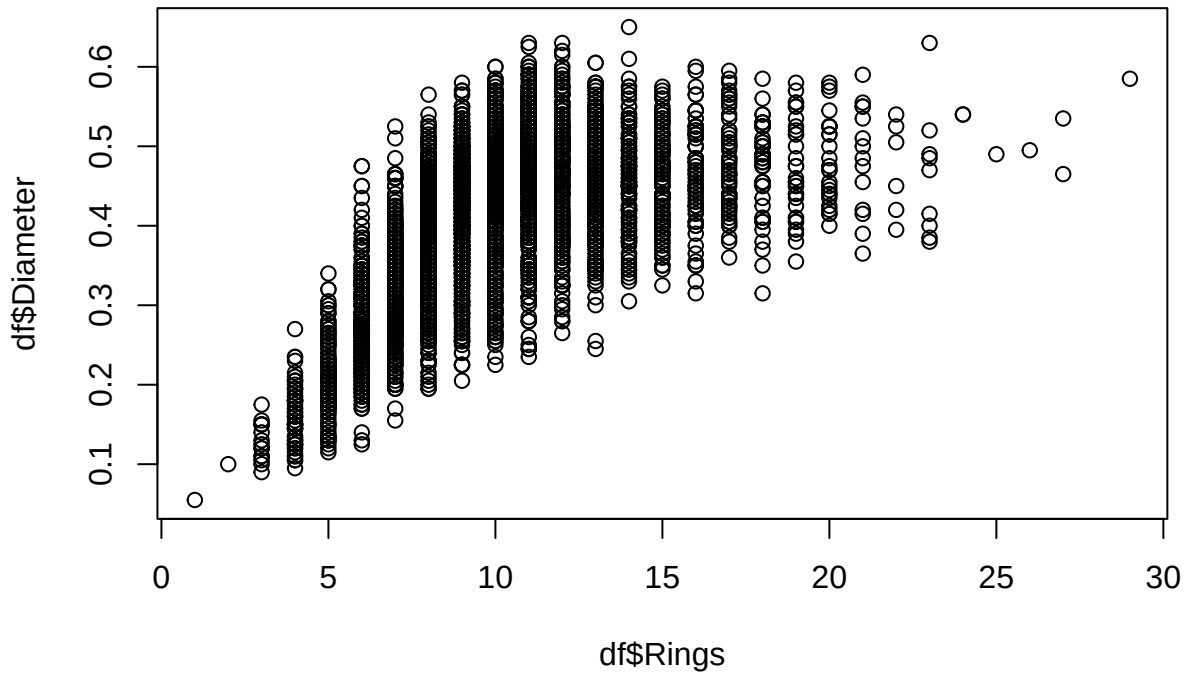
Standardized Length

```
Length_z_scores <- scale(df$Length)
summary(Length_z_scores)
```

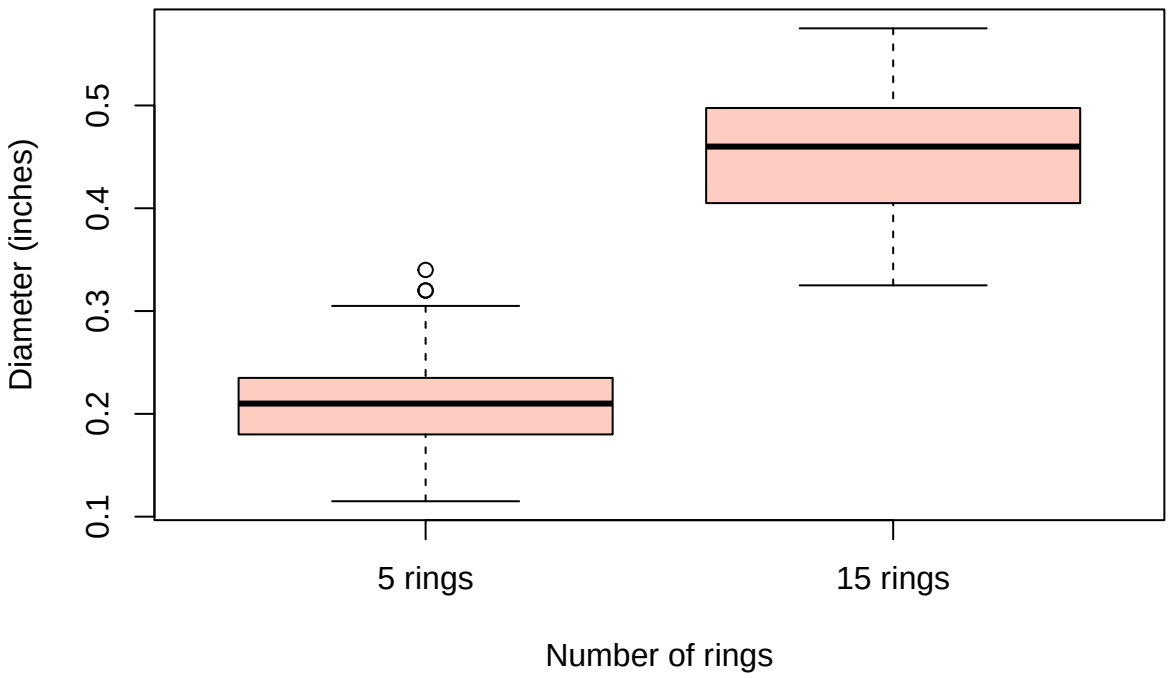
```
##        V1
##  Min.   :-3.7383
##  1st Qu.:-0.6167
##  Median : 0.1742
##  Mean   : 0.0000
##  3rd Qu.: 0.7569
##  Max.   : 2.4217
##  NA's   :7
```

Diameter of 5 and 15 ring midias

```
plot(df$Rings,df$Diameter)
```




```
rings515 <- list(df[df$Rings==5,]$Diameter,df[df$Rings==15,]$Diameter)
names(rings515) <- c("5 rings", "15 rings")
boxplot(rings515, col = "#ffccc2", ylab = "Diameter (inches)", xlab = "Number of rings")
```




```
t.test(df[df$Rings==15,]$Diameter, df[df$Rings==15,]$Diameter)
```

```
##
##  Welch Two Sample t-test
##
## data:  df[df$Rings == 15, ]$Diameter and df[df$Rings == 15, ]$Diameter
## t = 0, df = 204, p-value = 1
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01621616  0.01621616
## sample estimates:
## mean of x mean of y
## 0.4562621 0.4562621
```

Conclusion:

Midias with 15 rings are significantly bigger then midias with 5 rings. Its important to understand that after some age the diameter doest change the same (linear) way.

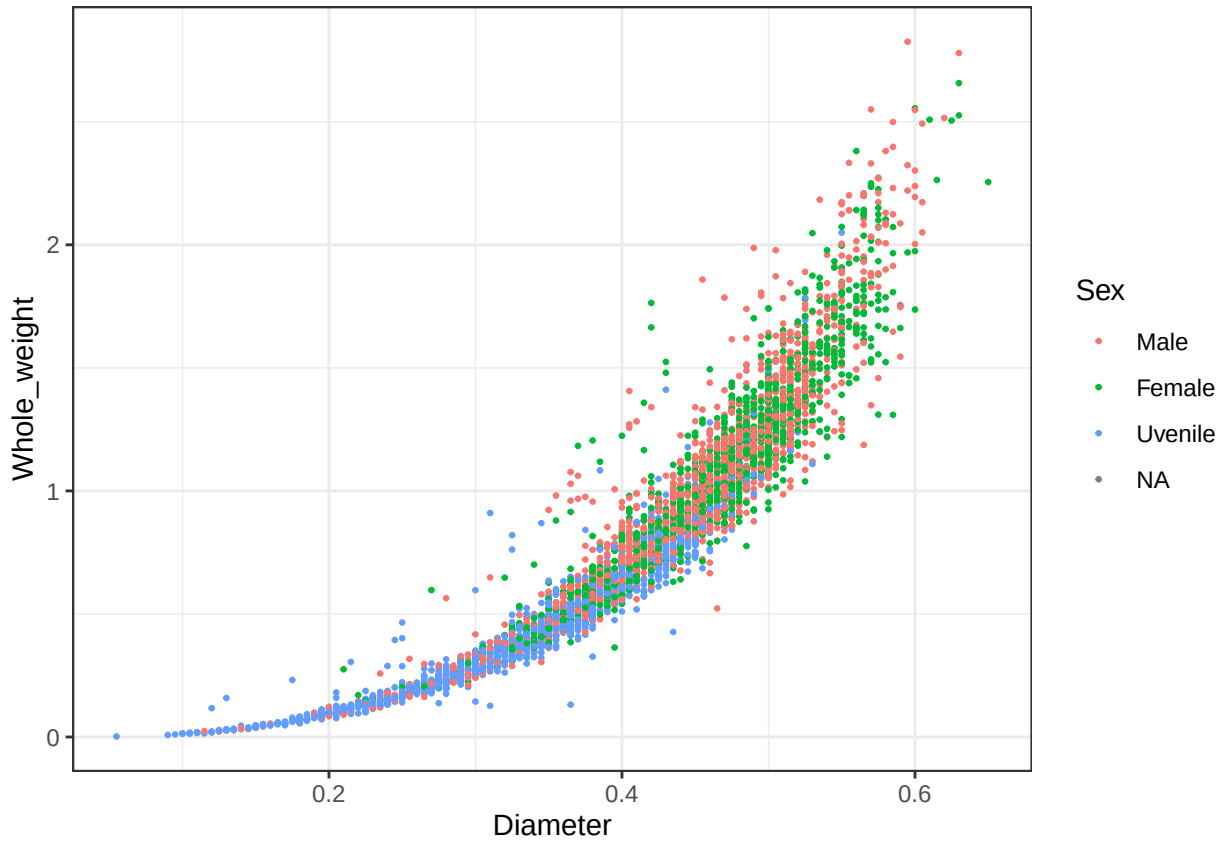
Diameter and Whole weight

Accordind to pearson test, Diameter and weight are correlated:

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

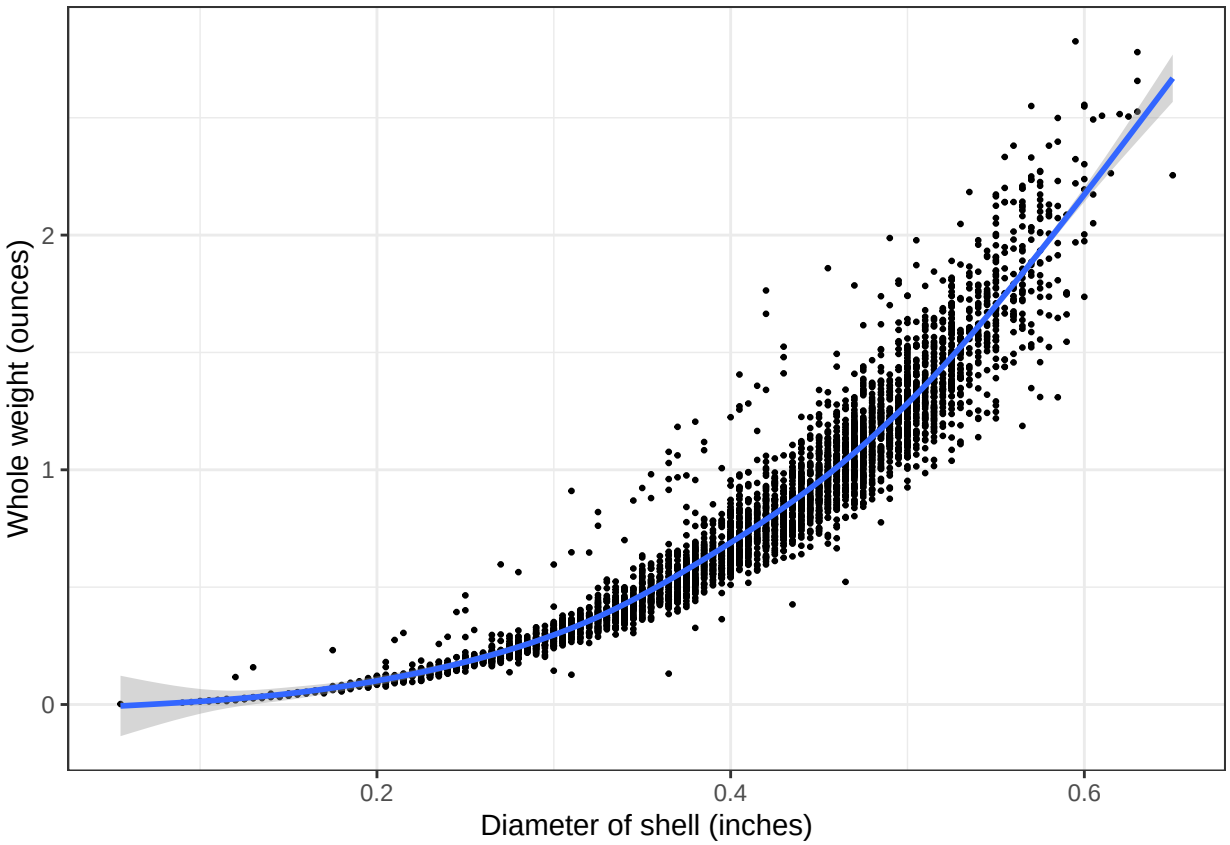
```
##
##  Pearson's product-moment correlation
##
## data:  df$Diameter and df$Whole_weight
## t = 157.86, df = 4169, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9210936 0.9297999
## sample estimates:
##      cor
## 0.925569
```

```
ggplot(df, aes(Diameter, Whole_weight, color = Sex))+
  geom_point(size = 0.5)+
  scale_fill_discrete(labels = c("Male", "Female","Uvenile"))
```




```
ggplot(df, aes(Diameter, Whole_weight))+
  geom_point(size = 0.5)+
  geom_smooth()+
  xlab("Diameter of shell (inches)") + ylab("Whole weight (ounces)")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



But the relationship isn't linear, we cant make a simple liner model. Most probably it's cubic, considering the wariables' phisical meaning.

Hypothesis cheching:

1. Age doesnt really influence the size and the weight of midia afters some point. While the uvenile age it grows gradually, but then it can stop growing, for example, because of lack of food or high competition or smth else... **CONFIRMED** (dotplot in 5 vs 15 ring midias)
2. Length and Diameters are highly correlated, it might be good to include only one factor in some model. **CONFIRMED** (cor.test)
3. Midias grow evenly: if the shell weights more, it means the visceral part weights more as well. **CONFIRMED** (0.91 correlation, cor.test)
4. Sex (except of uvenile stage) doesnt mean much (we need some more tests about this). **CONFUSION** (Sex and Length section, sex and ring chi.test)
5. Midia stops growing in diameter after some point of gained weight. The only explanation of this - it grows in height. Also maybe the weight grows and sizes don't because of some parasites on the shell of old midias or pearls inside? :D **FALSE** (analises below)

```
df_z <- as.data.frame(sapply(df[,3:9], scale))
mod_1 <- lm(Whole_weight ~ Height*Diameter*Length, df_z)
summary(mod_1)
```

```
##
## Call:
## lm(formula = Whole_weight ~ Height * Diameter * Length, data = df_z)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.87552 -0.10977 -0.01530  0.08388  1.61797
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -0.226175   0.004634 -48.811  < 2e-16 ***
## Height                   0.128659   0.006603  19.484  < 2e-16 ***
## Diameter                 0.390855   0.022788  17.152  < 2e-16 ***
## Length                   0.525696   0.022886  22.970  < 2e-16 ***
## Height:Diameter         -0.102670   0.026240  -3.913 9.27e-05 ***
## Height:Length            0.216960   0.026589   8.160 4.40e-16 ***
## Diameter:Length          0.147648   0.007862  18.781  < 2e-16 ***
## Height:Diameter:Length   0.024636   0.002332  10.566  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2264 on 4154 degrees of freedom
##   (15 observations deleted due to missingness)
## Multiple R-squared:  0.9489, Adjusted R-squared:  0.9488
## F-statistic: 1.101e+04 on 7 and 4154 DF,  p-value: < 2.2e-16
```

With this model we can predict Whole weight with height, diameter and length with a pretty high precision. Seems like theres no parasites or pearls((

6. It might be possible to make a linear model that will predict if the midia is uvenile by it's size and(or) weight. Or it might even predict age (ring amount), but only below 10 rings, if midia is older, the model will make mistakes.

Notion

It is confusing a bit, that the Sex variable kinda includes an age factor too, so mb it would be better to perform some of the previous tests without uvenile midias. Then sex won't correlate with some wariables.