Fish Market Code

Yufei Yin

Load Package

```
library(tidyverse)
library(leaps)
library(caret)
library(car)
```

Load Dataset

```
fish <- read_csv("fish.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Species = col_character(),
##   Weight = col_double(),
##   Length1 = col_double(),
##   Length2 = col_double(),
##   Length3 = col_double(),
##   Height = col_double(),
##   Width = col_double()
## )
```

```
head(fish)
```

```
## # A tibble: 6 x 7
##   Species Weight Length1 Length2 Length3 Height Width
##   <chr>    <dbl>   <dbl>   <dbl>   <dbl>  <dbl> <dbl>
## 1 Bream      242    23.2    25.4    30     11.5  4.02
## 2 Bream      290    24      26.3    31.2   12.5  4.31
```

```
## 3 Bream      340    23.9    26.5    31.1   12.4  4.70
## 4 Bream      363    26.3    29      33.5   12.7  4.46
## 5 Bream      430    26.5    29      34     12.4  5.13
## 6 Bream      450    26.8    29.7    34.7   13.6  4.93
```

Check Missing Value

```
any(is.na(fish))
```

```
## [1] FALSE
```

Extra point (outlier)

```
extrapoint <- tibble(Species = "Bream", Length1 = 20, Length2 = 10, Length3 = 30, Height = 15, Width = 5
fish <- bind_rows(fish, extrapoint)
```

Data Cleaning

```
fish <- fish %>%
  filter(Weight != 0) %>%                          # Observation 41 has weight 0
  mutate(Observations = row_number()) %>%
  mutate(Species = factor(Species)) %>%
  rename(Vlength = Length1) %>%                    # Vertical Length
  rename(Dlength = Length2) %>%                    # Diagonal Length
  rename(Clength = Length3) %>%                    # Cross Length
  select(Observations, Species, Vlength, Dlength, Clength, Height, Width, Weight)
```

Observation 41 has weight 0 which is not reasonable, so we drop this observation.

Change variable Species to factor

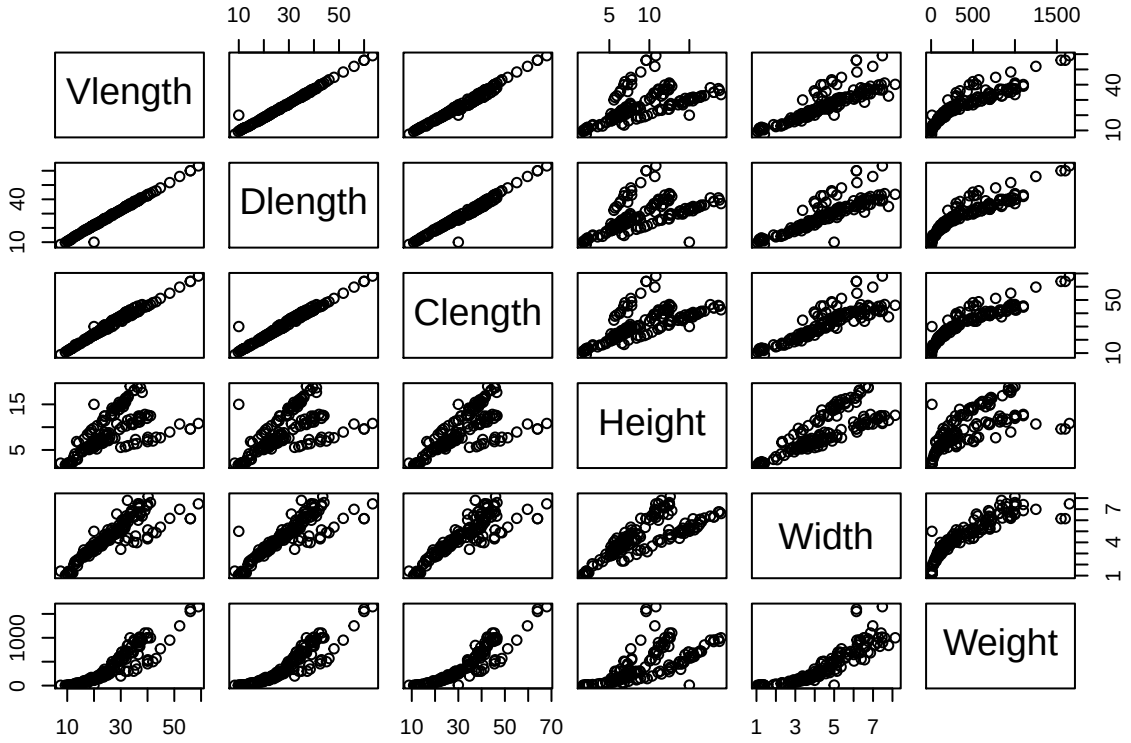
Create a new variable Observations contain the observation number.

Length1, Length2, Length3 stand for Vertical Length, Diagonal Length, and Cross Length, So we change the variable name to Vlength, Dlength, and Clength which makes more sense.

Finally we sort our data.

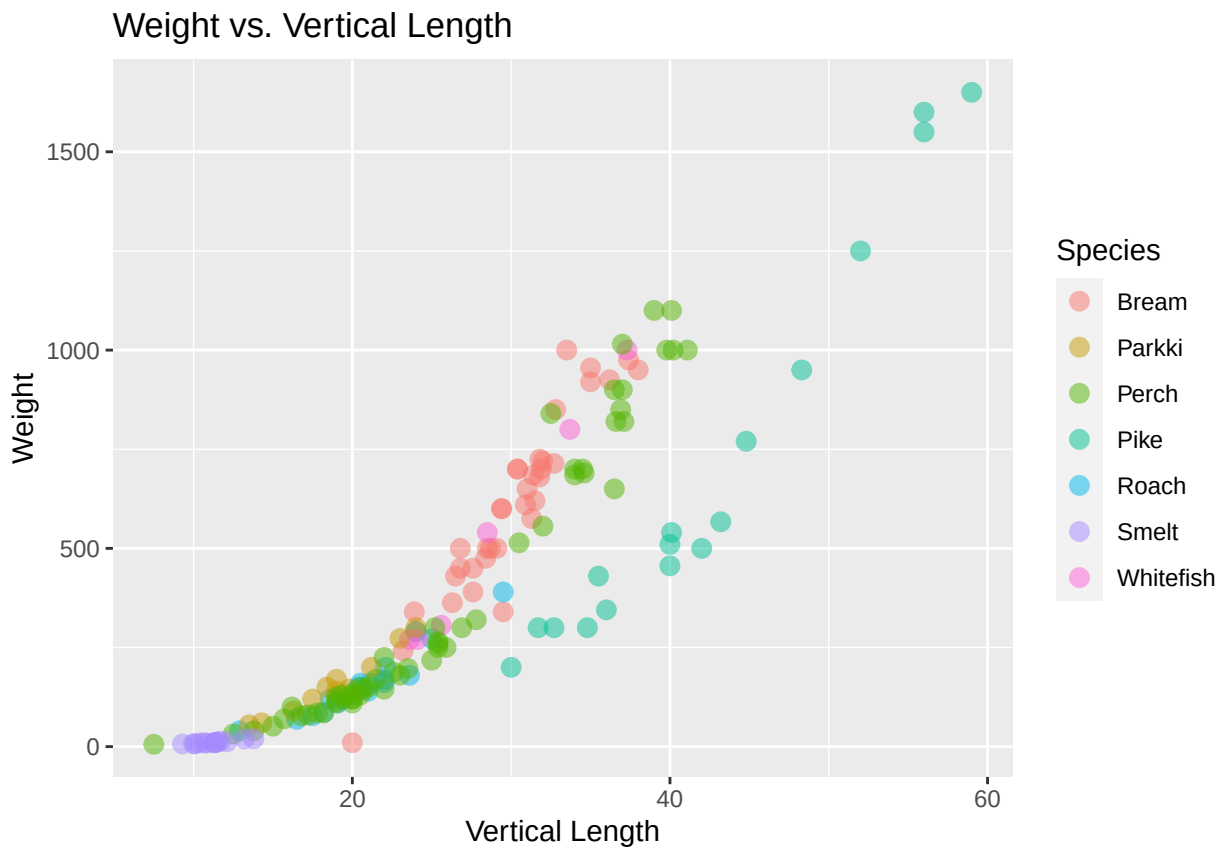
Pair Plot

```
pairs(fish[3:8])
```



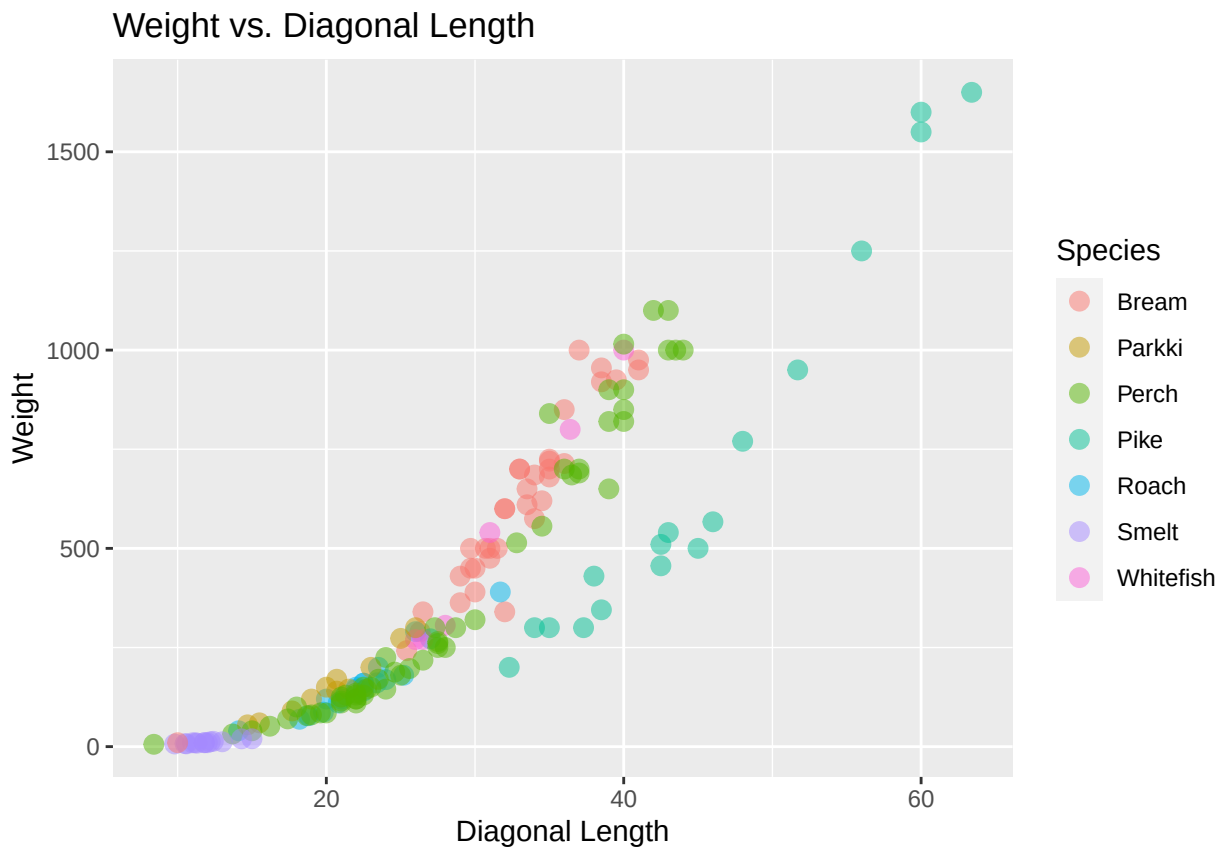
Weight vs. Vertical Length

```
ggplot(data = fish, aes(x = Vlength, y = Weight, color = Species)) +
  geom_point(size = 3, alpha = 0.5) +
  labs(x = "Vertical Length", title = "Weight vs. Vertical Length")
```



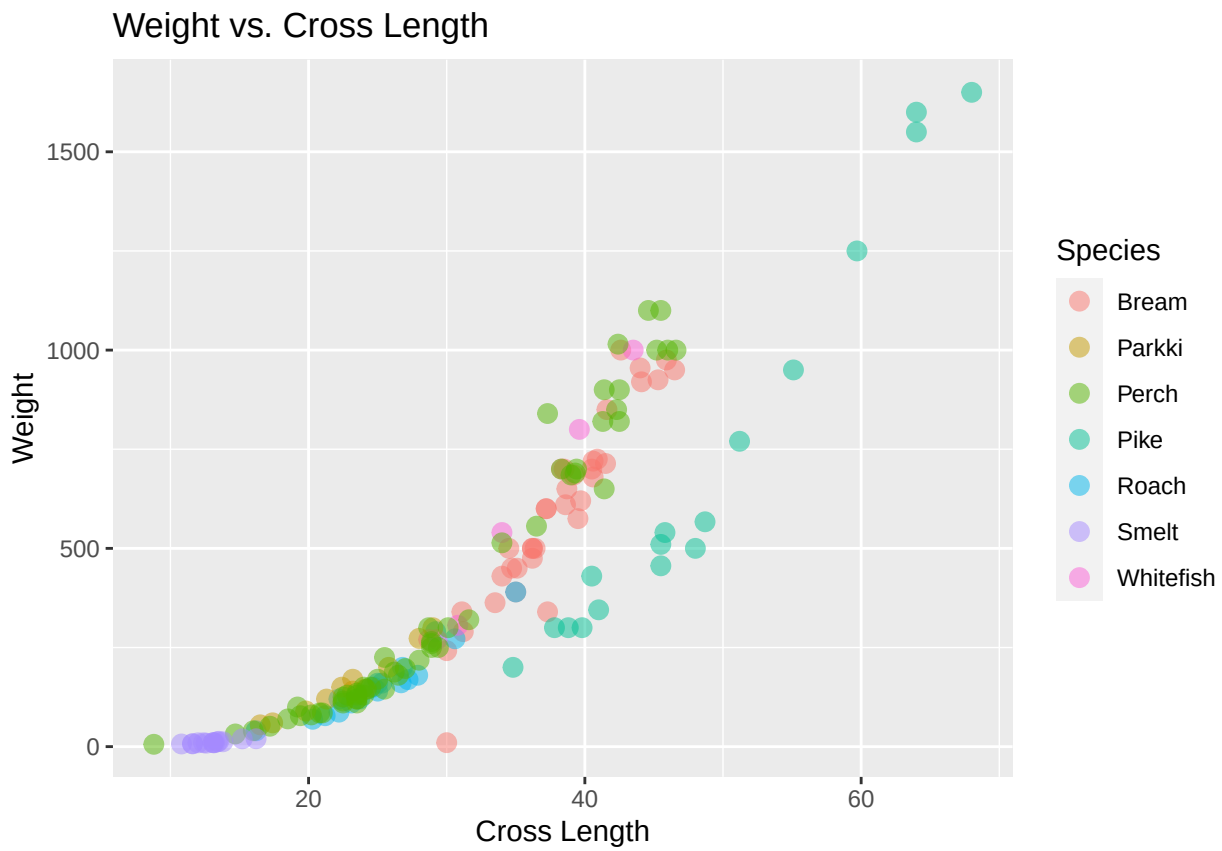
Weight vs. Diagonal Length

```
ggplot(data = fish, aes(x = Dlength, y = Weight, color = Species)) +
  geom_point(size = 3, alpha = 0.5) +
  labs(x = "Diagonal Length", title = "Weight vs. Diagonal Length")
```



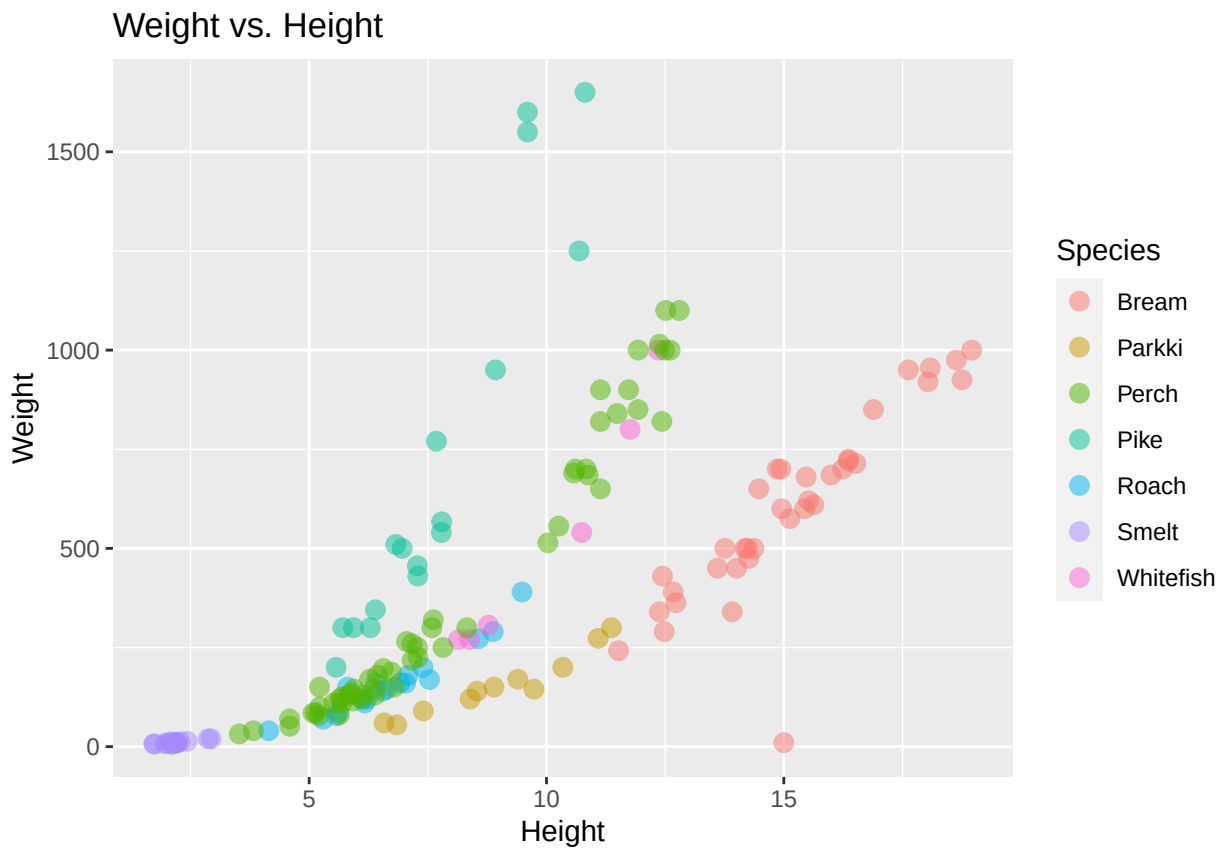
Weight vs. Cross Length

```
ggplot(data = fish, aes(x = Clength, y = Weight, color = Species)) +
  geom_point(size = 3, alpha = 0.5) +
  labs(x = "Cross Length", title = "Weight vs. Cross Length")
```



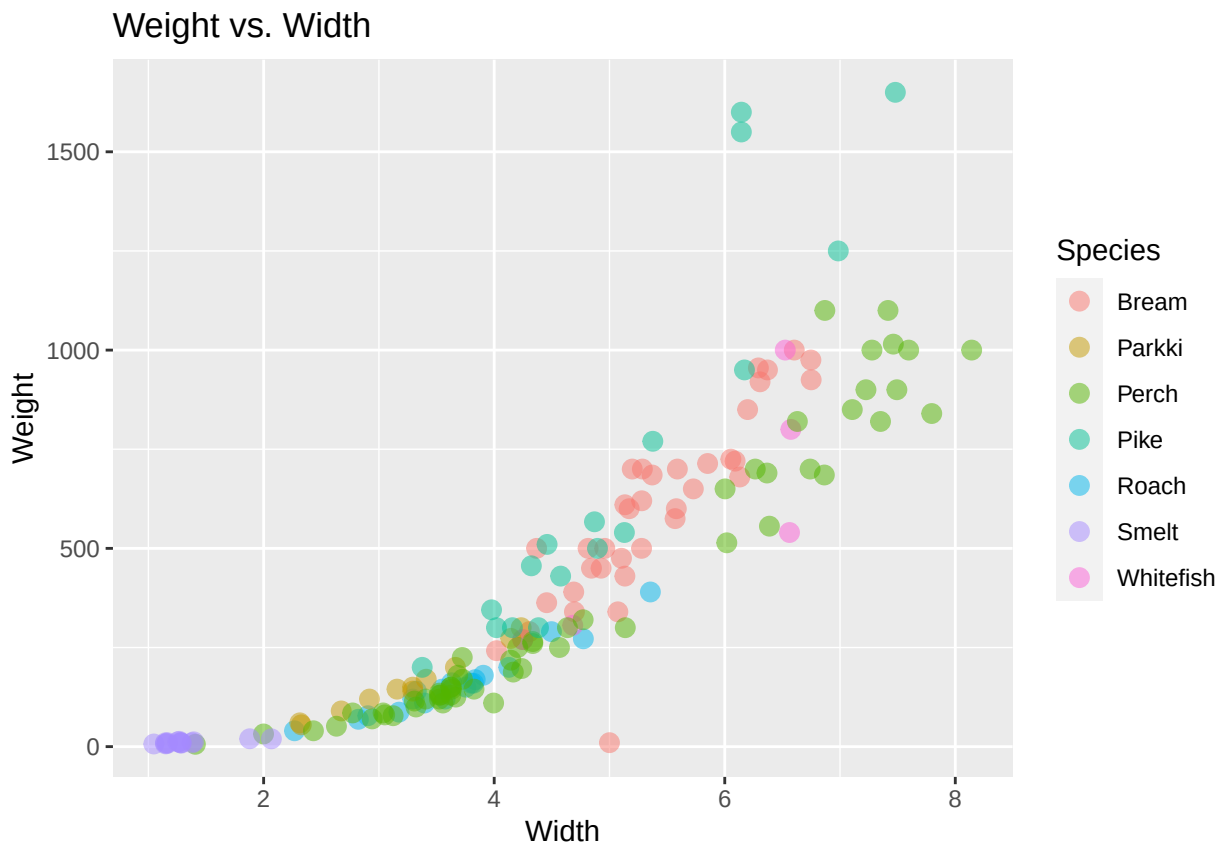
Weight vs. Height

```
ggplot(data = fish, aes(x = Height, y = Weight, color = Species)) +
  geom_point(size = 3, alpha = 0.5) +
  labs(title = "Weight vs. Height")
```



Weight vs. Width

```
ggplot(data = fish, aes(x = Width, y = Weight, color = Species)) +
  geom_point(size = 3, alpha = 0.5) +
  labs(title = "Weight vs. Width")
```

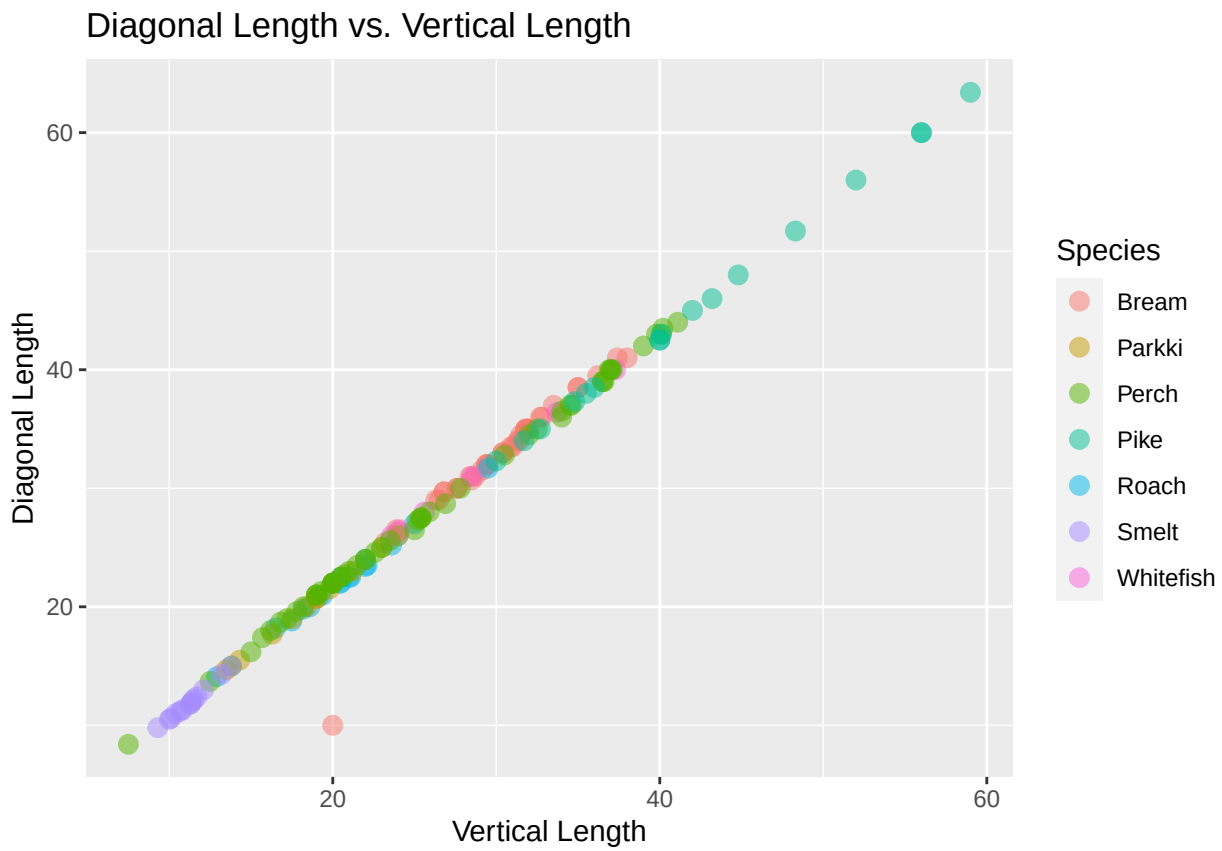


Multicollinearity

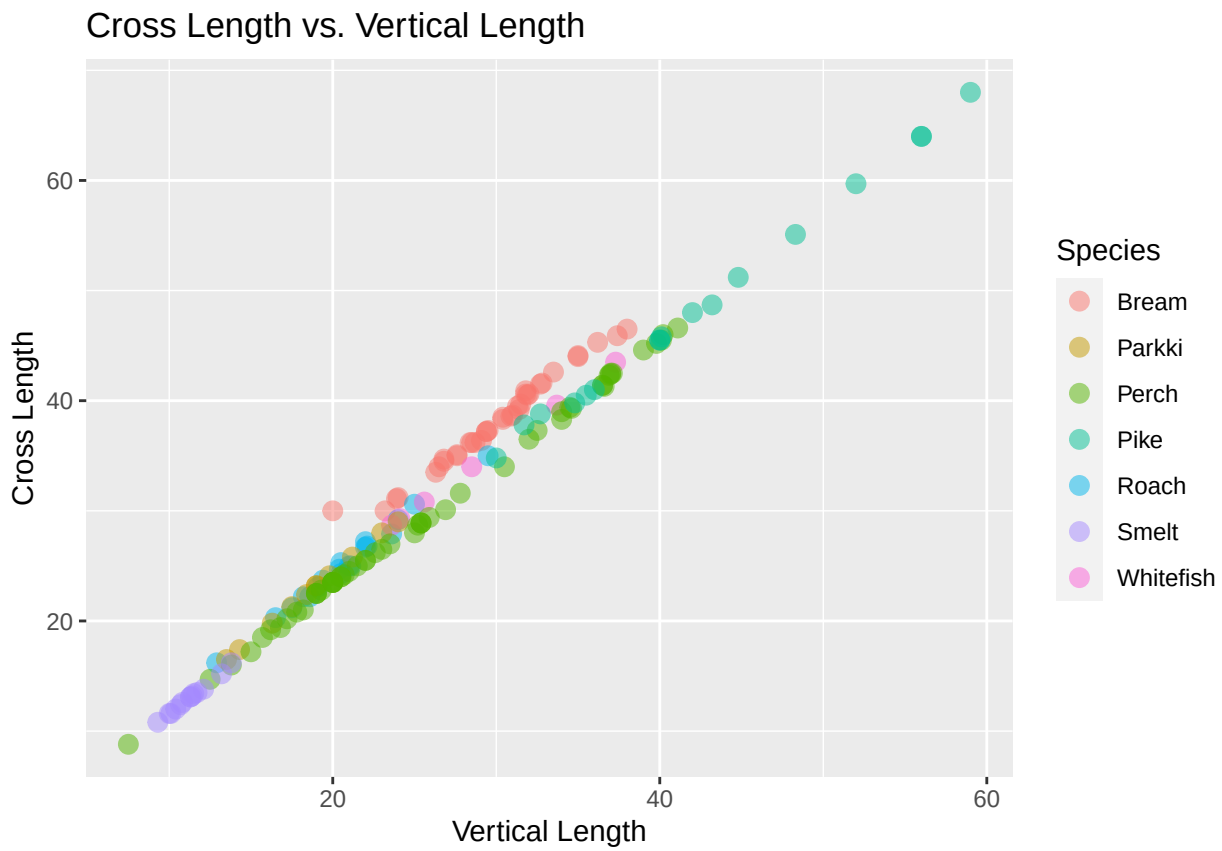
```
cor(fish[3:8])
```

```
##            Vlength   Dlength   Clength    Height     Width    Weight
## Vlength 1.0000000 0.9958187 0.9911656 0.6142680 0.8639304 0.9154659
## Dlength 0.9958187 1.0000000 0.9860478 0.6146870 0.8611377 0.9184415
## Clength 0.9911656 0.9860478 1.0000000 0.6973401 0.8775947 0.9203455
## Height  0.6142680 0.6146870 0.6973401 1.0000000 0.7902423 0.7071193
## Width   0.8639304 0.8611377 0.8775947 0.7902423 1.0000000 0.8806498
## Weight  0.9154659 0.9184415 0.9203455 0.7071193 0.8806498 1.0000000
```

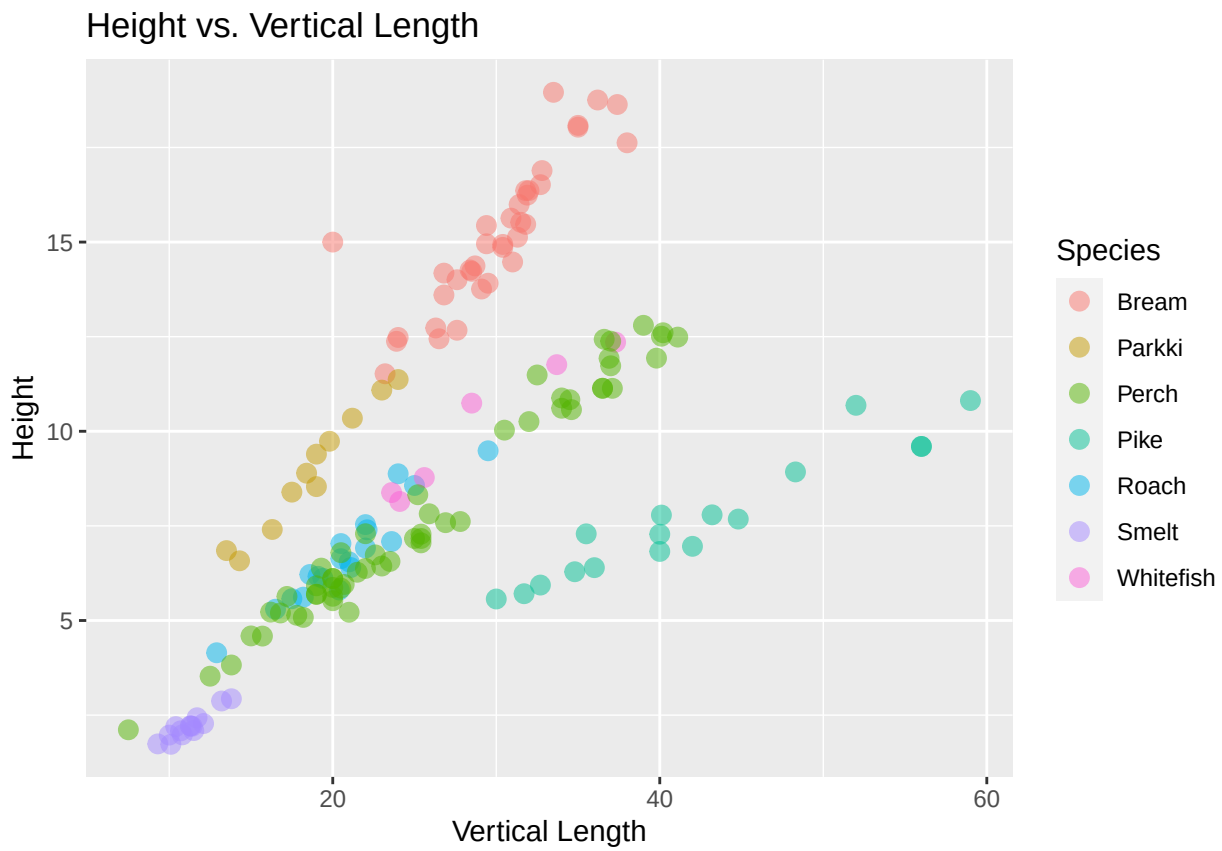
```
ggplot(data = fish, aes(x = Vlength, y = Dlength, color = Species)) +
  geom_point(size = 3, alpha = 0.5) +
  labs(x = "Vertical Length", y = "Diagonal Length", title = "Diagonal Length vs. Vertical Length")
```



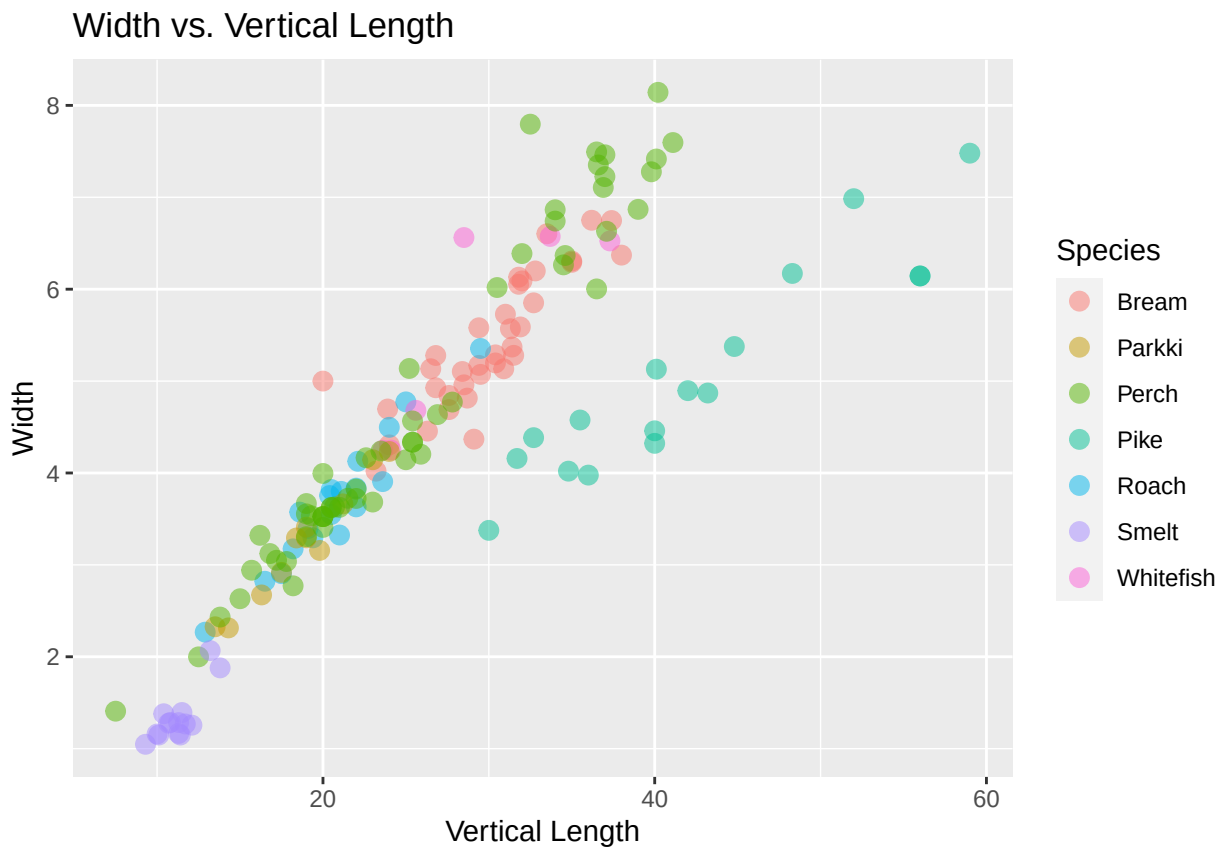

```
ggplot(data = fish, aes(x = Vlength, y = Clength, color = Species)) +
  geom_point(size = 3, alpha = 0.5) +
  labs(x = "Vertical Length", y = "Cross Length", title = "Cross Length vs. Vertical Length")
```



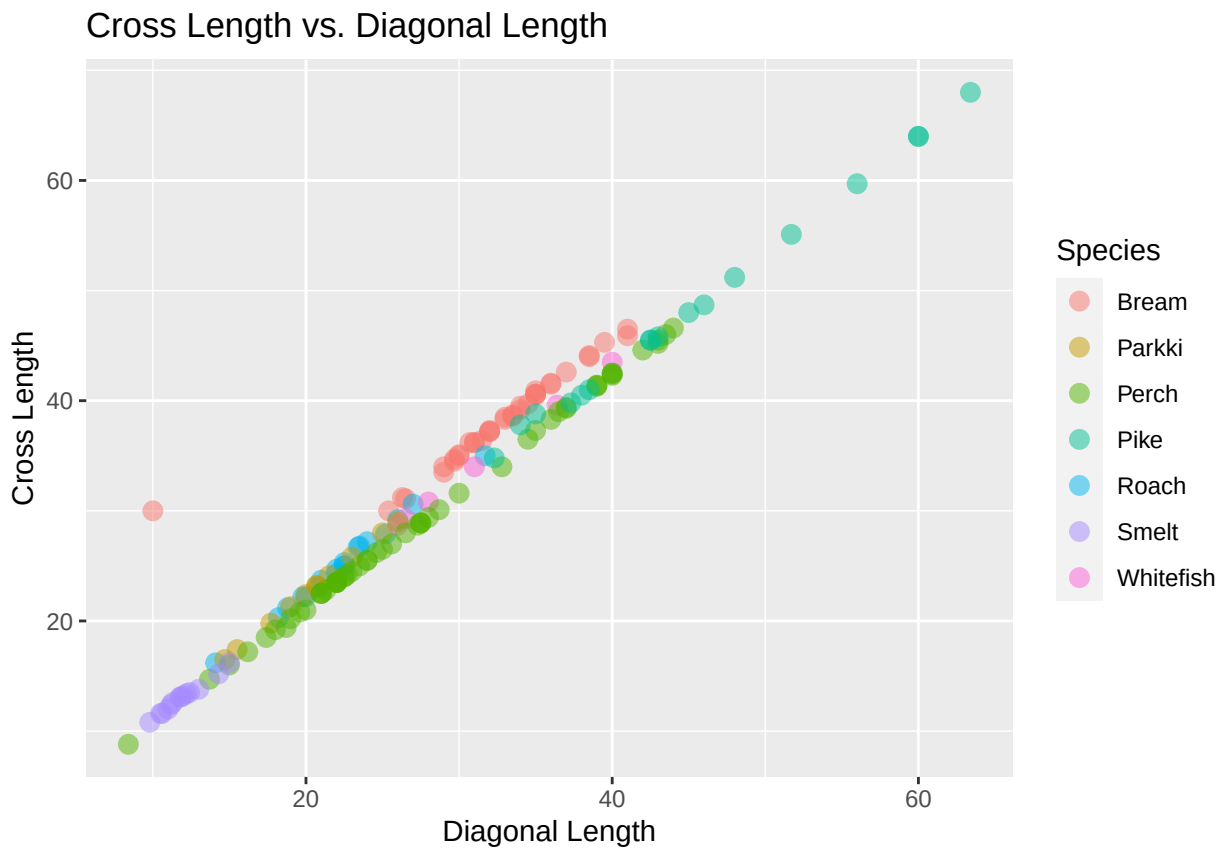

```
ggplot(data = fish, aes(x = Vlength, y = Height, color = Species)) +
  geom_point(size = 3, alpha = 0.5) +
  labs(x = "Vertical Length", title = "Height vs. Vertical Length")
```



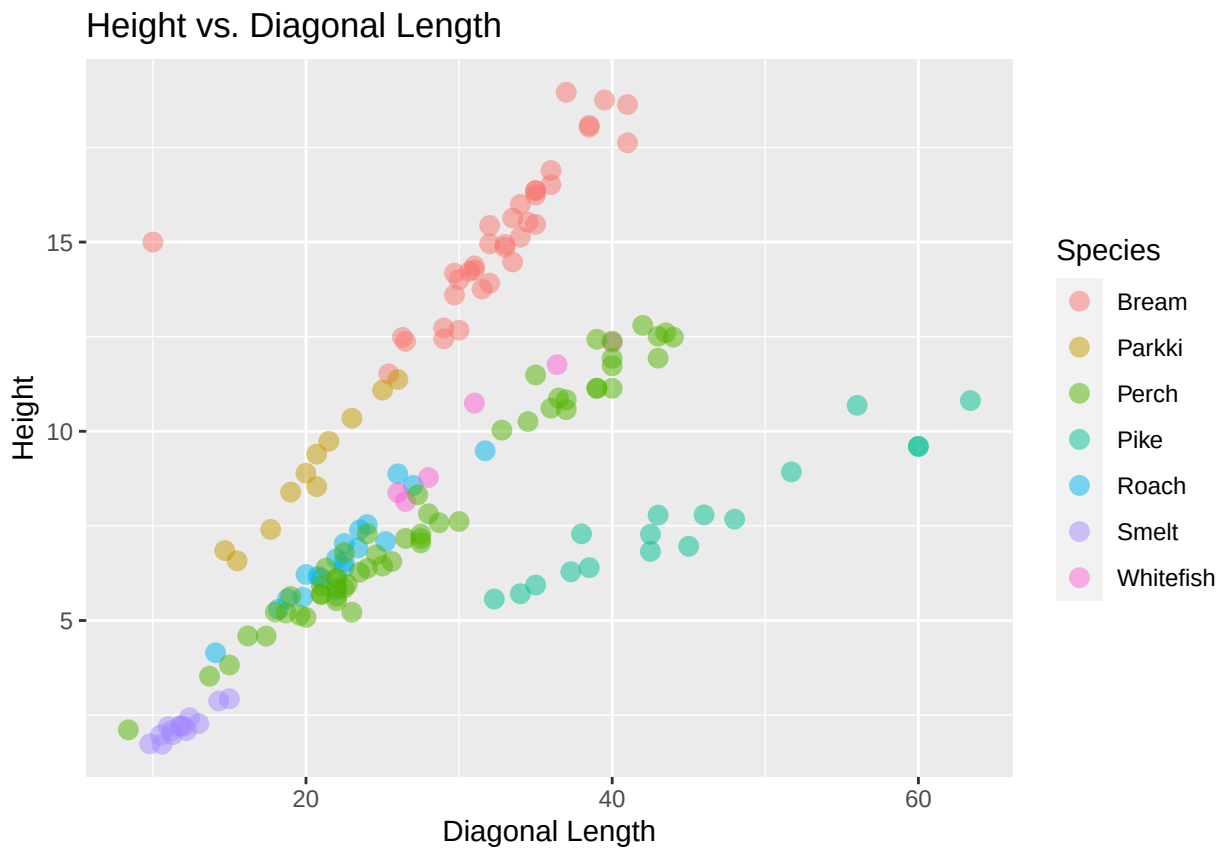

```
ggplot(data = fish, aes(x = Vlength, y = Width, color = Species)) +
  geom_point(size = 3, alpha = 0.5) +
  labs(x = "Vertical Length", title = "Width vs. Vertical Length")
```



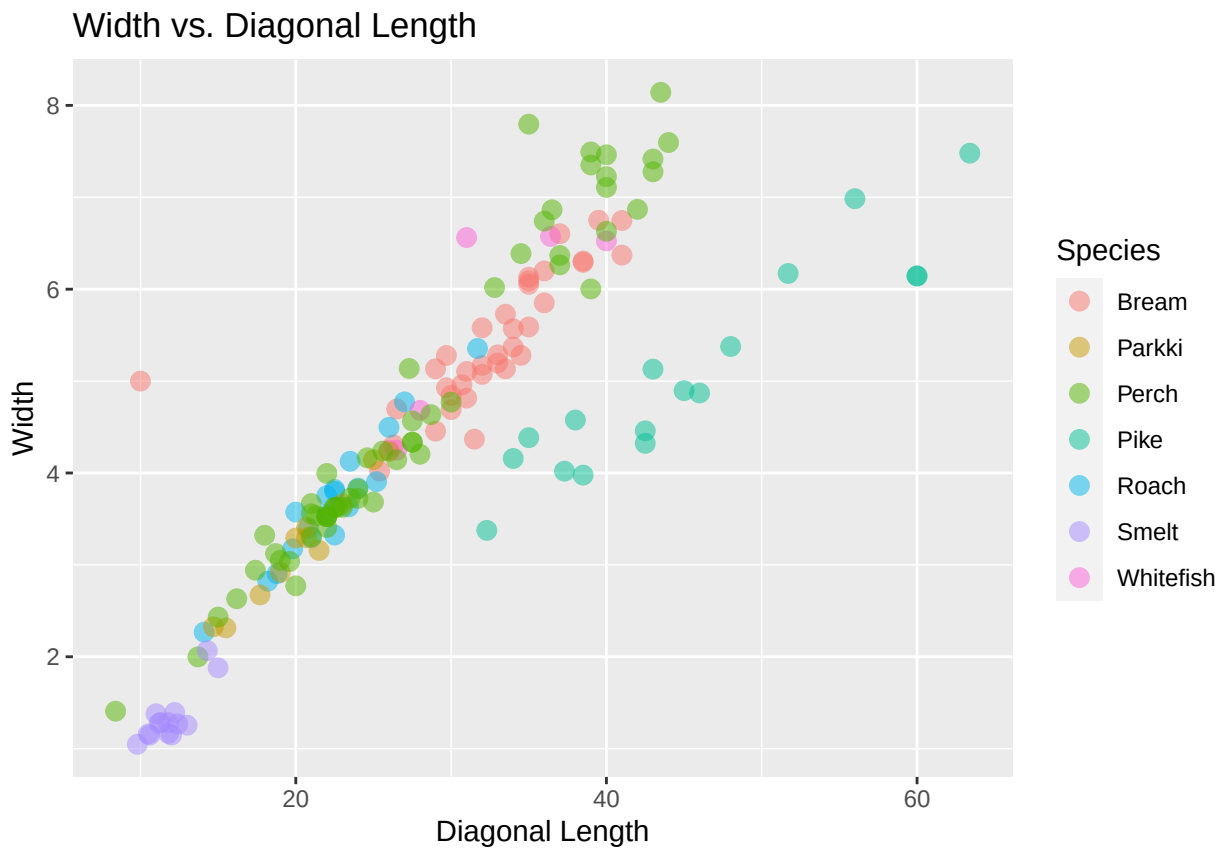

```
ggplot(data = fish, aes(x = Dlength, y = Clength, color = Species)) +
  geom_point(size = 3, alpha = 0.5) +
  labs(x = "Diagonal Length", y = "Cross Length", title = "Cross Length vs. Diagonal Length")
```




```
ggplot(data = fish, aes(x = Dlength, y = Height, color = Species)) +
  geom_point(size = 3, alpha = 0.5) +
  labs(x = "Diagonal Length", title = "Height vs. Diagonal Length")
```

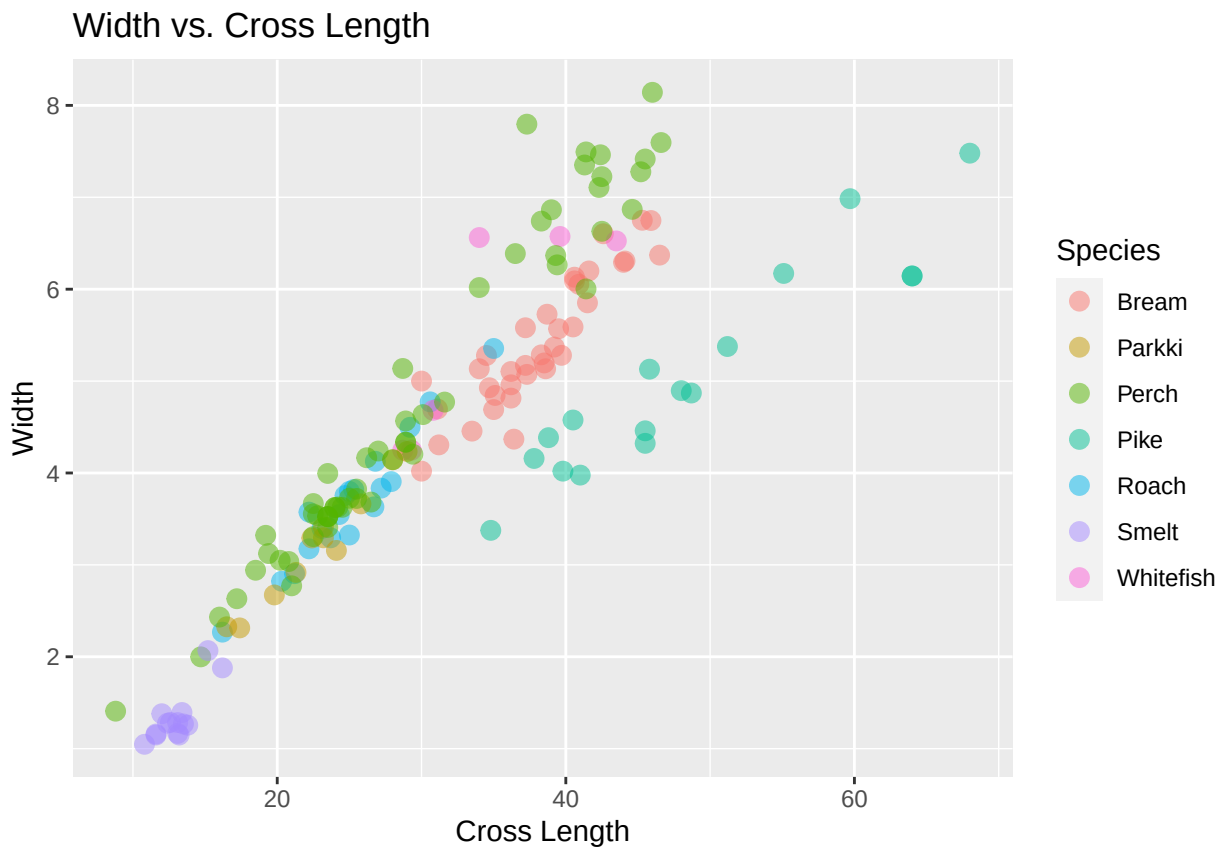



```
ggplot(data = fish, aes(x = Dlength, y = Width, color = Species)) +
  geom_point(size = 3, alpha = 0.5) +
  labs(x = "Diagonal Length", title = "Width vs. Diagonal Length")
```

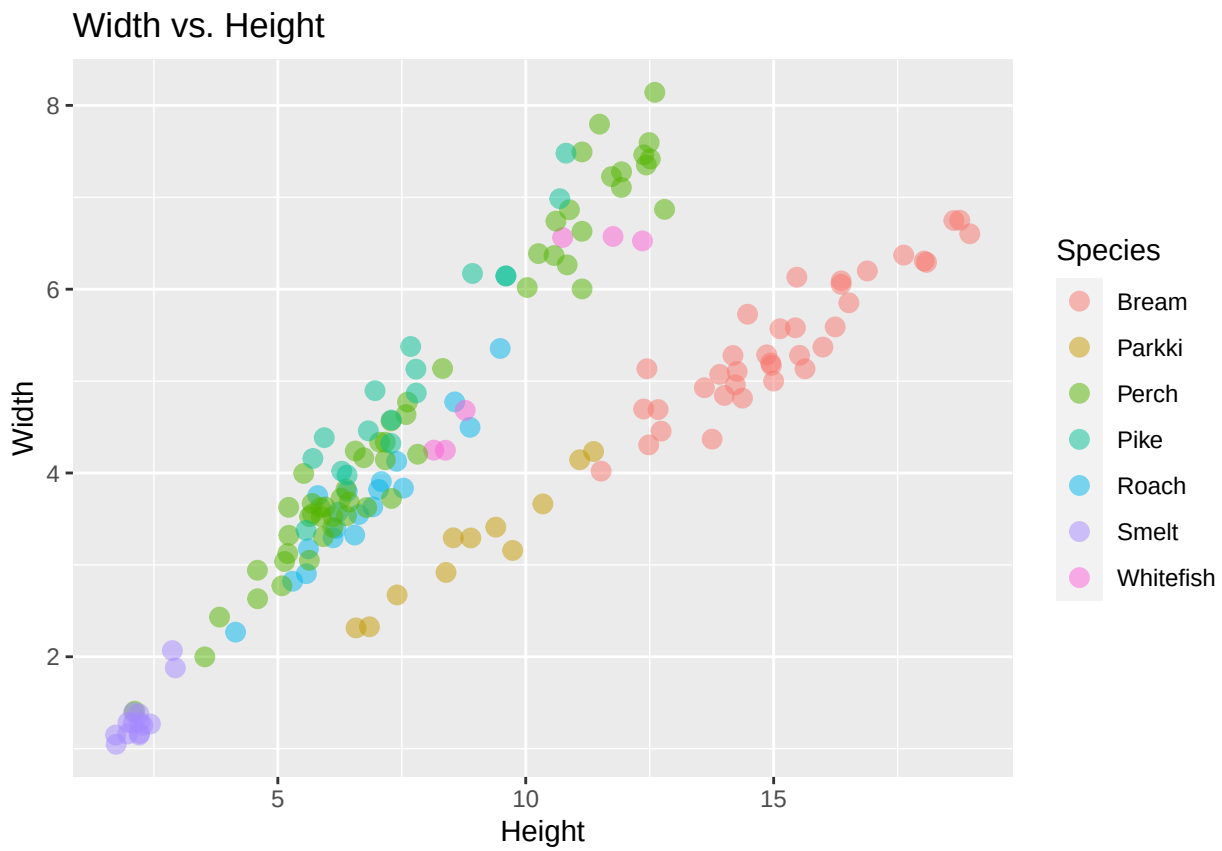



```
ggplot(data = fish, aes(x = Clength, y = Height, color = Species)) +
  geom_point(size = 3, alpha = 0.5) +
  labs(x = "Cross Length", title = "Height vs. Cross Length")
```


```
ggplot(data = fish, aes(x = Clength, y = Width, color = Species)) +
  geom_point(size = 3, alpha = 0.5) +
  labs(x = "Cross Length", title = "Width vs. Cross Length")
```




```
ggplot(data = fish, aes(x = Height, y = Width, color = Species)) +
  geom_point(size = 3, alpha = 0.5) +
  labs(title = "Width vs. Height")
```



Full Model

```
full.mdl <- lm(data = fish,
               Weight ~ Species*Vlength + Species*Dlength + Species*Clength + Species*Height + Species*
summary(full.mdl)
```

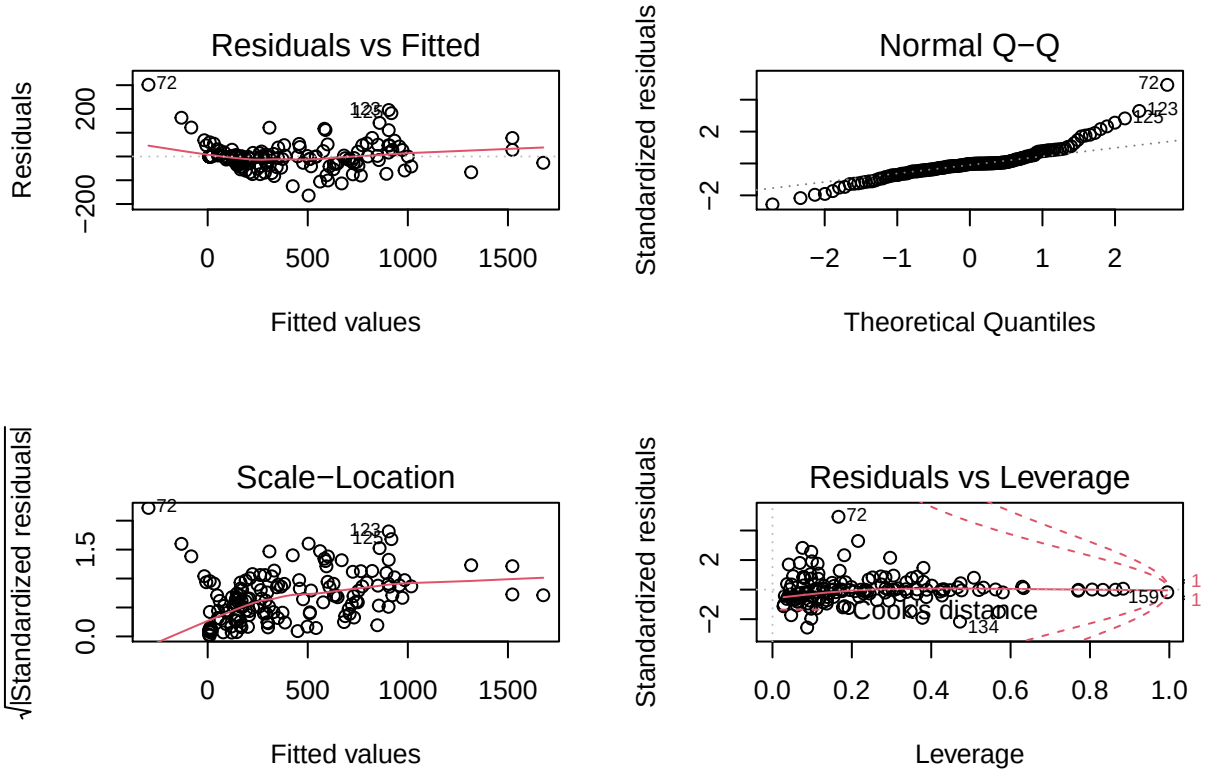
```
##
## Call:
## lm(formula = Weight ~ Species * Vlength + Species * Dlength +
##     Species * Clength + Species * Height + Species * Width, data = fish)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -164.411  -27.909   -0.864   14.585  301.659
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)               -925.610    186.650  -4.959 2.42e-06 ***
## SpeciesParkki              729.828    251.450   2.902 0.004426 **
## SpeciesPerch               369.024    192.399   1.918 0.057547 .
## SpeciesPike               -779.220    214.224  -3.637 0.000411 ***
## SpeciesRoach               655.697    223.972   2.928 0.004106 **
## SpeciesSmelt               912.421    360.588   2.530 0.012723 *
## SpeciesWhitefish         -1732.734   2720.059  -0.637 0.525356
## Vlength                      9.987     50.653   0.197 0.844034
## Dlength                     29.762     10.428   2.854 0.005107 **
## Clength                    -26.220     44.331  -0.591 0.555358
## Height                      63.226     21.486   2.943 0.003925 **
## Width                       55.463     51.587   1.075 0.284531
## SpeciesParkki:Vlength      215.598    787.390   0.274 0.784711
## SpeciesPerch:Vlength       -13.118     67.451  -0.194 0.846140
## SpeciesPike:Vlength       -406.206    138.237  -2.938 0.003974 **
## SpeciesRoach:Vlength        -4.802     99.668  -0.048 0.961653
## SpeciesSmelt:Vlength       -14.791    369.619  -0.040 0.968148
## SpeciesWhitefish:Vlength  -411.084   1116.179  -0.368 0.713318
## SpeciesParkki:Dlength     -101.626   1334.951  -0.076 0.939448
## SpeciesPerch:Dlength       -68.264     68.938  -0.990 0.324105
## SpeciesPike:Dlength        181.376    136.357   1.330 0.186056
## SpeciesRoach:Dlength       -27.148     99.551  -0.273 0.785560
## SpeciesSmelt:Dlength       -25.272    208.596  -0.121 0.903777
## SpeciesWhitefish:Dlength  -186.394   1020.001  -0.183 0.855319
## SpeciesParkki:Clength     -103.539    710.180  -0.146 0.884336
## SpeciesPerch:Clength        69.137     64.058   1.079 0.282676
## SpeciesPike:Clength        229.691     63.995   3.589 0.000486 ***
## SpeciesRoach:Clength        20.872     75.489   0.276 0.782656
## SpeciesSmelt:Clength        27.163    200.825   0.135 0.892640
## SpeciesWhitefish:Clength   563.034    846.415   0.665 0.507232
## SpeciesParkki:Height       -50.673     80.607  -0.629 0.530812
## SpeciesPerch:Height          2.430     31.537   0.077 0.938721
## SpeciesPike:Height          84.076     61.048   1.377 0.171073
## SpeciesRoach:Height        -40.017     55.318  -0.723 0.470885
## SpeciesSmelt:Height        -59.650    177.559  -0.336 0.737514
## SpeciesWhitefish:Height    120.277    299.928   0.401 0.689138
```

```
## SpeciesParkki:Width          78.953    223.058   0.354 0.724007
## SpeciesPerch:Width            9.439     58.836   0.160 0.872815
## SpeciesPike:Width          -330.409     94.098  -3.511 0.000634 ***
## SpeciesRoach:Width           10.539    105.088   0.100 0.920292
## SpeciesSmelt:Width          -51.944    159.908  -0.325 0.745886
## SpeciesWhitefish:Width     -167.937    222.861  -0.754 0.452633
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 67.03 on 117 degrees of freedom
## Multiple R-squared:  0.974,  Adjusted R-squared:  0.9649
## F-statistic:   107 on 41 and 117 DF,  p-value: < 2.2e-16
```

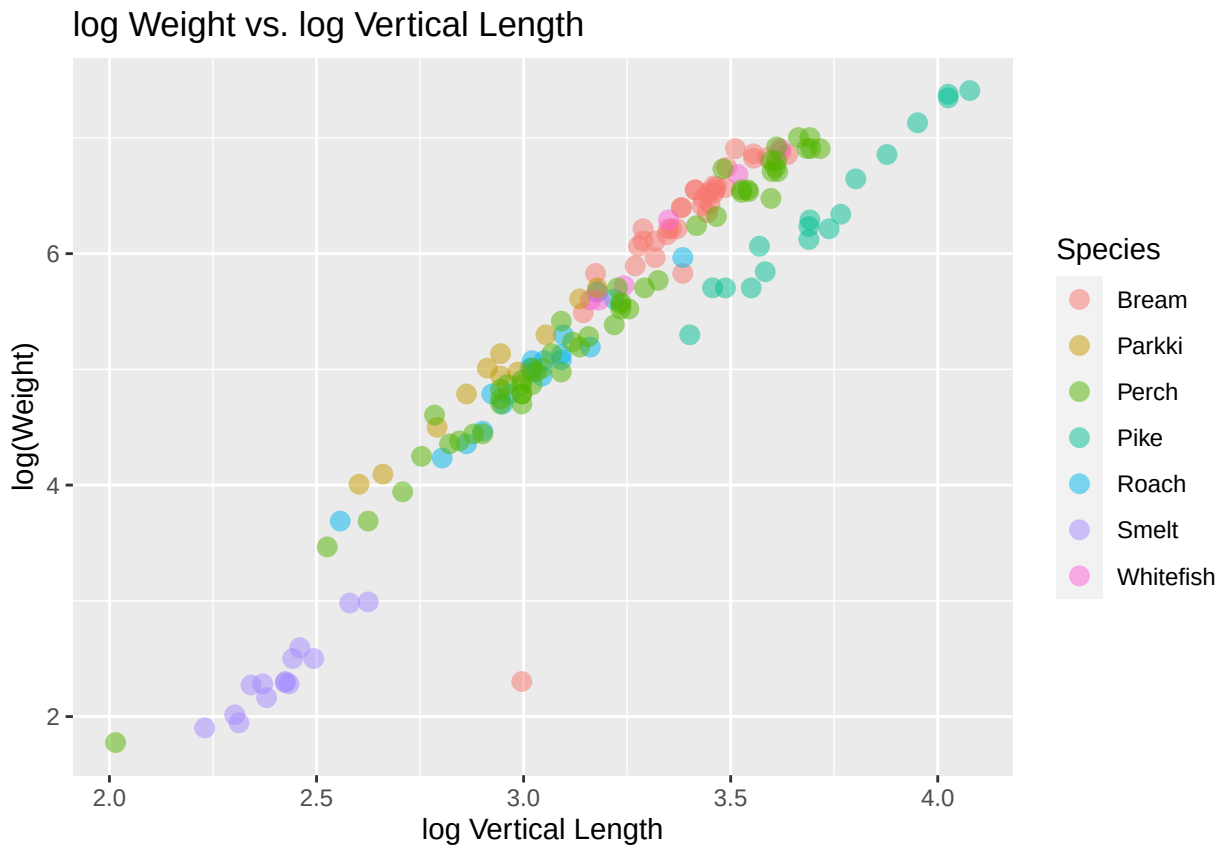
Residual Analysis

```
par(mfrow = c(2,2))
plot(full.mdl)
```

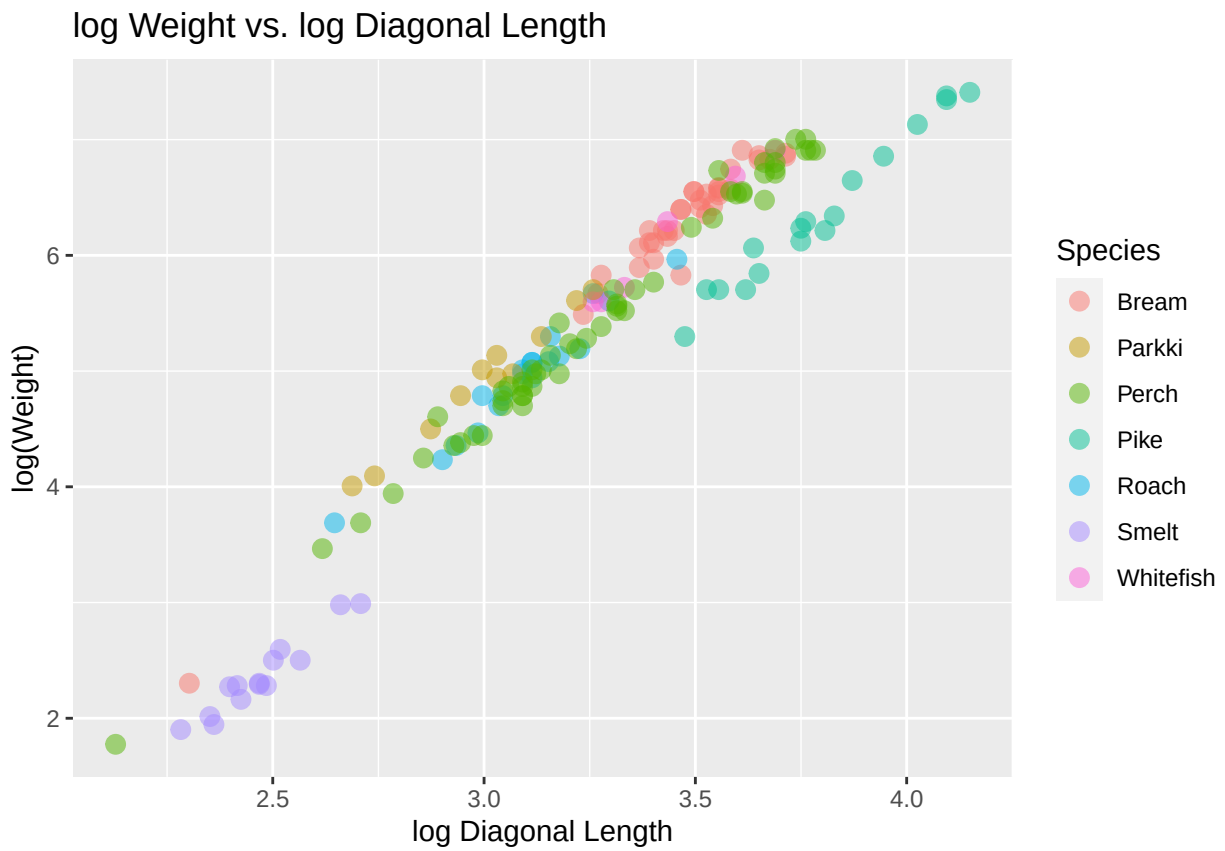


Transformation

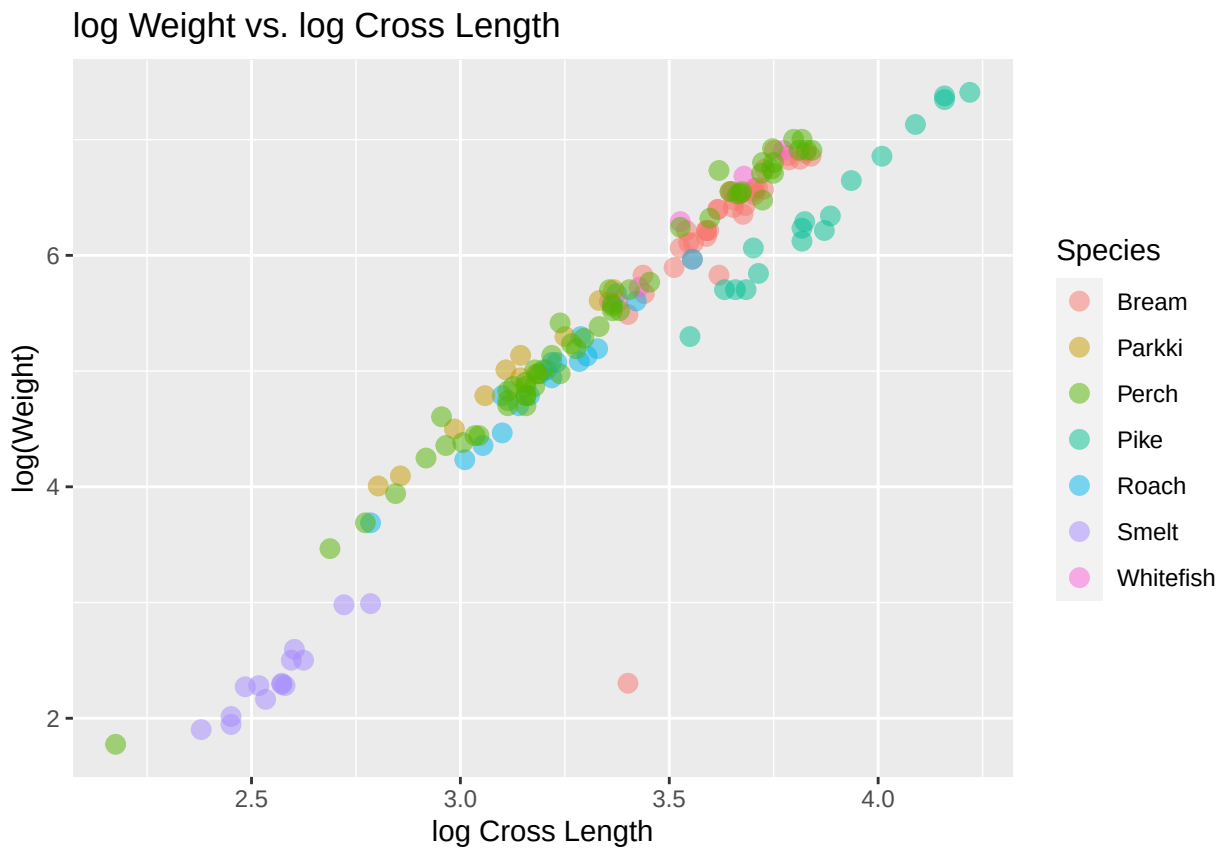
```
ggplot(data = fish, aes(x = log(Vlength), y = log(Weight), color = Species)) +
  geom_point(size = 3, alpha = 0.5) +
  labs(x = "log Vertical Length", title = "log Weight vs. log Vertical Length")
```



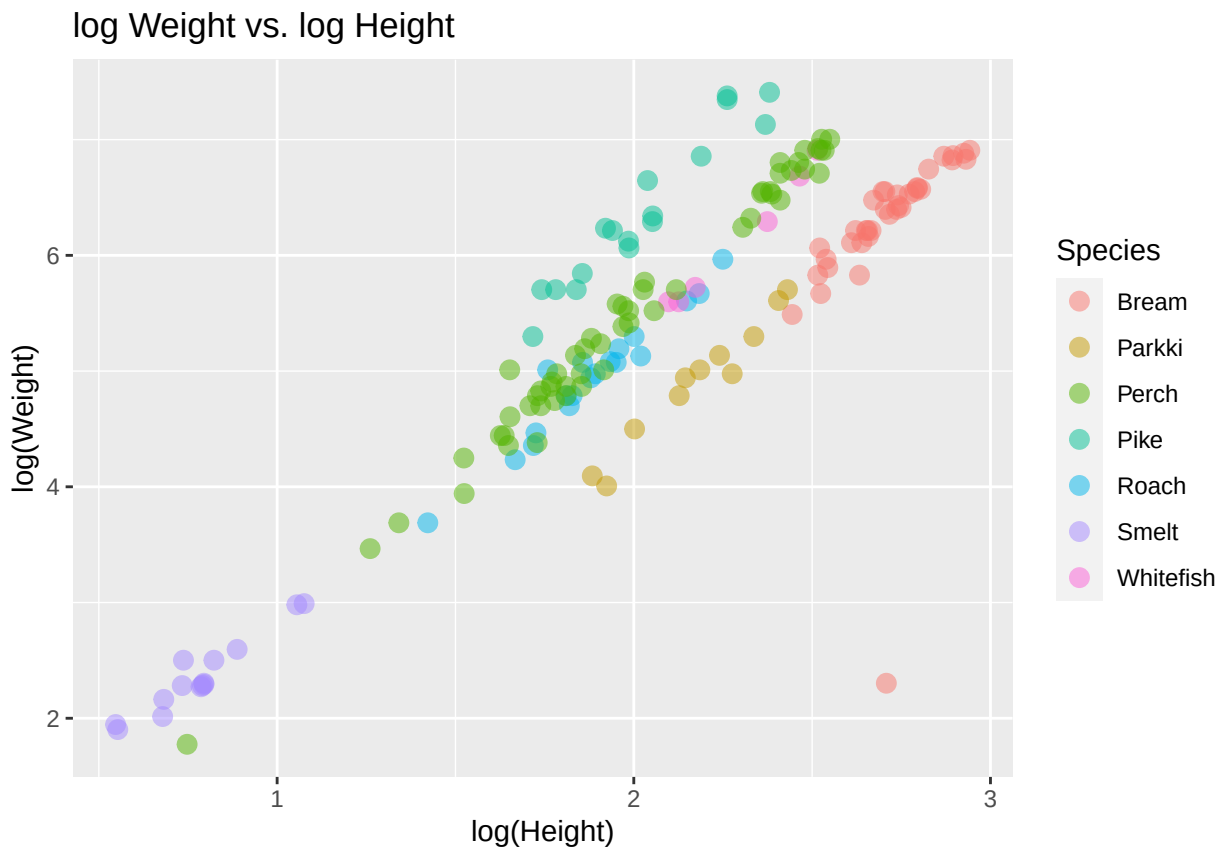

```
ggplot(data = fish, aes(x = log(Dlength), y = log(Weight), color = Species)) +
  geom_point(size = 3, alpha = 0.5) +
  labs(x = "log Diagonal Length", title = "log Weight vs. log Diagonal Length")
```



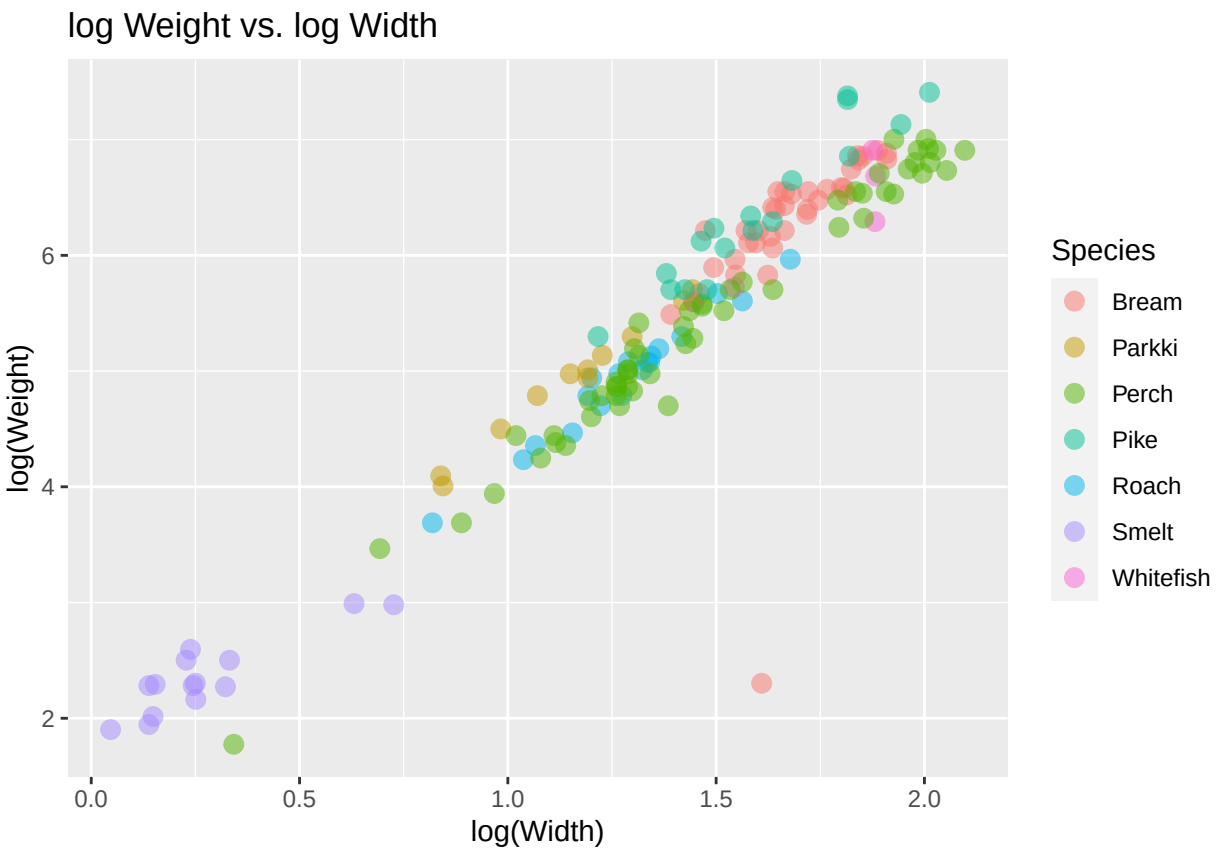

```
ggplot(data = fish, aes(x = log(Clength), y = log(Weight), color = Species)) +
  geom_point(size = 3, alpha = 0.5) +
  labs(x = "log Cross Length", title = "log Weight vs. log Cross Length")
```




```
ggplot(data = fish, aes(x = log(Height), y = log(Weight), color = Species)) +
  geom_point(size = 3, alpha = 0.5) +
  labs(title = "log Weight vs. log Height")
```




```
ggplot(data = fish, aes(x = log(Width), y = log(Weight), color = Species)) +
  geom_point(size = 3, alpha = 0.5) +
  labs(title = "log Weight vs. log Width")
```



Model After Transformation

```
loglog.mdl <- lm(data = fish,
                 log(Weight) ~ Species*log(Vlength) + Species*log(Dlength) + Species*log(Clength) +
                   Species*log(Height) + Species*log(Width))
summary(loglog.mdl)
```

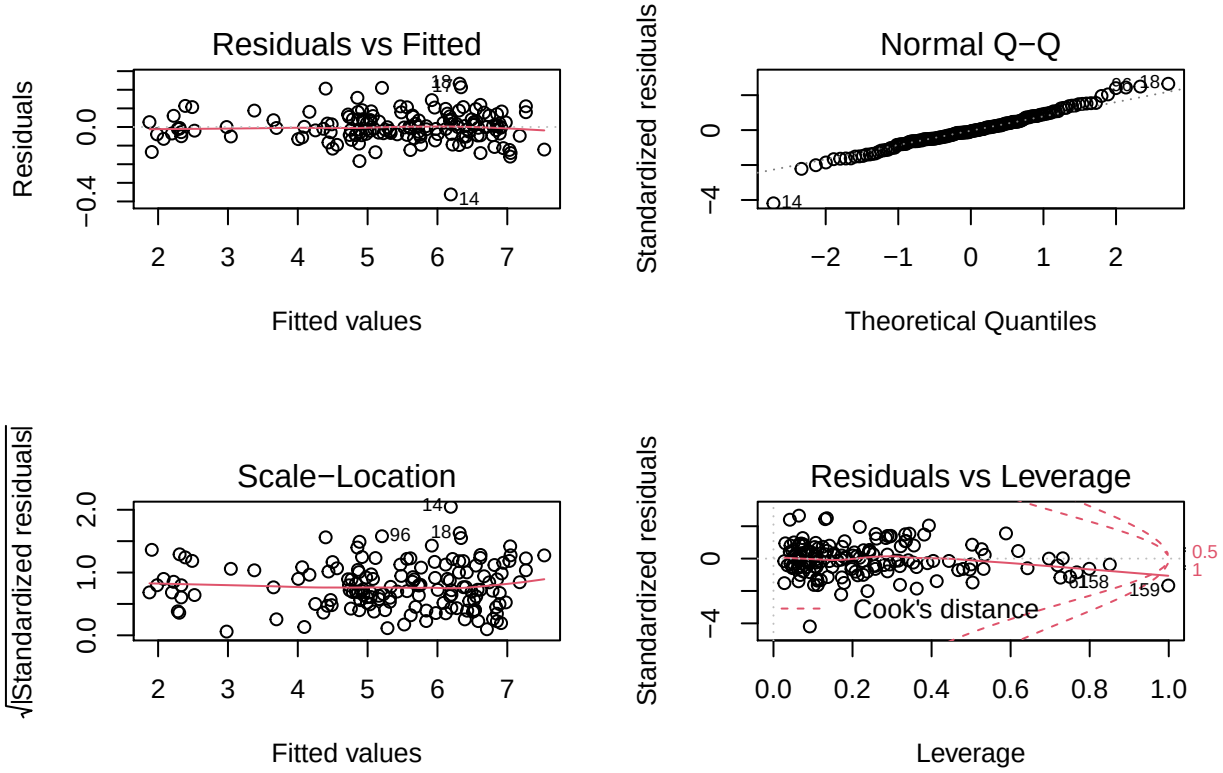
```
##
## Call:
## lm(formula = log(Weight) ~ Species * log(Vlength) + Species *
##     log(Dlength) + Species * log(Clength) + Species * log(Height) +
##     Species * log(Width), data = fish)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36240 -0.04209 -0.00261  0.04087  0.23271
##
```

```
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -0.04231    1.97364  -0.021  0.98293
## SpeciesParkki                    1.42295    4.26351   0.334  0.73917
## SpeciesPerch                    -2.30076    2.03907  -1.128  0.26149
## SpeciesPike                     -4.64114    2.34922  -1.976  0.05055 .
## SpeciesRoach                    -0.11890    2.37363  -0.050  0.96013
## SpeciesSmelt                    -1.54552    2.42779  -0.637  0.52563
## SpeciesWhitefish               -12.56109   14.27409  -0.880  0.38067
## log(Vlength)                     0.03353    2.19090   0.015  0.98782
## log(Dlength)                     4.04747    0.32243  12.553  < 2e-16 ***
## log(Clength)                    -3.40897    2.49665  -1.365  0.17474
## log(Height)                      1.36084    0.44845   3.035  0.00297 **
## log(Width)                       0.52155    0.37147   1.404  0.16296
## SpeciesParkki:log(Vlength)       5.38379   19.40733   0.277  0.78195
## SpeciesPerch:log(Vlength)        0.74442    2.65630   0.280  0.77978
## SpeciesPike:log(Vlength)         1.88305    7.01981   0.268  0.78898
## SpeciesRoach:log(Vlength)        3.04397    3.18785   0.955  0.34161
## SpeciesSmelt:log(Vlength)       -2.43723    6.49583  -0.375  0.70819
## SpeciesWhitefish:log(Vlength) -22.26834   41.18533  -0.541  0.58975
## SpeciesParkki:log(Dlength)      -3.19940   35.31450  -0.091  0.92797
## SpeciesPerch:log(Dlength)       -5.18127    2.48120  -2.088  0.03895 *
## SpeciesPike:log(Dlength)        -7.93261    7.43832  -1.066  0.28841
## SpeciesRoach:log(Dlength)       -7.19154    2.97988  -2.413  0.01736 *
## SpeciesSmelt:log(Dlength)       -0.69935    3.52287  -0.199  0.84298
## SpeciesWhitefish:log(Dlength)    5.80981   52.14507   0.111  0.91148
## SpeciesParkki:log(Clength)      -2.27655   23.08844  -0.099  0.92162
## SpeciesPerch:log(Clength)        5.38759    3.00161   1.795  0.07525 .
## SpeciesPike:log(Clength)         7.81572    3.66428   2.133  0.03502 *
## SpeciesRoach:log(Clength)        4.18434    3.23256   1.294  0.19806
## SpeciesSmelt:log(Clength)        3.72695    4.59041   0.812  0.41850
## SpeciesWhitefish:log(Clength)   19.08398   31.37073   0.608  0.54414
## SpeciesParkki:log(Height)       -0.98995    0.96775  -1.023  0.30844
## SpeciesPerch:log(Height)        -0.55396    0.50275  -1.102  0.27279
## SpeciesPike:log(Height)         -0.67565    0.73691  -0.917  0.36109
## SpeciesRoach:log(Height)        -0.68178    0.64888  -1.051  0.29556
```

```
## SpeciesSmelt:log(Height)       -0.61576    0.66617  -0.924  0.35722
## SpeciesWhitefish:log(Height)    0.25564    3.68390   0.069  0.94479
## SpeciesParkki:log(Width)        1.28871    1.04325   1.235  0.21920
## SpeciesPerch:log(Width)         0.03845    0.41674   0.092  0.92664
## SpeciesPike:log(Width)         -0.37540    0.63899  -0.587  0.55801
## SpeciesRoach:log(Width)         0.86841    0.58276   1.490  0.13887
## SpeciesSmelt:log(Width)        -0.22925    0.47440  -0.483  0.62982
## SpeciesWhitefish:log(Width)    -0.30050    1.63059  -0.184  0.85411
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09075 on 117 degrees of freedom
## Multiple R-squared:  0.9966, Adjusted R-squared:  0.9954
## F-statistic: 843.3 on 41 and 117 DF,  p-value: < 2.2e-16
```

Residual Analysis After Transformation

```
par(mfrow = c(2,2))
plot(loglog.mdl)
```



high leverage points

```
# X matrix
attach(fish)
X <- cbind(rep(1,158), log(Vlength), log(Dlength), log(Clength), log(Height), log(Width))
# Hat matrix
H <- X %*% solve(t(X) %*% X) %*% t(X)
# hii
hii <- diag(H)
# Identify points of high Leverage
p<-ncol(X)
n<-nrow(X)
which(hii>2*p/n)
```

```
## [1]  72 142 143 153 159
```

```r
which(abs(rstandard(loglog.mdl)) >3 )
```

```
## 14
## 14
```

outliers

```r
sort(cooks.distance(loglog.mdl), decreasing = TRUE)
```

```
##          159          158           61           43          132           66
## 5.509202e+01 8.989855e-02 8.723032e-02 8.222378e-02 6.435633e-02 5.663286e-02
##          134           72           14           71          137           52
## 5.283244e-02 4.651339e-02 4.234942e-02 3.758511e-02 3.258445e-02 3.224775e-02
##           34          154          144          142           94           90
## 3.001718e-02 2.961937e-02 2.942839e-02 2.685649e-02 2.541880e-02 2.423511e-02
##           96           77          155          111           20            5
## 2.307877e-02 2.128525e-02 1.984188e-02 1.918404e-02 1.887962e-02 1.865403e-02
##           68          129           13          128           46          143
## 1.849816e-02 1.774105e-02 1.522086e-02 1.519307e-02 1.416574e-02 1.371317e-02
##          139            1           18          147          141           40
## 1.278501e-02 1.171340e-02 1.154424e-02 1.093790e-02 1.069533e-02 1.066233e-02
##          156           36          130           67          145          117
## 1.028830e-02 9.029468e-03 8.897479e-03 8.552111e-03 8.439870e-03 8.069897e-03
##          126          138           38            2          135           53
## 7.187913e-03 7.070635e-03 6.875335e-03 6.723936e-03 6.598652e-03 6.035387e-03
##           73           17           33           79           39          146
## 6.006200e-03 5.947190e-03 5.757835e-03 5.534441e-03 4.918772e-03 4.809639e-03
##          118           75           54          153          127          104
## 4.627870e-03 4.067041e-03 3.702333e-03 3.336922e-03 3.173729e-03 3.061206e-03
##           69          123          149           64          131          150
## 3.055454e-03 2.698131e-03 2.644833e-03 2.494754e-03 2.454158e-03 2.438031e-03
##           44          109          107           86           45           35
## 2.182330e-03 1.877478e-03 1.793404e-03 1.774098e-03 1.711186e-03 1.700468e-03
##            4           76           97           29           51           63
## 1.677562e-03 1.554247e-03 1.554142e-03 1.513311e-03 1.493191e-03 1.458048e-03
##           48          116          101          103          151           15
## 1.455943e-03 1.402731e-03 1.385951e-03 1.321008e-03 1.319266e-03 1.300520e-03
```

```
##           50          121           42           82           41           47
## 1.234567e-03 1.227027e-03 1.221970e-03 1.219880e-03 1.100249e-03 1.058157e-03
##           26           22           37           28           10          106
## 9.021344e-04 8.625108e-04 8.077146e-04 7.982258e-04 6.998498e-04 6.822708e-04
##           95          112           30           65           21           84
## 6.556411e-04 6.046625e-04 5.911899e-04 5.808126e-04 5.328522e-04 5.154839e-04
##          119          113          110          148           87           12
## 4.761039e-04 4.393950e-04 4.182780e-04 4.031792e-04 4.000768e-04 3.849905e-04
##          108           49           91           89            6          136
## 3.478837e-04 3.294536e-04 2.971407e-04 2.646013e-04 2.549970e-04 2.091937e-04
##           81           23           99            7          152           78
## 2.089916e-04 2.074522e-04 1.979370e-04 1.819983e-04 1.819671e-04 1.802034e-04
##           93           92            3           11           24           31
## 1.746574e-04 1.713209e-04 1.679545e-04 1.530376e-04 1.287619e-04 1.282337e-04
##           70           88          102           80          125           19
## 1.279157e-04 1.259732e-04 1.171381e-04 1.077075e-04 1.072024e-04 9.280784e-05
##           83           85            8           25           32          114
## 8.060014e-05 7.612613e-05 7.287500e-05 7.267299e-05 3.885354e-05 3.741122e-05
##          115            9           16           74           62           98
## 3.078506e-05 2.633382e-05 2.184159e-05 2.072023e-05 1.995379e-05 1.722826e-05
##          140          133          120           27          124          105
## 1.681915e-05 1.432719e-05 6.717488e-06 4.535527e-06 4.166449e-06 1.247819e-06
##          157          100          122
## 6.810336e-07 2.633529e-07 1.442005e-07
```

```
fish_without159 <- filter(fish, Observations != 159)
loglog.mdl_without159 <-lm(data = fish_without159,
                           log(Weight) ~ Species*log(Vlength) + Species*log(Dlength) + Species*log(Cleng
                             Species*log(Height) + Species*log(Width))


coef(loglog.mdl)
```

```
##                   (Intercept)                SpeciesParkki
##                   -0.04230745                   1.42294626
##                  SpeciesPerch                  SpeciesPike
##                   -2.30075949                  -4.64113827
##                  SpeciesRoach                  SpeciesSmelt
```

```
##                  -0.11890117                   -1.54552311
##             SpeciesWhitefish                  log(Vlength)
##                 -12.56109012                    0.03352575
##                 log(Dlength)                  log(Clength)
##                   4.04746566                   -3.40897259
##                  log(Height)                   log(Width)
##                   1.36084162                    0.52155414
##    SpeciesParkki:log(Vlength)     SpeciesPerch:log(Vlength)
##                   5.38379042                    0.74442498
##      SpeciesPike:log(Vlength)     SpeciesRoach:log(Vlength)
##                   1.88304663                    3.04397450
##     SpeciesSmelt:log(Vlength) SpeciesWhitefish:log(Vlength)
##                  -2.43723279                  -22.26834141
##    SpeciesParkki:log(Dlength)     SpeciesPerch:log(Dlength)
##                  -3.19940066                   -5.18126955
##      SpeciesPike:log(Dlength)     SpeciesRoach:log(Dlength)
##                  -7.93261310                   -7.19154486
##     SpeciesSmelt:log(Dlength) SpeciesWhitefish:log(Dlength)
##                  -0.69935283                    5.80980894
##    SpeciesParkki:log(Clength)     SpeciesPerch:log(Clength)
##                  -2.27655298                    5.38758627
##      SpeciesPike:log(Clength)     SpeciesRoach:log(Clength)
##                   7.81572048                    4.18433697
##     SpeciesSmelt:log(Clength) SpeciesWhitefish:log(Clength)
##                   3.72695155                   19.08398398
##     SpeciesParkki:log(Height)      SpeciesPerch:log(Height)
##                  -0.98995317                   -0.55395669
##       SpeciesPike:log(Height)      SpeciesRoach:log(Height)
##                  -0.67565290                   -0.68178237
##      SpeciesSmelt:log(Height)  SpeciesWhitefish:log(Height)
##                  -0.61575560                    0.25564205
##      SpeciesParkki:log(Width)       SpeciesPerch:log(Width)
##                   1.28871333                    0.03845143
##        SpeciesPike:log(Width)       SpeciesRoach:log(Width)
##                  -0.37539721                    0.86841342
##       SpeciesSmelt:log(Width)   SpeciesWhitefish:log(Width)
```

```
##                   -0.22925358                   -0.30049561
```

```
coef(loglog.mdl_without159)
```

```
##                   (Intercept)                 SpeciesParkki
##                   -0.70675310                    2.08739191
##                  SpeciesPerch                   SpeciesPike
##                   -1.63631384                   -3.97669261
##                  SpeciesRoach                   SpeciesSmelt
##                    0.54554449                   -0.88107745
##              SpeciesWhitefish                  log(Vlength)
##                  -11.89664447                    2.11509993
##                  log(Dlength)                  log(Clength)
##                   -0.91246824                   -0.55811465
##                   log(Height)                   log(Width)
##                    1.49945449                    0.60875925
##    SpeciesParkki:log(Vlength)    SpeciesPerch:log(Vlength)
##                    3.30221624                   -1.33714920
##      SpeciesPike:log(Vlength)    SpeciesRoach:log(Vlength)
##                   -0.19852755                    0.96240032
##     SpeciesSmelt:log(Vlength) SpeciesWhitefish:log(Vlength)
##                   -4.51880697                  -24.34991559
##    SpeciesParkki:log(Dlength)    SpeciesPerch:log(Dlength)
##                    1.76053325                   -0.22133564
##      SpeciesPike:log(Dlength)    SpeciesRoach:log(Dlength)
##                   -2.97267919                   -2.23161095
##     SpeciesSmelt:log(Dlength) SpeciesWhitefish:log(Dlength)
##                    4.26058108                   10.76974285
##    SpeciesParkki:log(Clength)    SpeciesPerch:log(Clength)
##                   -5.12741092                    2.53672833
##      SpeciesPike:log(Clength)    SpeciesRoach:log(Clength)
##                    4.96486254                    1.33347903
##     SpeciesSmelt:log(Clength) SpeciesWhitefish:log(Clength)
##                    0.87609361                   16.23312605
##     SpeciesParkki:log(Height)     SpeciesPerch:log(Height)
##                   -1.12856604                   -0.69256956
##       SpeciesPike:log(Height)     SpeciesRoach:log(Height)
```

```
##                   -0.81426578                   -0.82039524
##      SpeciesSmelt:log(Height)  SpeciesWhitefish:log(Height)
##                   -0.75436847                    0.11702918
##      SpeciesParkki:log(Width)       SpeciesPerch:log(Width)
##                    1.20150822                   -0.04875368
##        SpeciesPike:log(Width)       SpeciesRoach:log(Width)
##                   -0.46260232                    0.78120830
##       SpeciesSmelt:log(Width)   SpeciesWhitefish:log(Width)
##                   -0.31645870                   -0.38770072
```

```
# Remove outliers
fish <- filter(fish, Observations != 159)
```

Model Selection

```
loglog.mdl <- lm(data = fish,
                 log(Weight) ~ Species*log(Vlength) + Species*log(Dlength) + Species*log(Clength) +
                   Species*log(Height) + Species*log(Width))
step(loglog.mdl, direction = "both")
```

```
## Start:  AIC=-725.54
## log(Weight) ~ Species * log(Vlength) + Species * log(Dlength) +
##     Species * log(Clength) + Species * log(Height) + Species *
##     log(Width)
##
##                         Df Sum of Sq     RSS     AIC
## - Species:log(Vlength)  6  0.012679 0.95345 -735.42
## - Species:log(Clength)  6  0.018367 0.95914 -734.48
## - Species:log(Dlength)  6  0.018577 0.95935 -734.45
## - Species:log(Height)   6  0.022048 0.96282 -733.88
## - Species:log(Width)    6  0.054039 0.99482 -728.71
## <none>                             0.94078 -725.54
##
## Step:  AIC=-735.42
## log(Weight) ~ Species + log(Vlength) + log(Dlength) + log(Clength) +
##     log(Height) + log(Width) + Species:log(Dlength) + Species:log(Clength) +
##     Species:log(Height) + Species:log(Width)
```

```
##
##                         Df Sum of Sq     RSS     AIC
## - Species:log(Dlength)  6  0.024095 0.97755 -743.48
## - Species:log(Height)   6  0.025587 0.97904 -743.24
## - Species:log(Clength)  6  0.031406 0.98486 -742.30
## - Species:log(Width)    6  0.057795 1.01125 -738.12
## <none>                             0.95345 -735.42
## - log(Vlength)          1  0.014876 0.96833 -734.97
## + Species:log(Vlength)  6  0.012679 0.94078 -725.54
##
## Step:  AIC=-743.48
## log(Weight) ~ Species + log(Vlength) + log(Dlength) + log(Clength) +
##     log(Height) + log(Width) + Species:log(Clength) + Species:log(Height) +
##     Species:log(Width)
##
##                         Df Sum of Sq     RSS     AIC
## - Species:log(Height)   6  0.032412 1.00996 -750.32
## - Species:log(Width)    6  0.050897 1.02845 -747.46
## - Species:log(Clength)  6  0.065189 1.04274 -745.28
## - log(Dlength)          1  0.002517 0.98007 -745.07
## - log(Vlength)          1  0.007571 0.98512 -744.26
## <none>                             0.97755 -743.48
## + Species:log(Dlength)  6  0.024095 0.95345 -735.42
## + Species:log(Vlength)  6  0.018197 0.95935 -734.45
##
## Step:  AIC=-750.32
## log(Weight) ~ Species + log(Vlength) + log(Dlength) + log(Clength) +
##     log(Height) + log(Width) + Species:log(Clength) + Species:log(Width)
##
##                         Df Sum of Sq     RSS     AIC
## - Species:log(Width)    6  0.057222 1.06718 -753.62
## - log(Dlength)          1  0.003059 1.01302 -751.85
## - log(Vlength)          1  0.005827 1.01579 -751.41
## - Species:log(Clength)  6  0.076330 1.08629 -750.81
## <none>                             1.00996 -750.32
## + Species:log(Height)   6  0.032412 0.97755 -743.48
```

```
## + Species:log(Dlength)  6  0.030921 0.97904 -743.24
## + Species:log(Vlength)  6  0.025346 0.98462 -742.34
## - log(Height)           1  0.202717 1.21268 -723.42
##
## Step:  AIC=-753.62
## log(Weight) ~ Species + log(Vlength) + log(Dlength) + log(Clength) +
##     log(Height) + log(Width) + Species:log(Clength)
##
##                         Df Sum of Sq    RSS     AIC
## - Species:log(Clength)  6  0.064666 1.1319 -756.32
## - log(Dlength)          1  0.001914 1.0691 -755.33
## - log(Vlength)          1  0.005713 1.0729 -754.77
## <none>                             1.0672 -753.62
## + Species:log(Width)    6  0.057222 1.0100 -750.32
## + Species:log(Height)   6  0.038737 1.0285 -747.46
## + Species:log(Dlength)  6  0.019518 1.0477 -744.53
## + Species:log(Vlength)  6  0.015565 1.0516 -743.94
## - log(Width)            1  0.187522 1.2547 -730.04
## - log(Height)           1  0.199915 1.2671 -728.49
##
## Step:  AIC=-756.32
## log(Weight) ~ Species + log(Vlength) + log(Dlength) + log(Clength) +
##     log(Height) + log(Width)
##
##                         Df Sum of Sq    RSS     AIC
## - log(Dlength)          1   0.00957 1.1414 -756.99
## <none>                             1.1319 -756.32
## - log(Vlength)          1   0.01604 1.1479 -756.10
## + Species:log(Clength)  6   0.06467 1.0672 -753.62
## - log(Clength)          1   0.03536 1.1672 -753.46
## + Species:log(Vlength)  6   0.06140 1.0704 -753.13
## + Species:log(Dlength)  6   0.05951 1.0723 -752.85
## + Species:log(Height)   6   0.04781 1.0840 -751.14
## + Species:log(Width)    6   0.04556 1.0863 -750.81
## - log(Height)           1   0.17913 1.3110 -735.11
## - log(Width)            1   0.18016 1.3120 -734.98
```

```
## - Species               6   0.32951 1.4614 -727.95
##
## Step:  AIC=-756.99
## log(Weight) ~ Species + log(Vlength) + log(Clength) + log(Height) +
##     log(Width)
##
##                         Df Sum of Sq    RSS     AIC
## - log(Vlength)           1   0.00667 1.1481 -758.07
## <none>                               1.1414 -756.99
## + log(Dlength)           1   0.00957 1.1319 -756.32
## - log(Clength)           1   0.02582 1.1672 -755.46
## + Species:log(Clength)   6   0.07232 1.0691 -755.33
## + Species:log(Vlength)   6   0.06897 1.0724 -754.84
## + Species:log(Height)    6   0.05472 1.0867 -752.75
## + Species:log(Width)     6   0.05408 1.0873 -752.66
## - log(Width)             1   0.17275 1.3142 -736.72
## - log(Height)            1   0.17354 1.3150 -736.63
## - Species                6   0.36175 1.5032 -725.49
##
## Step:  AIC=-758.07
## log(Weight) ~ Species + log(Clength) + log(Height) + log(Width)
##
##                         Df Sum of Sq    RSS     AIC
## <none>                               1.1481 -758.07
## + log(Vlength)           1   0.00667 1.1414 -756.99
## + Species:log(Clength)   6   0.07486 1.0732 -756.73
## + log(Dlength)           1   0.00019 1.1479 -756.10
## + Species:log(Height)    6   0.05873 1.0894 -754.37
## + Species:log(Width)     6   0.05498 1.0931 -753.82
## - log(Width)             1   0.16662 1.3147 -738.66
## - log(Height)            1   0.17508 1.3232 -737.65
## - Species                6   0.57760 1.7257 -705.68
## - log(Clength)           1   1.03945 2.1875 -658.21
##
## Call:
## lm(formula = log(Weight) ~ Species + log(Clength) + log(Height) +
```

```
##     log(Width), data = fish)
##
## Coefficients:
##      (Intercept)     SpeciesParkki      SpeciesPerch       SpeciesPike
##         -2.96739           0.15545           0.26013           0.12893
##     SpeciesRoach      SpeciesSmelt  SpeciesWhitefish      log(Clength)
##          0.15256           0.03308           0.27848           1.79408
##      log(Height)        log(Width)
##          0.70609           0.52629
```

```
loglog.mdl <- lm(data = fish, log(Weight) ~ Species + log(Clength) + log(Height) + log(Width))
summary(loglog.mdl)
```

```
##
## Call:
## lm(formula = log(Weight) ~ Species + log(Clength) + log(Height) +
##     log(Width), data = fish)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41008 -0.05420 -0.00258  0.05766  0.21814
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2.96739    0.26835 -11.058  < 2e-16 ***
## SpeciesParkki     0.15545    0.03501   4.440 1.75e-05 ***
## SpeciesPerch      0.26013    0.06915   3.762 0.000242 ***
## SpeciesPike       0.12893    0.12894   1.000 0.318991
## SpeciesRoach      0.15256    0.06359   2.399 0.017684 *
## SpeciesSmelt      0.03308    0.10963   0.302 0.763290
## SpeciesWhitefish  0.27848    0.06478   4.299 3.10e-05 ***
## log(Clength)      1.79408    0.15499  11.576  < 2e-16 ***
## log(Height)       0.70609    0.14863   4.751 4.75e-06 ***
## log(Width)        0.52629    0.11356   4.635 7.79e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

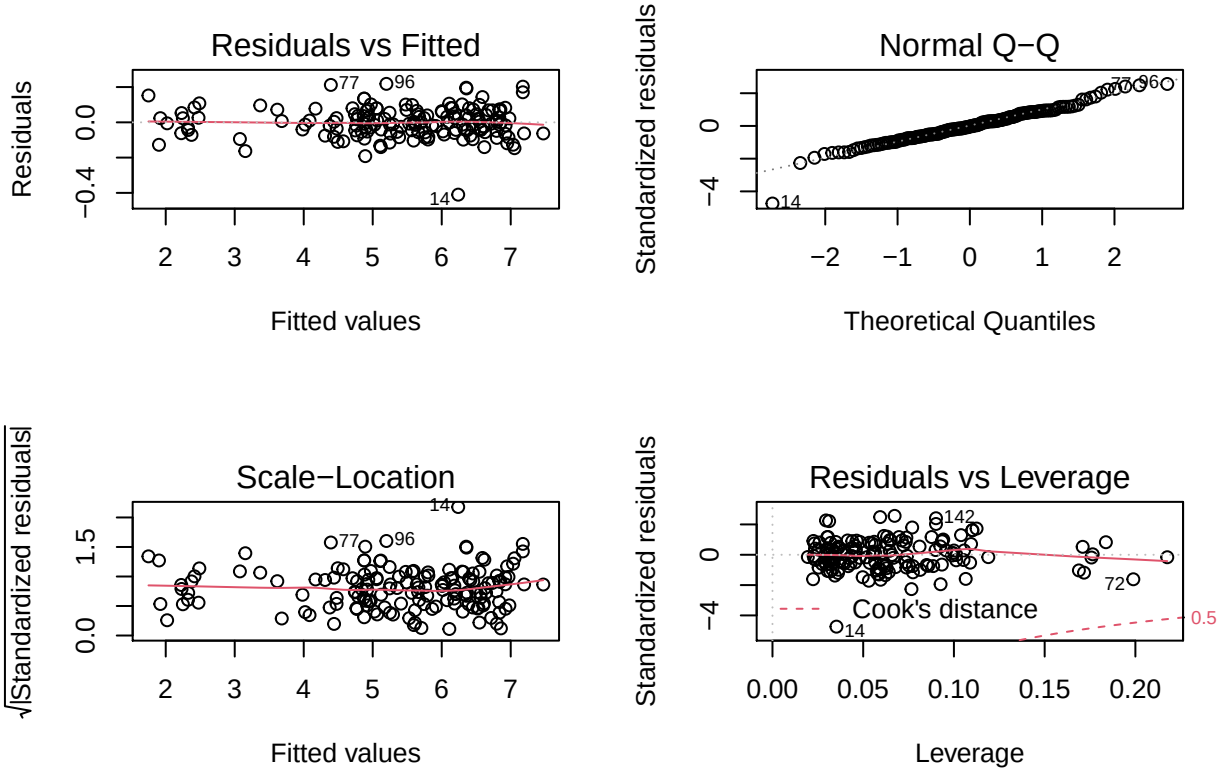
```
## Residual standard error: 0.08808 on 148 degrees of freedom
## Multiple R-squared:  0.9958, Adjusted R-squared:  0.9956
## F-statistic:  3939 on 9 and 148 DF,  p-value: < 2.2e-16
```

Further Analysis after Model Selection

```
# Multicollinearity
loglog.mdl <- lm(data = fish, log(Weight) ~ Species + log(Clength) + log(Height) + log(Width))
vif(loglog.mdl)
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## Species       475.10874  6        1.671345
## log(Clength)   80.52344  1        8.973486
## log(Height)   142.64291  1       11.943321
## log(Width)     57.34860  1        7.572886
```

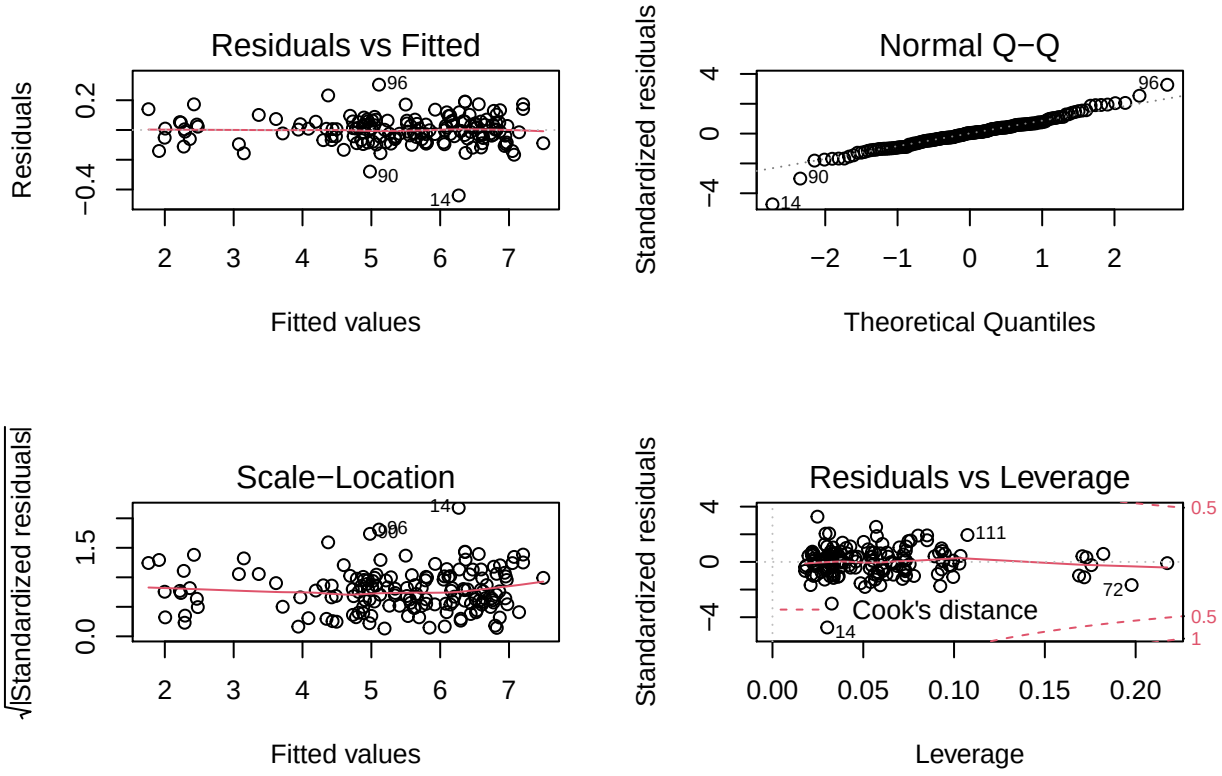
```
par(mfrow = c(2,2))
plot(loglog.mdl)
```



```r
# Refit Without variable Height
loglog.mdl <- lm(data = fish, log(Weight) ~ Species + log(Clength) + log(Width))
vif(loglog.mdl)
```

```
##                  GVIF Df GVIF^(1/(2*Df))
## Species      17.79288  6        1.271122
## log(Clength) 44.22697  1        6.650336
## log(Width)   44.42533  1        6.665233
```

```r
plot(loglog.mdl)
```




```r
# Check high standardized residual point 14
fish_without14 <- fish[-14,]
loglog.mdl_without14 <- lm(data = fish_without14, log(Weight) ~ Species + log(Clength) + log(Width))
coef(loglog.mdl)
```

```
##      (Intercept)     SpeciesParkki      SpeciesPerch       SpeciesPike
##      -3.28333900       0.17591364      -0.04493055      -0.44947271
##      SpeciesRoach     SpeciesSmelt SpeciesWhitefish     log(Clength)
##      -0.11548255      -0.43618404       0.04059123       2.28842164
```

```
##       log(Width)
##       0.78238352
```

```
coef(loglog.mdl_without14)
```

```
##      (Intercept)    SpeciesParkki     SpeciesPerch      SpeciesPike
##      -3.31360818       0.16344760      -0.05531334      -0.46879330
##     SpeciesRoach     SpeciesSmelt SpeciesWhitefish     log(Clength)
##      -0.12733482      -0.45464163       0.03005336       2.30959719
##       log(Width)
##       0.76225683
```

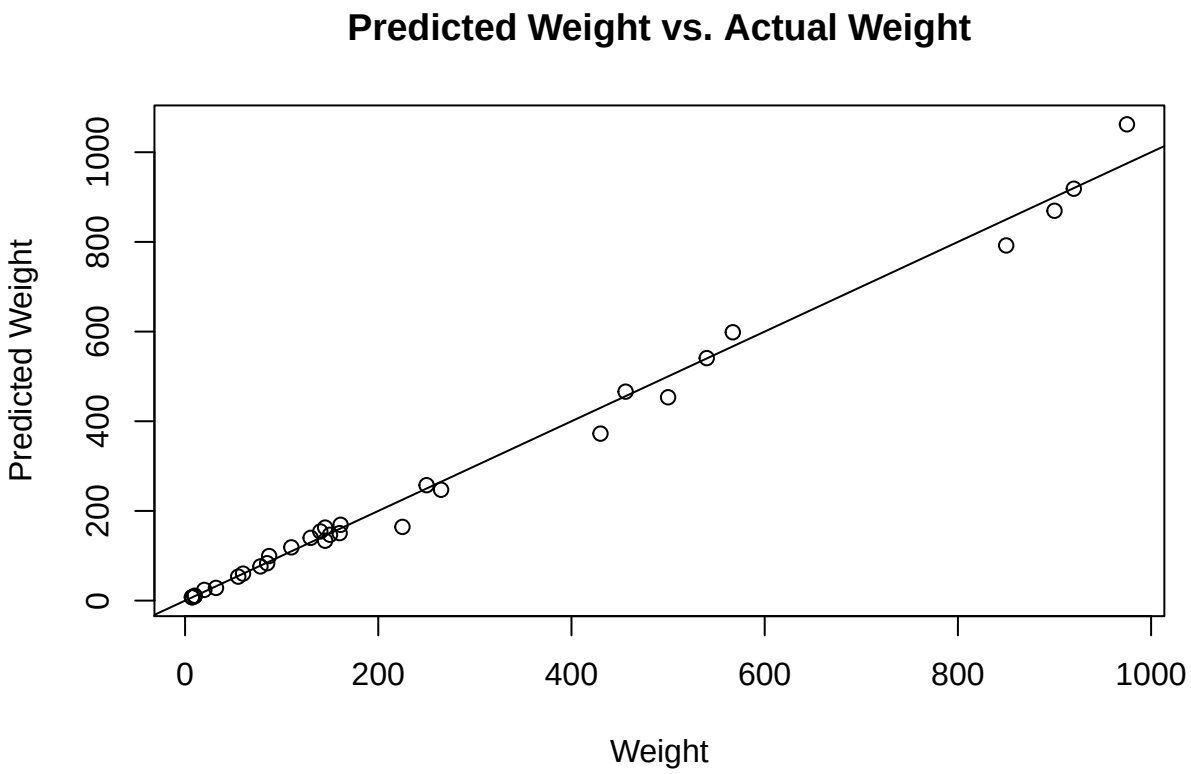
```
# final model
final_mdl <- lm(data = fish, log(Weight) ~ Species + log(Clength) + log(Width))
summary(final_mdl)
```

```
##
## Call:
## lm(formula = log(Weight) ~ Species + log(Clength) + log(Width),
##     data = fish)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44001 -0.05091  0.00274  0.05304  0.30444
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -3.28334    0.27816 -11.804  < 2e-16 ***
## SpeciesParkki     0.17591    0.03718   4.732 5.13e-06 ***
## SpeciesPerch     -0.04493    0.02745  -1.637 0.103791
## SpeciesPike      -0.44947    0.04544  -9.892  < 2e-16 ***
## SpeciesRoach     -0.11548    0.03139  -3.679 0.000326 ***
## SpeciesSmelt     -0.43618    0.05089  -8.572 1.20e-14 ***
## SpeciesWhitefish  0.04059    0.04397   0.923 0.357386
## log(Clength)      2.28842    0.12290  18.621  < 2e-16 ***
## log(Width)        0.78238    0.10694   7.316 1.46e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

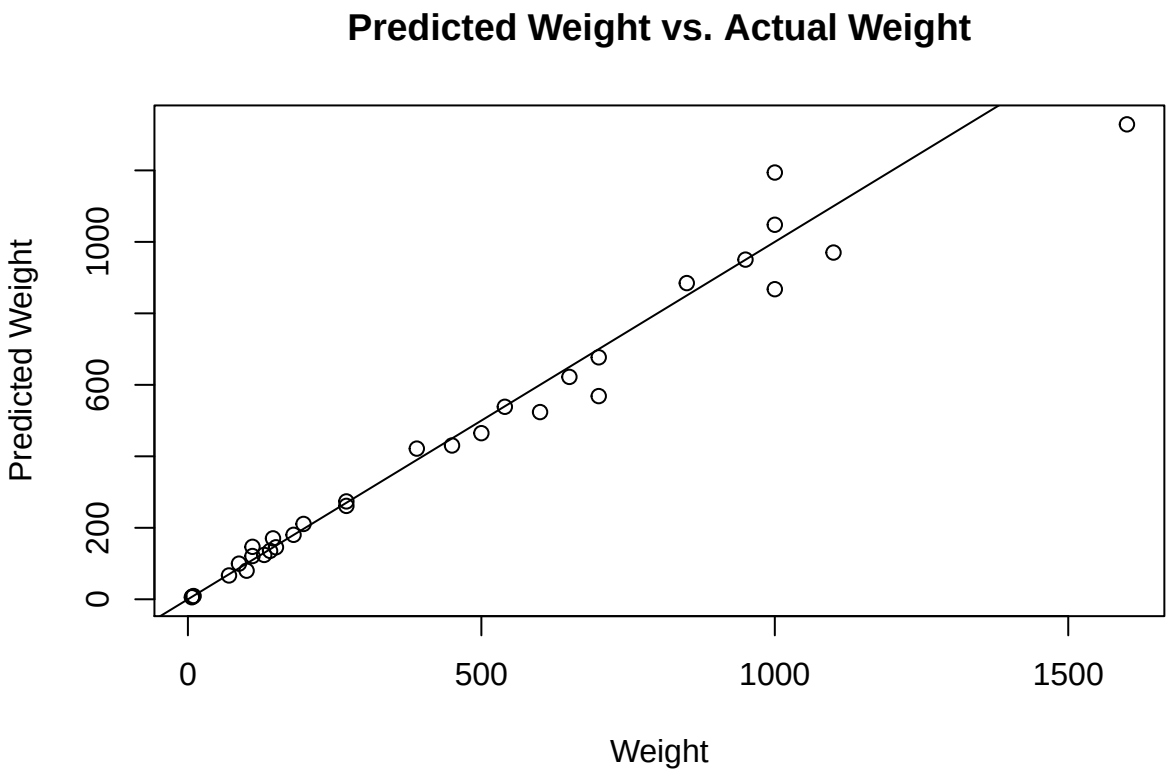
```
##
## Residual standard error: 0.09424 on 149 degrees of freedom
## Multiple R-squared:  0.9952, Adjusted R-squared:  0.995
## F-statistic:  3868 on 8 and 149 DF,  p-value: < 2.2e-16
```

Cross Validation

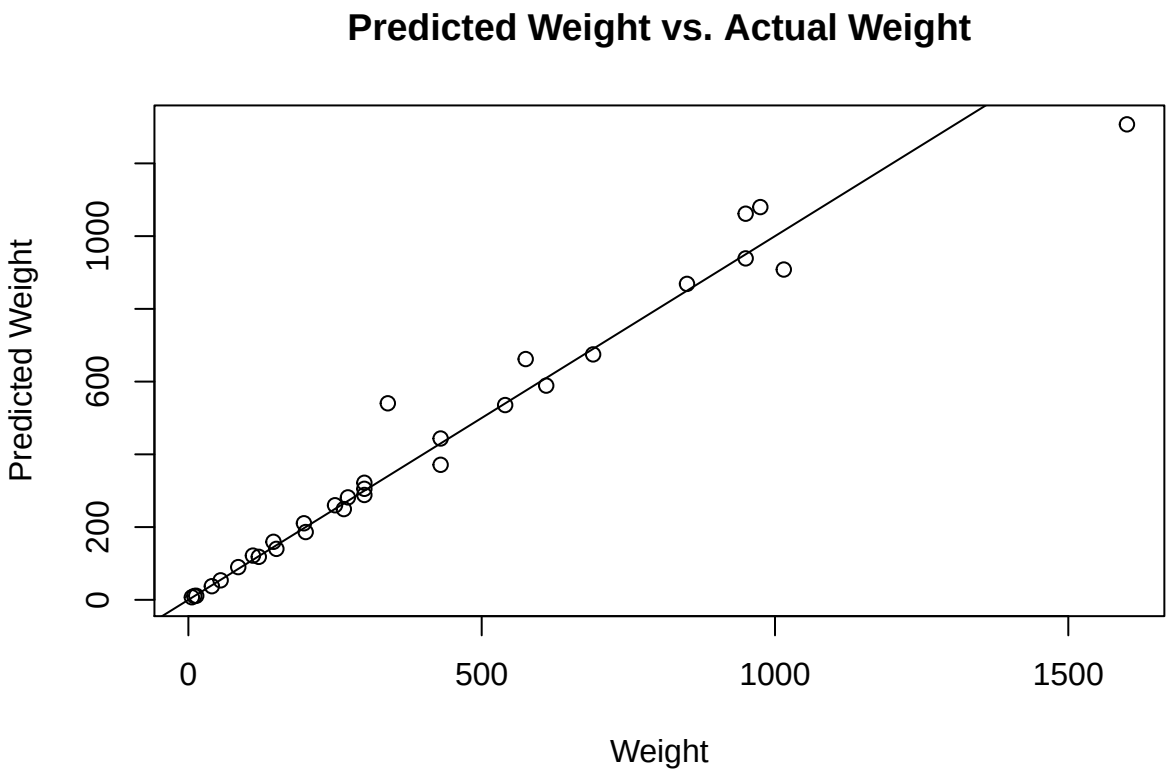
```
set.seed(10086)
for (i in 1:5){
nsamp <- ceiling(0.8*length(fish$Weight))
training_samps <- sample(c(1:length(fish$Weight)),nsamp)
train_data  <- fish[training_samps, ]
test_data <-   fish[-training_samps, ]
train.lm <- lm(data = train_data, log(Weight) ~ Species + log(Clength)  + log(Width))
summary(train.lm)
preds <- exp(predict(train.lm,test_data))
plot(test_data$Weight, preds,
     xlab = "Weight", ylab = "Predicted Weight", main = "Predicted Weight vs. Actual Weight")
abline(c(0,1))
R.sq <- R2(preds, test_data$Weight)
RMSPE <- RMSE(preds, test_data$Weight)
MAPE <- MAE(preds, test_data$Weight)
print(c(i,R.sq,RMSPE,MAPE))
}
```



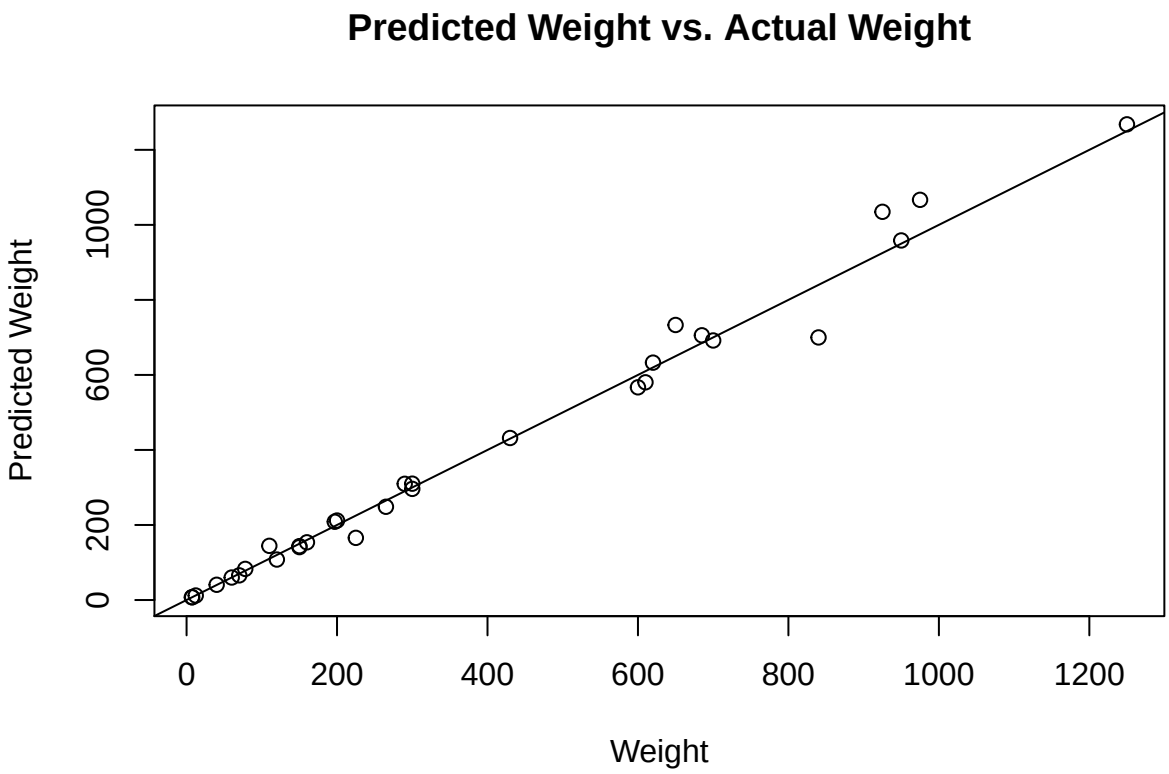

```
## [1]  1.0000000  0.9912097 27.6397102 16.7467734
```



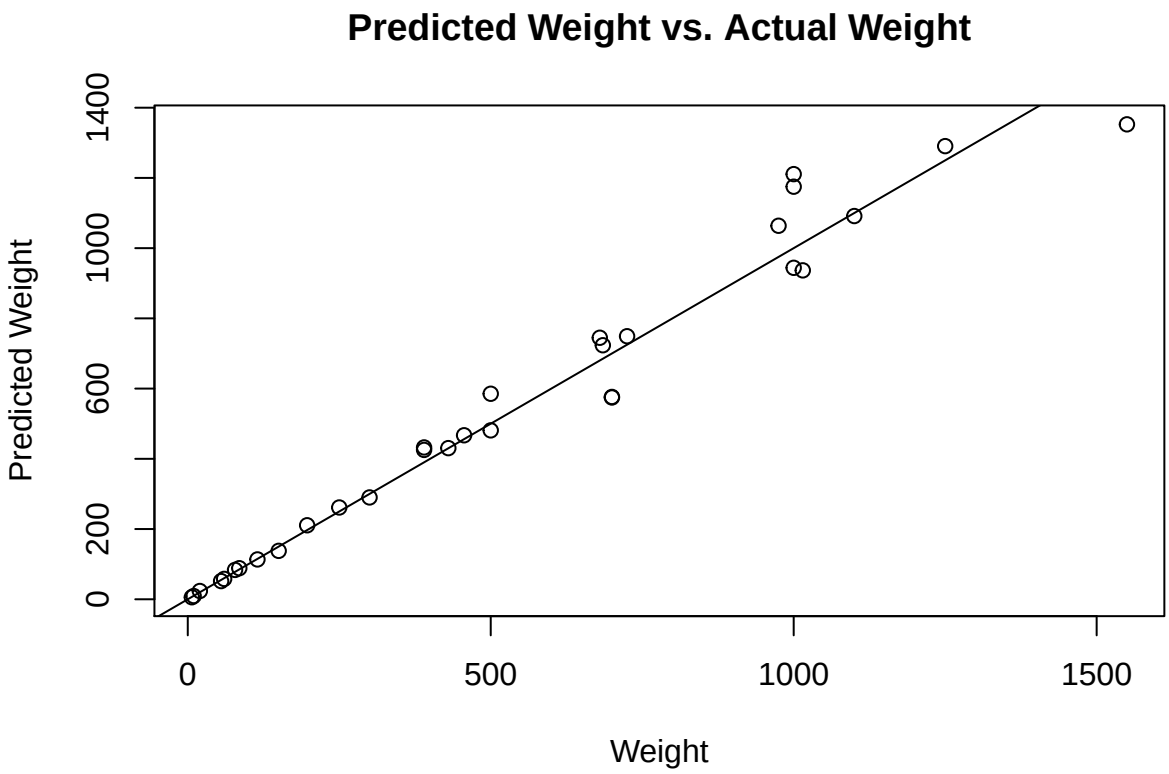

```
## [1]  2.0000000  0.9683909 76.0219676 42.2484542
```




```
## [1]  3.000000  0.961259 75.128820 38.647781
```

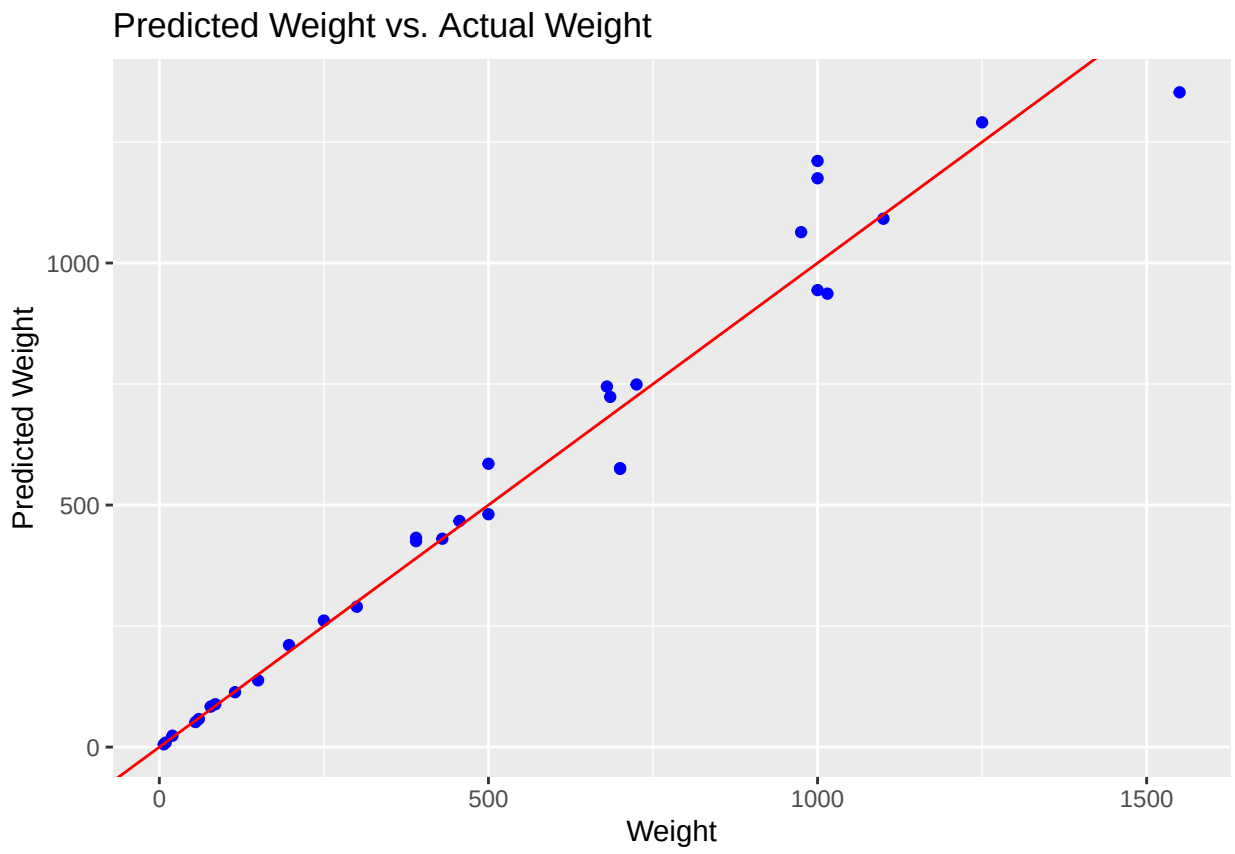



```
## [1]  4.0000000  0.9860467 42.5577402 24.8099605
```




```
## [1]  5.0000000  0.9670864 76.5467564 48.1780445
```

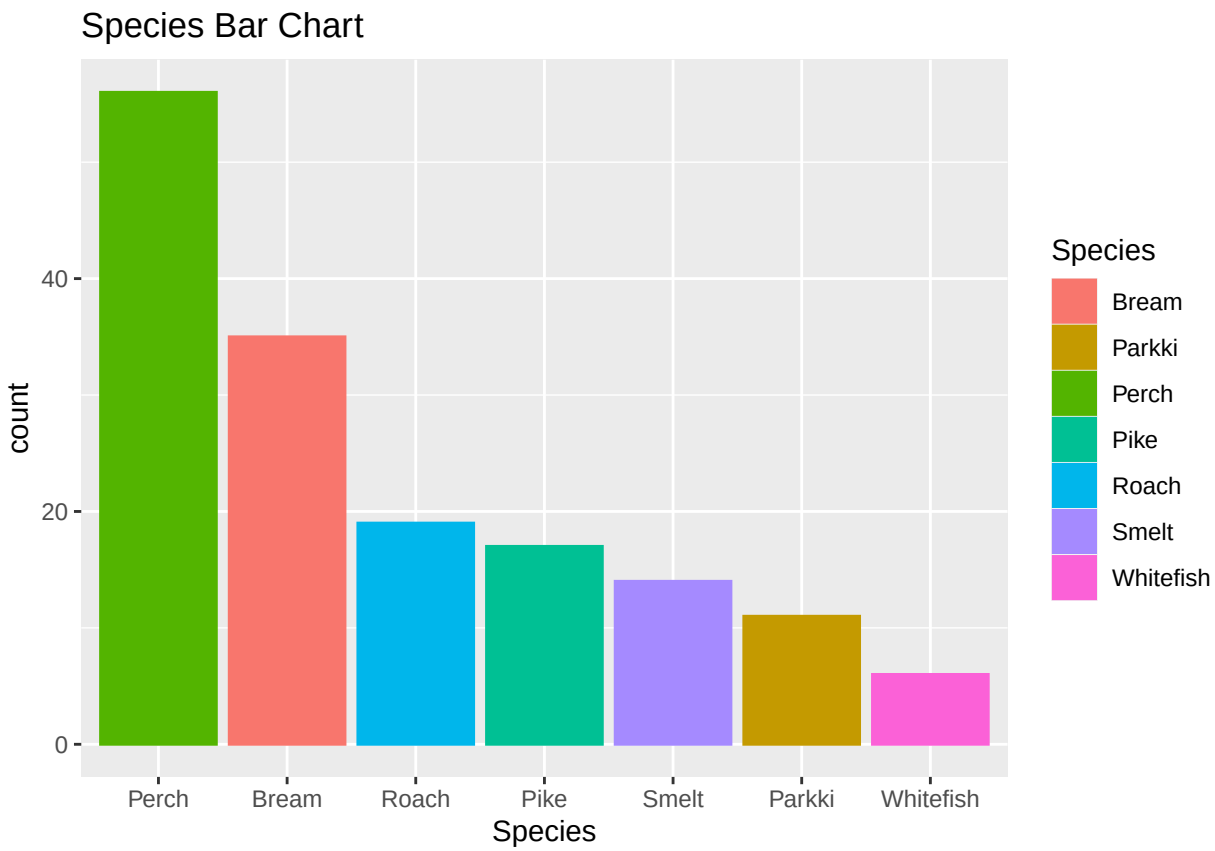
```
point.data <- tibble(preds, test_data$Weight)
ggplot(data = point.data, aes(x = test_data$Weight, y = preds)) +
  geom_point(color = "blue") +
  geom_abline(slope = 1, intercept = 0, color = "red") +
  labs(x = "Weight", y = "Predicted Weight", title = "Predicted Weight vs. Actual Weight")
```



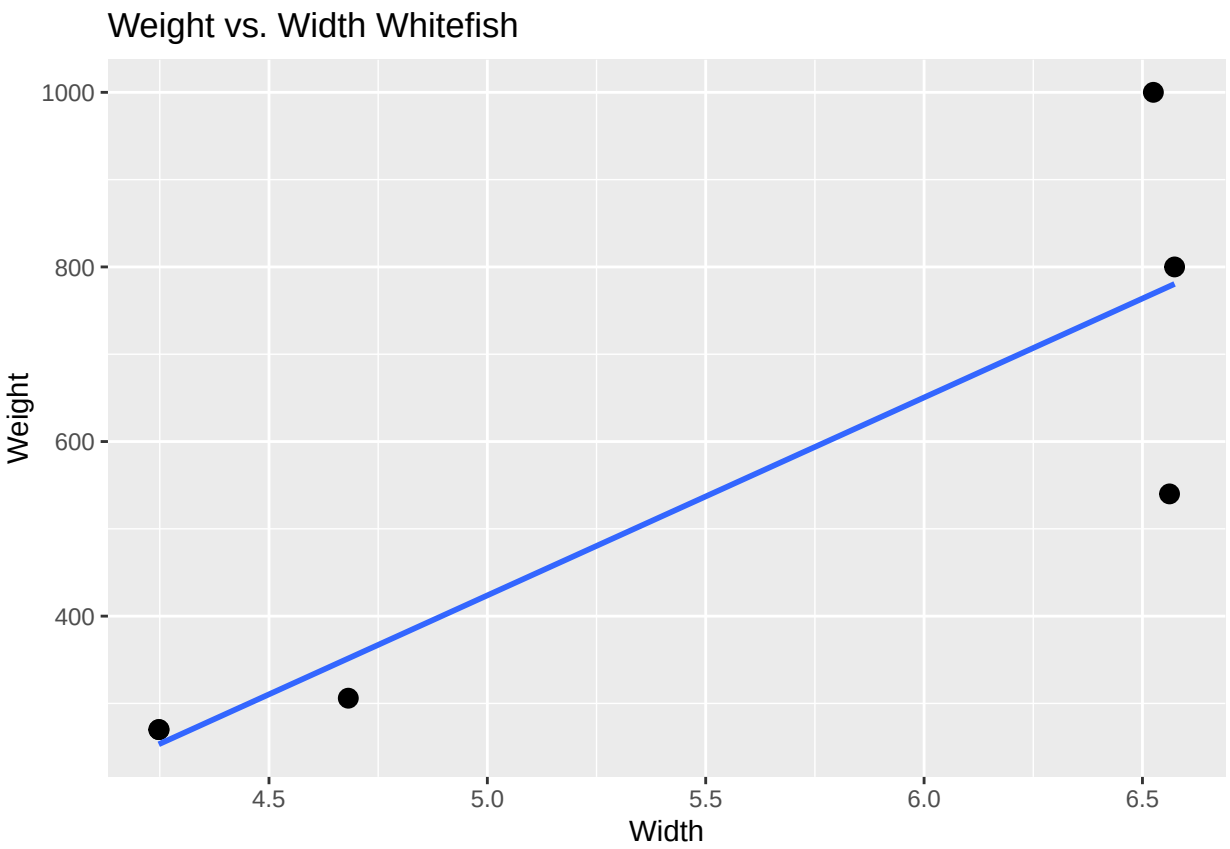

```
fish %>% group_by(Species) %>% count()
```

```
## # A tibble: 7 x 2
## # Groups:   Species [7]
##   Species       n
##   <fct>     <int>
## 1 Bream        35
## 2 Parkki       11
## 3 Perch        56
## 4 Pike         17
## 5 Roach        19
## 6 Smelt        14
## 7 Whitefish     6
```

```
ggplot(data = fish, aes(x = fct_infreq(Species), color = Species, fill = Species)) +
  geom_bar() +
  labs(x = "Species", title = "Species Bar Chart")
```




```
Whitefish <- filter(fish, Species %in% "Whitefish")
ggplot(data = Whitefish, aes(x = Width, y = Weight)) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Weight vs. Width Whitefish")
```


Weight vs. Width Whitefish
1000
800
600
400
Weight
4.5
5.0
5.5
6.0
6.5
Width