Descriptive Analysis of the Data

The purpose of the descriptive analysis is to get a better understanding of the raw data before the actual regression analysis is started. In general, this means determining how clean the data is, its size and scale, and take note of any possible patterns or inconsistencies that may impact the regression analysis.

```
data <- read.csv("fish-market-data.csv", fileEncoding="UTF-8-BOM")
```

Structure

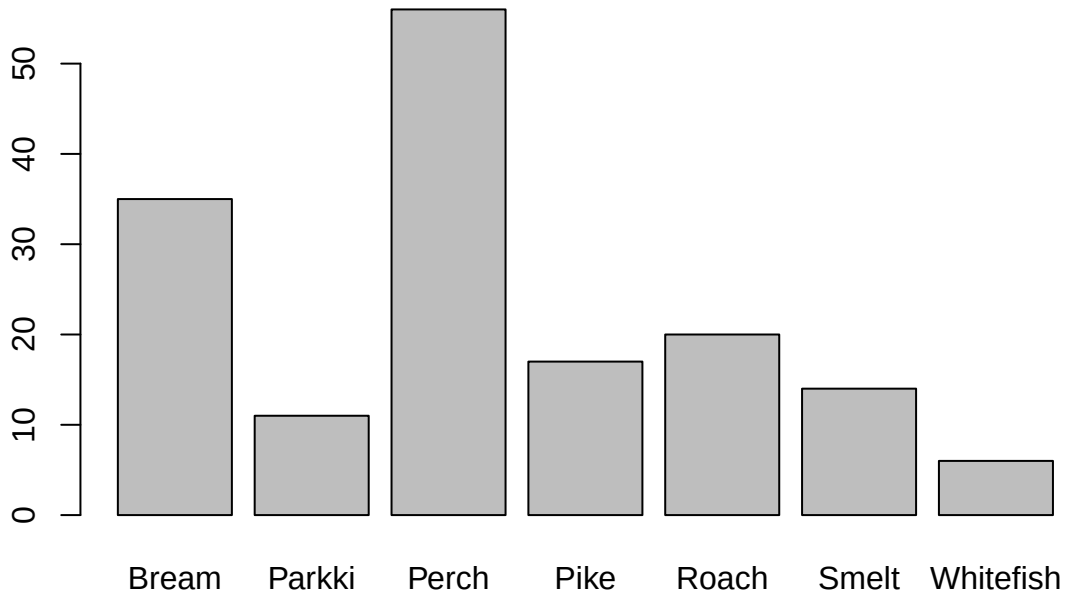
```
str(data)
```

```
## 'data.frame':    159 obs. of  7 variables:
##  $ Species: chr  "Bream" "Bream" "Bream" "Bream" ...
##  $ Weight : num  242 290 340 363 430 450 500 390 450 500 ...
##  $ Length1: num  23.2 24 23.9 26.3 26.5 26.8 26.8 27.6 27.6 28.5 ...
##  $ Length2: num  25.4 26.3 26.5 29 29 29.7 29.7 30 30 30.7 ...
##  $ Length3: num  30 31.2 31.1 33.5 34 34.7 34.5 35 35.1 36.2 ...
##  $ Height : num  11.5 12.5 12.4 12.7 12.4 ...
##  $ Width  : num  4.02 4.31 4.7 4.46 5.13 ...
```

There are 7 total variables, 6 of which are numbers and one of which is an indicator variable.

Qualitative Variables

```
barplot(table(data$Species))
```



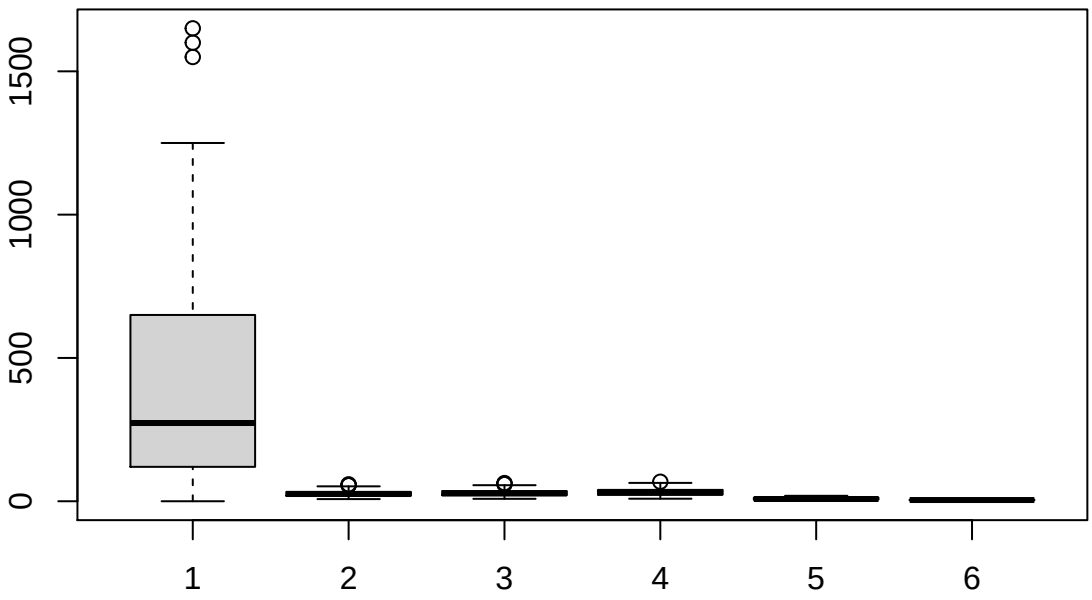
There are more Perch samples than any other type of fish, while there are very few Whitefish samples.

Summary Statistics

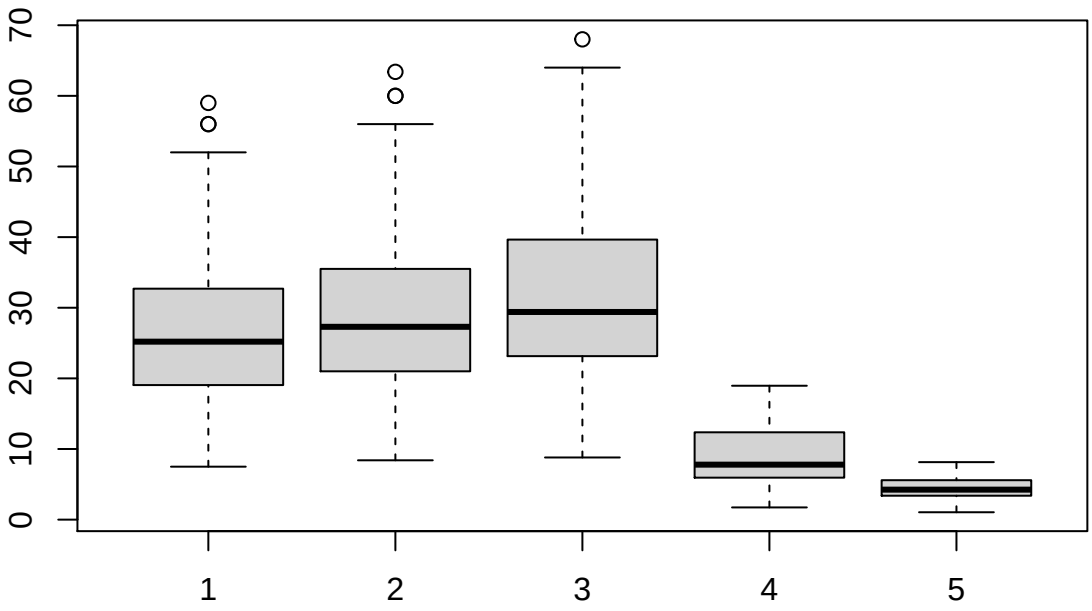
```
summary(data)
```

```
##    Species              Weight          Length1         Length2
##  Length:159         Min.   :   0.0   Min.   : 7.50   Min.   : 8.40
##  Class :character   1st Qu.: 120.0   1st Qu.:19.05   1st Qu.:21.00
##  Mode  :character   Median : 273.0   Median :25.20   Median :27.30
##                     Mean   : 398.3   Mean   :26.25   Mean   :28.42
##                     3rd Qu.: 650.0   3rd Qu.:32.70   3rd Qu.:35.50
##                     Max.   :1650.0   Max.   :59.00   Max.   :63.40
##     Length3          Height           Width
##  Min.   : 8.80   Min.   : 1.728   Min.   :1.048
##  1st Qu.:23.15   1st Qu.: 5.945   1st Qu.:3.386
##  Median :29.40   Median : 7.786   Median :4.248
##  Mean   :31.23   Mean   : 8.971   Mean   :4.417
##  3rd Qu.:39.65   3rd Qu.:12.366   3rd Qu.:5.585
##  Max.   :68.00   Max.   :18.957   Max.   :8.142
```

```
boxplot(data$Weight, data$Length1, data$Length2, data$Length3, data$Height, data$Width)
```



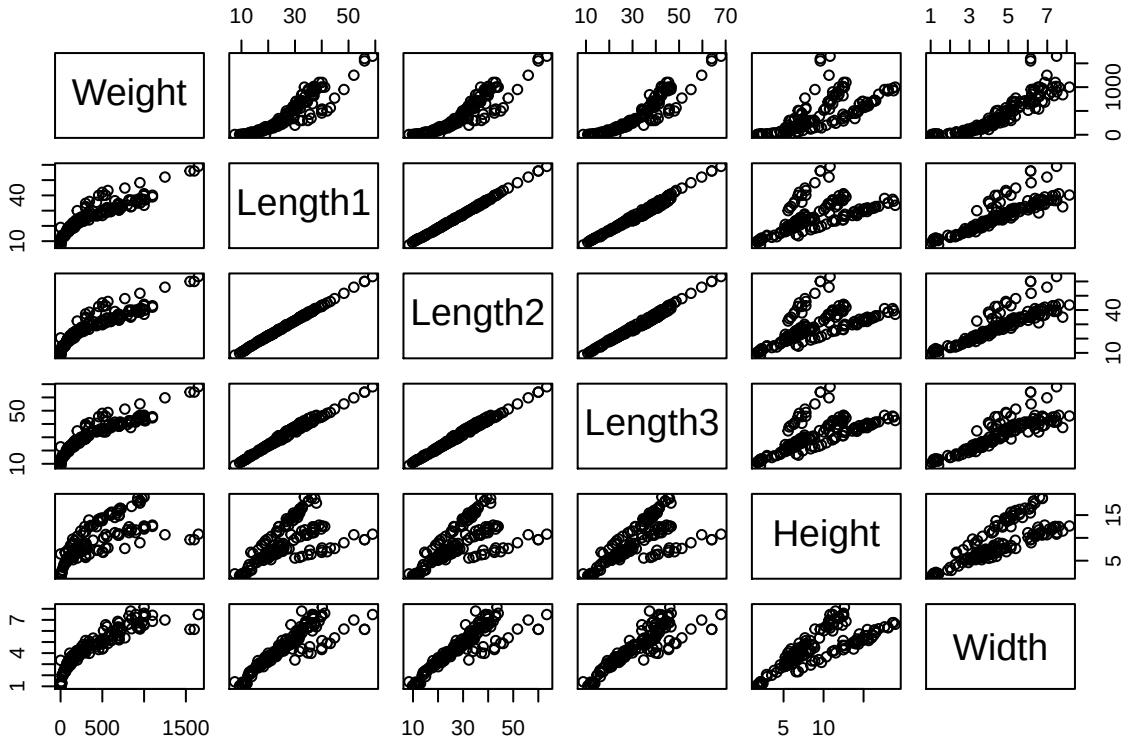

```
boxplot(data$Length1, data$Length2, data$Length3, data$Height, data$Width)
```



Clearly weight is in a different unit than the other predictors. It will make sense to scale the data for comparability of the regression coefficients. If Length 1 2 and 3 are the same measurement units as Width and Height, then this group may be measuring the length of the fish in slightly different ways. Weight seems to be highly variable and heavily skewed to the right with large outliers

Summary Statistics

```
pairs(data[,c(2:7)])
```



Looking at the pairs plot, there appears to be some challenges ahead. Assuming weight is the response variable variable, it does appear to have a relation with every other predictor. However, the best relationship does not appear to be linear. This means that the data may have to be transformed to reflect a linear relationship between the response and predictors. Length 1, Length 2, and Length 3 all appear to be impacted by multicollinearity. Width and Height may have a the same issue, but have different 'groups' of correlation which may be explained by the categorical variable Species.