

Data Cleaning

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

Manually browsing the data, it seems pretty clean. Some issues came up in the descriptive analysis though. In the summary data, one or more observations have zero weight.

```
data[data$Weight == 0, ]
```

```
##      Species Weight Length1 Length2 Length3 Height  Width
## 41    Roach      0      19    20.5    22.8 6.4752 3.3516
```

Clearly, a fish cannot have zero weight. Since its other attributes seem fine, I don't think removing the observation is necessary. A better option would be to interpolate for weight given that the Species type is a Roach. Weight has a similar relationship to all other variables, so ordering by one of these and using it as a basis for linear interpolation should be fine. Even though we currently see that the relationship is not quite linear, interpolating in this way should be fine and not significantly impact the model results.

```
roach_data <- data[data$Species == 'Roach', ]
rownames(roach_data) <- 1:nrow(roach_data)
zero_index <- as.numeric(rownames(roach_data[roach_data$Weight == 0, ]))
pred_val <- roach_data[c(zero_index), 3]

roach_data <- roach_data[-c(zero_index),]
roach_lm <- lm(Weight ~ Length1, data = roach_data)

inter_value <- predict(roach_lm, data.frame(Length1 = pred_val))
data[data$Weight == 0, 2] <- inter_value
```

From the descriptive analysis it was also evident that the data is in different units. To make the regression analysis simpler a unit normal scaling can be performed.

```
scaled_data <- as.data.frame(scale(data[, -1], center = TRUE, scale = TRUE))
```

This is also a good time to setup our indicator variable species.

```
indicator_data <- fastDummies::dummy_cols(data['Species'], select_columns = "Species")
indicator_data$Species <- NULL
```

Combine the scaled and indicator data.

```
data <- data.frame(indicator_data, scaled_data)
```

Unique Data Point Introduction

Output the new dataframe as a .csv so it can be used in the regression analysis. It may not be necessary to use Species in the analysis of residuals and checking for collinearity, so scaled_data will also be saved.

```
write.csv(data, 'cleaned_data_2.csv', row.names = FALSE)
write.csv(scaled_data, 'cleaned_data_scaled_only_2.csv', row.names = FALSE)
```