Fishy Business

Kankshat Patel

6/9/2022

Abstract

This analysis is motivated by my interest in aquatic life and a hope to gain a better understanding of the causation of weight fluctuations of numerous fish species off the coast of China. The data examined is composed of two categorical and two numerical variables, all of which are randomly sampled. The data was assessed by performing a two-way ANOVA test using the categorical variables, species and width. Moreover, a multivariate regression was utilized to determine the predictability of quantitative variable change in weight of fish via the use of the other, change in height and lenght of fish. It was found that height and lenght have a significant effect of the weight of a fish and can be used to predict the change in weight. The results can explain the variability in sizes of all axes and how the one common alteration is weight.

Introduction

This dataset was gathered by researchers at Asia Pacific University (KL, Malaysia) Mandalay, and Mandalay Region in conjunction with Myanmar (Burma). This dataset focuses on fish attained within the Chinese sea and taken to nearby fish markets where each fish was measured and weighed. The data was generated by labeling and taking 159 measurements of each fish and determining which species the fish belongs to. Knowing the height, width, length, and species is vital to understanding how these factors can potentially affect or predict the weight of a fish. Though all 7 species performed well in SW (Saltwater), the study also revealed that farming of these species can be feasible if water quality parameters are properly monitored.

The goal for this analysis is to evaluate the MCR LTER dataset in order to investigate which independent variables or interactions between independent variables have a significant effect on the weight of fish, and which, if any, can be used to predict it. I hypothesize that species type and width will have a significant effect on the weight, but the interaction between species type and width will not. I also hypothesize that height and length variables wil be able to predict the weight of a fish regardless of species.

Exploratory Analysis

To begin understanding our data deeper, we must graph the data meaningfully. I created a scatterplot that compares the height of any given fish to their respective weight. The scatterplot indicates that there seems to be an increasing relationship between height and weight for all species fish. As the height of fish increase, the weight of the fish increase as well. However, there is this height bracket of 0cm - 5cm where the weight does not increase. In addition, there are a few outliers where the height is around 10cm where the weight is far heavier than the overall trend.

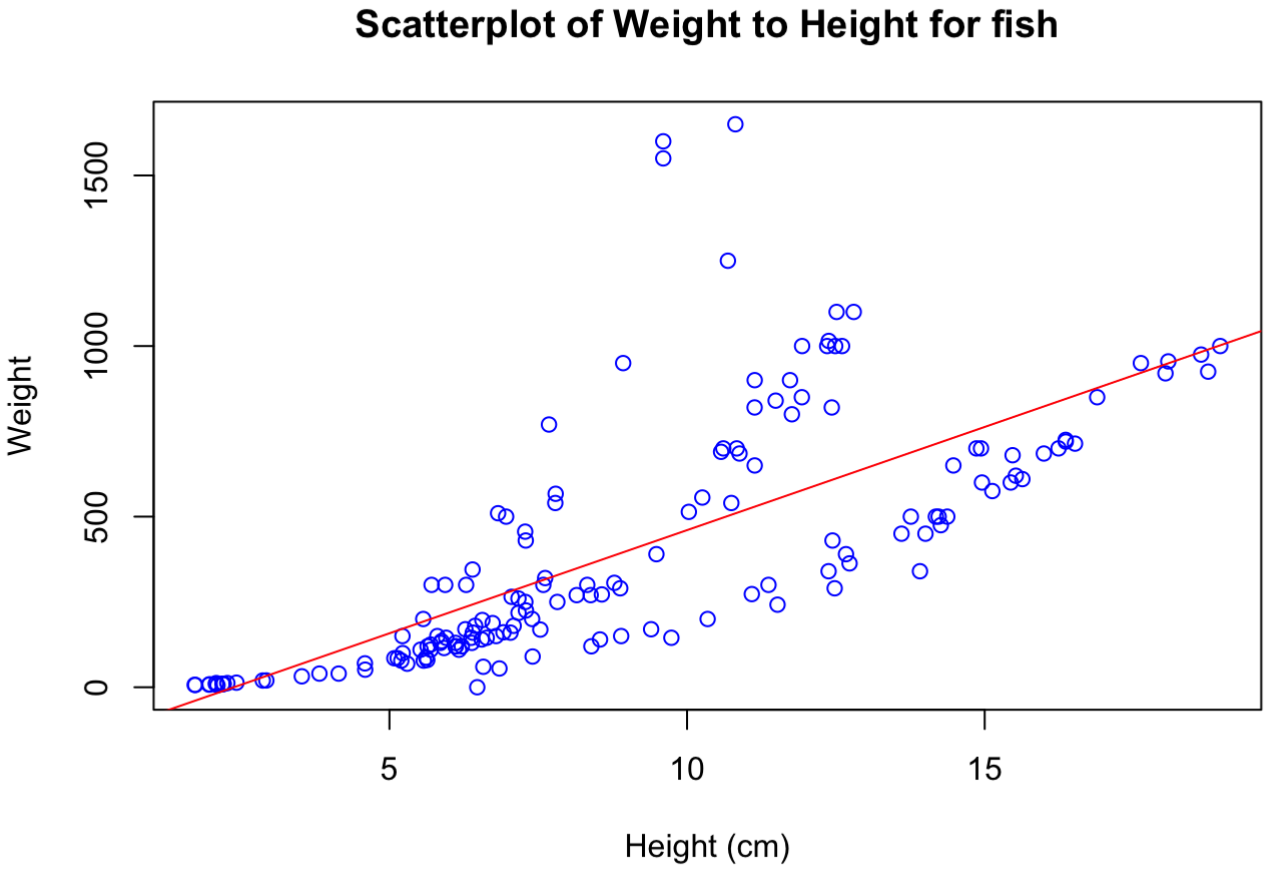


Figure 1: A scatterplot to indicate the relationship between the heights of fish and the weights. The graph indicates that there is an increasing relationship between the two variables.

Histograms

A histogram was used to explore the distribution of the heights of fish of all species. Upon viewing the histogram, the data exhibits a right-leaning skew. However, upon further analysis it was found that attempts to normalize the data via log and square root transformations increased the severity of the skew. This indicates the data is most normal when untampered. In addition, due to the large sample size of data points, it satisfies the criterion of the Central Limit Theorem and thus can be considered approximately normal.

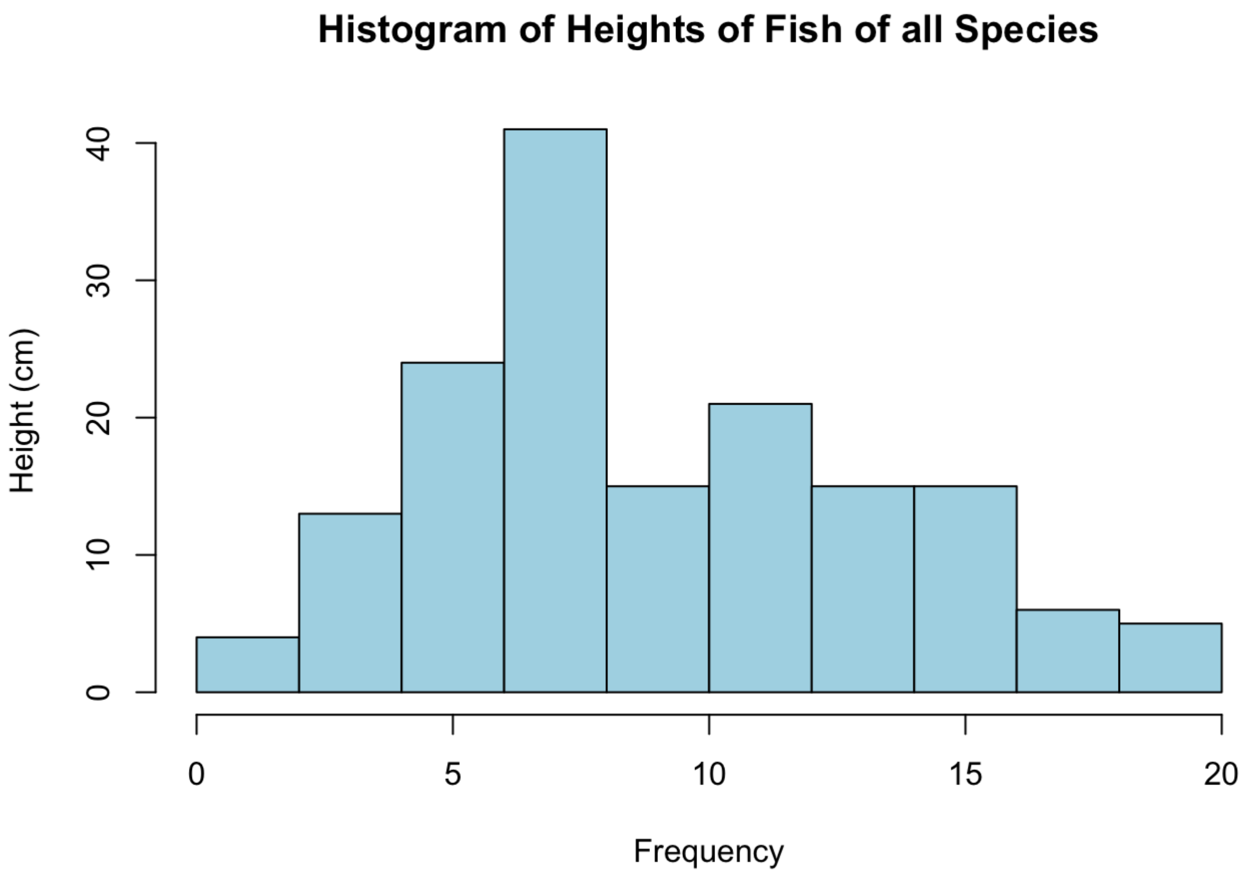


Figure 2. A histogram to determine the distribution of the heights of fish with no specific species. The histogram displays a slight right skew of the data.

A Histogram was used to explore the distribution of lengths of all species of fish. Upon examination of the histogram, it can be seen that the data is approximately normal with a slight right skew. My attempt to normalize the data led me to perform square root and log transformation to find that both transformations further skewed our data, indicating the data is most normal when non-transformed. Similarly to the previous numerical variable, the sample size is 159, meeting the criteria for the central limit theorem, thus we can treat the data as approximately normal.

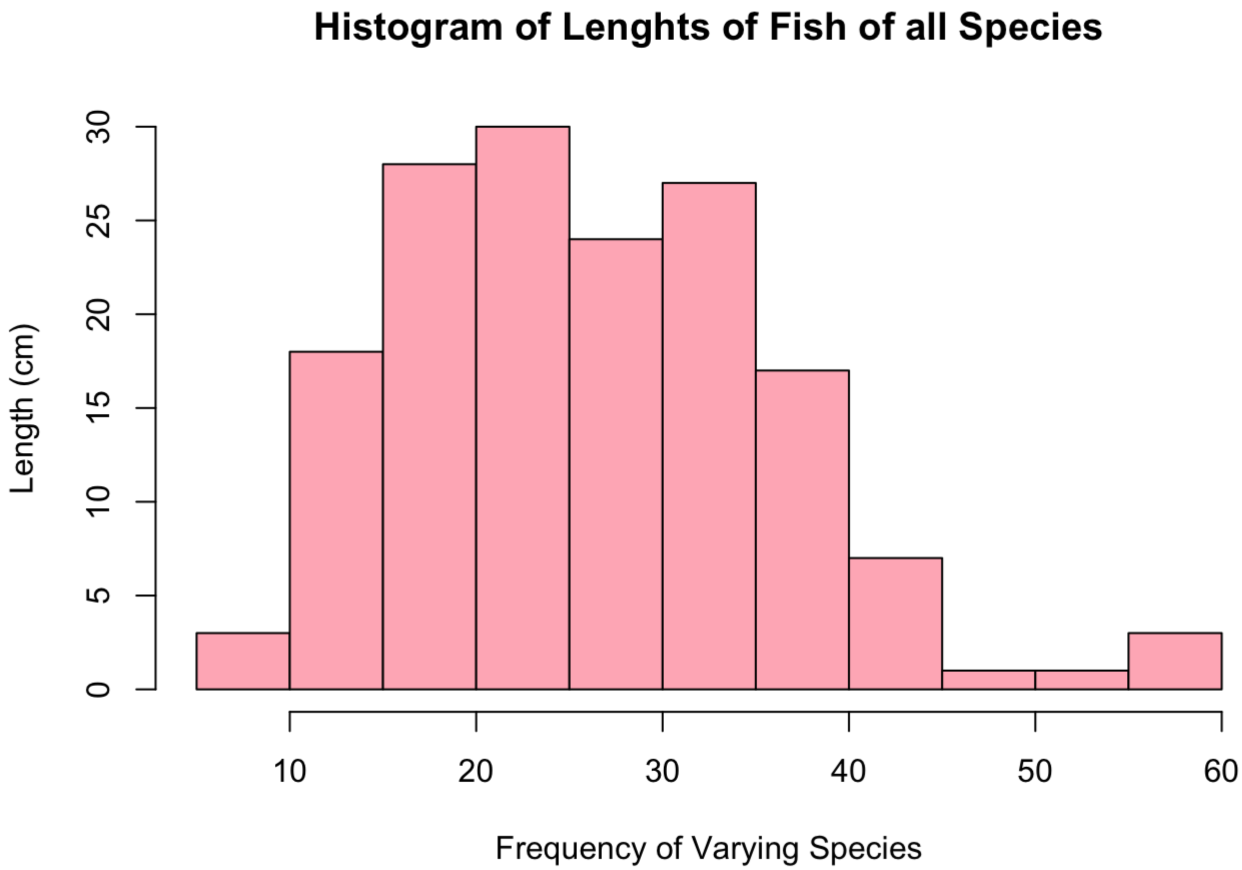


Figure 3. A histogram to determine the distribution of the lengths of fish with no specific species. The histogram displays a slight right skew of the data.

To graph the categorical data meaningfully, I created a boxplot that compares the weights of fish with varying widths. The widths have been coerced into three brackets; low (0-3.7), medium (3.8-5.1), and high (5.2-8.2). The boxplot shows what our intuition suggests: the higher the width, the larger the weight of the fish. In addition, our boxplot displays a fwe outliers with high height and high weight.

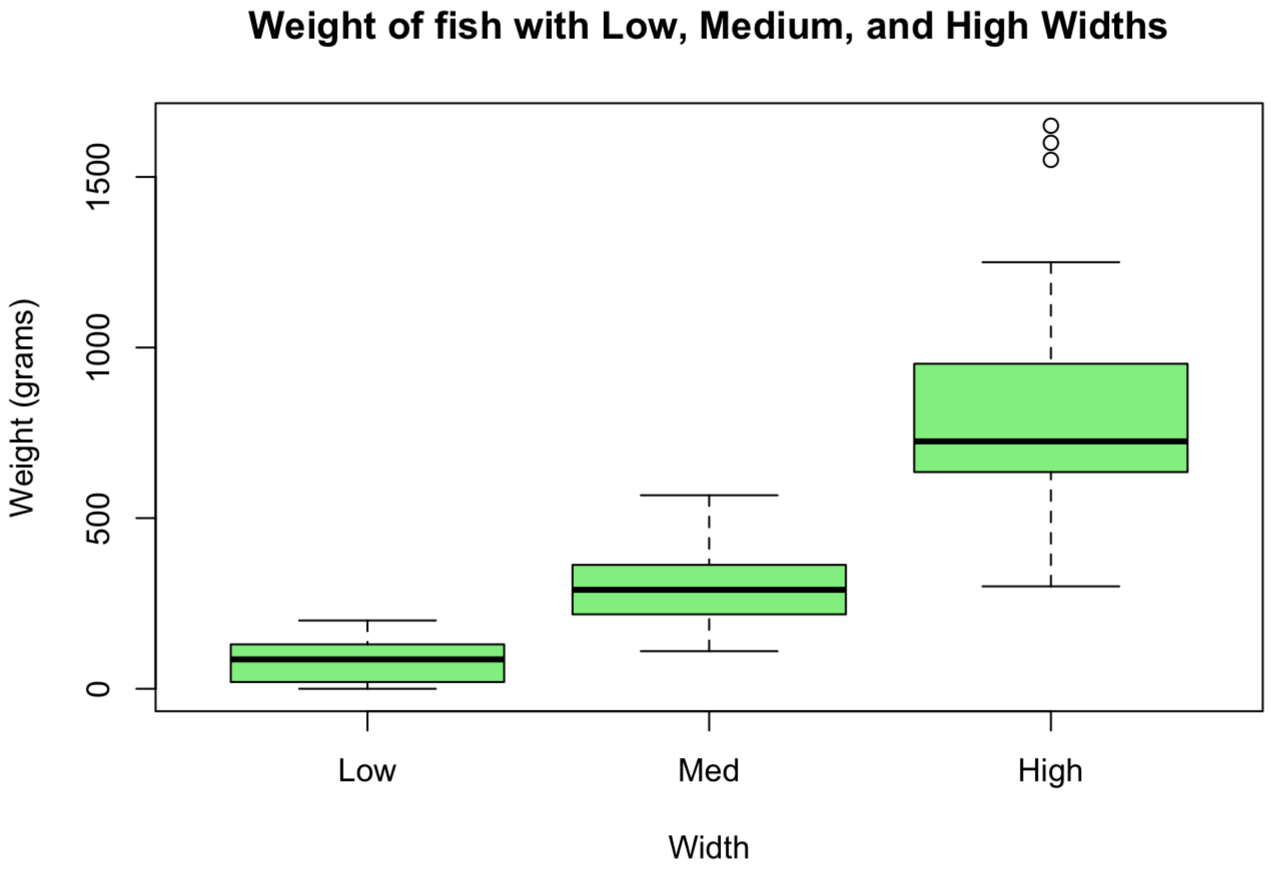


Figure 4. This boxplot is used to determine the relationship between the numerical variable, weights of fish, to the categorical variable, widths of fish. From the boxplot, we can deduce that that as the widths of fish increase, the weight increases as well.

To analyze the difference in weights of varying species of fish, I used a boxplot and found that Smelt, Parkki, and Roach species of fishes have consistently lighter fishes than all other species. The species Smelt and Roach are the only two species with outliers, neither of which are significant. In addition, I found that the data indicates Pike and Bream fishes have the largest mean weight while Pike fishes are exhibited to have a larger range of weight. In addition, it can be inferred that Whitefish follow a similar trend to Pike but at a lower weight. Lastly, the data tells us that Perch fishes have a large range of weights, but a relatively low mean.

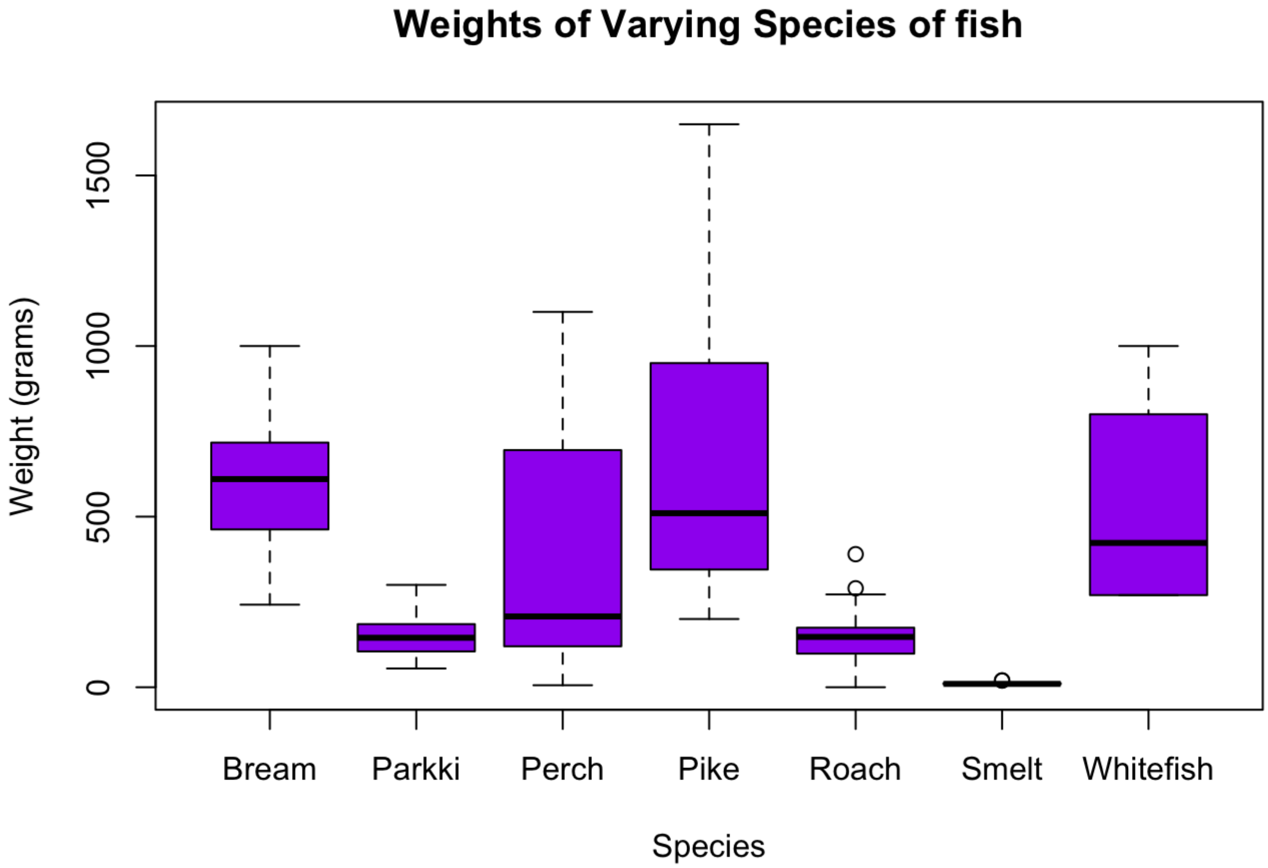


Figure 5. This boxplot shows the difference weights of 7 different species of fish. It can be seen that few species such as Perch and Pike have large variation in weights while other species, like Roach and Smelt, have very frigid variation in weight.

Statistical Methods

ANOVA

An ANOVA test is a statistical test that assesses how a quantitative variable changes according to categorical variables. Running an anova test will allow me to analyze how weight changes according to species and widths of fishes. For this ANOVA test, weight is the dependent variable, and species and width are the two independent variables. The Species variable has 7 levels, each a different species, while the width variable has three levels, low, medium, and high.

The ANOVA test has three assumptions. The ANOVA test assumes homogeneity of variance, normality of residuals, and random sampling. Variance is defined by the difference between the collected datapoints and the mean of the dataset and can be assessed by graphing the residuals. Residuals are defined by the difference between the experimental and theoretical data. The theoretical data is represented by a red trendline that runs through the residual graph.

The homogeneity of variance assumption is tested by the “Residuals vs. Fitted” plot, and the assumption is shown to be met by the datapoints on the graph being of relatively equal breadth about the 0 line, as well as the residuals having no distinct pattern about the 0 line, which provides evidence for the equality of variances.

The normality of residuals is tested and shown to be met by the “Normal Q-Q” plot, as when looking at the plot, it can be seen that the datapoints adhere closely to the trendline. There are several outliers, but given the general adherence to the trendline and large dataset, the residuals can be assumed to be normal.

The random sampling assumption is met because all datapoints were randomly sampled from a fish market in China.

The null hypothesis states that the p-value is greater than 0.05 (p>0.05), which means that the independent variable or interaction of species and width does not have a significant effect on the change in weight. The alternative hypothesis states that the p-value is less than 0.05 (p<0.05), which means that the independent variable or interaction does have a significant effect on the change in weight.

Multivariate Regression

A Multivariate regression is the method of modeling multiple responses, or dependent variables, with a single set of predictor variables. In this study, performing a Multivariate regression using my dataset allows me to understand whether or not height and length are able to predict the weights of fishes.

The Multivariate regression model has three assumptions. It assumes that the data come from a random sample, the Y variable is normally distributed with equal variance for all values of X, and the residuals are normal. The random sampling assumption is met because the data come from a random sample taken from a fish market in China.

The equality of variance assumption is tested by plotting the residuals and assessing the “Residuals vs. Fitted” plot. The data appears slightly heteroskedastic. However, running transformations to attempt to normalize the data only increased the heteroskedaticity, so the most homoskedastic model is the untransformed one. Due to this, as well as the very large, random sample size, we can still consider the assumption met with the acknowledgment of the slight heteroskedasticity.

The normality of the residuals is tested and shown to be met by the “Normal Q-Q” plot, as when looking at the plot, it can be seen that the datapoints adhere closely to the trendline. There are several outliers, but given the general high adherence to the trendline and large dataset, the residuals can be assumed to be normal.

The null hypothesis for this linear regression states that the p-value is greater than (p>0.05), which means that the change in weight cannot be predicted by the change in height and weight. The alternative hypothesis states that the p-value is less than 0.05 (p<0.05), which means that the change in weight can be predicted by the change in height and weight.

Results

ANOVA

The two-way ANOVA test yields three p-values. Every p-value is compared to an alpha level of 0.05. P-values below the alpha level result in a rejection of the null hypothesis, and p-values above the alpha level result in a failure to reject the null hypothesis.

The ANOVA assessment of weight of fish according to the width variable yields a p-value of 2e-16. Due to the small nature of the p-value, I reject the null hypothesis and conclude that width does have a significant effect on the weight of a fish.

The ANOVA assessment of weight of fish according to the species variable yields a p-value of 2.11e-13. Due to the small nature of the p-value, I reject the null hypothesis and conclude that the species of a fish does have a significant effect on the change in weight.

The ANOVA assessment of weight of fish according to the interaction between species and width yields a p-value of 0.058. Due to the large nature of this p-value, I fail to reject the null hypothesis and conclude that the interaction between width and height does not have a significant effect on the weight of fish.

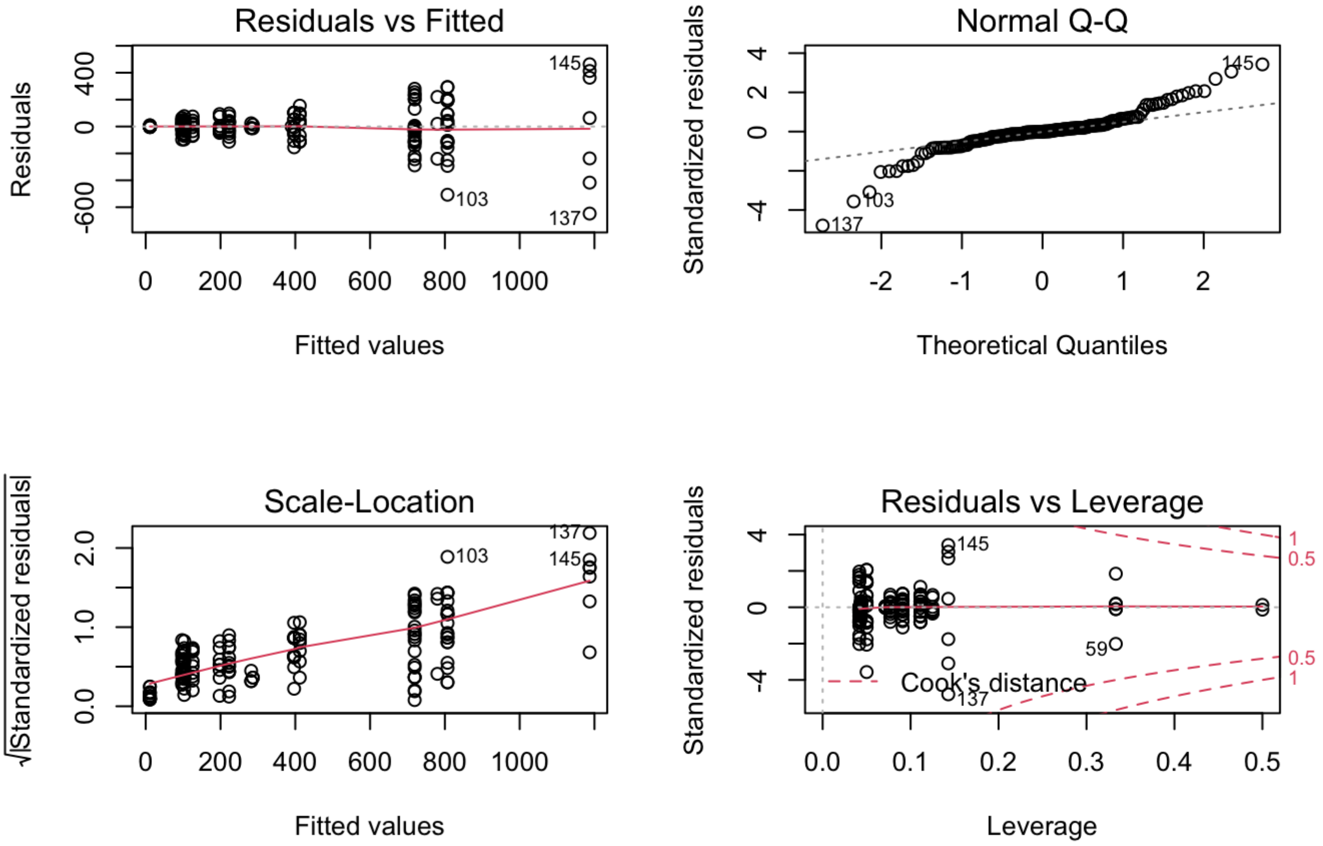


Figure 6. Residual plots for the ANOVA test. The residuals vs. fitted plot exhibits the homoskedasicity required to assume equality of variances due to the breadth and distribution of the data points. The normal Q-Q plot exhibits the normality required to assume normality of residuals via the close alignment of the datapoints to the generated trendline.

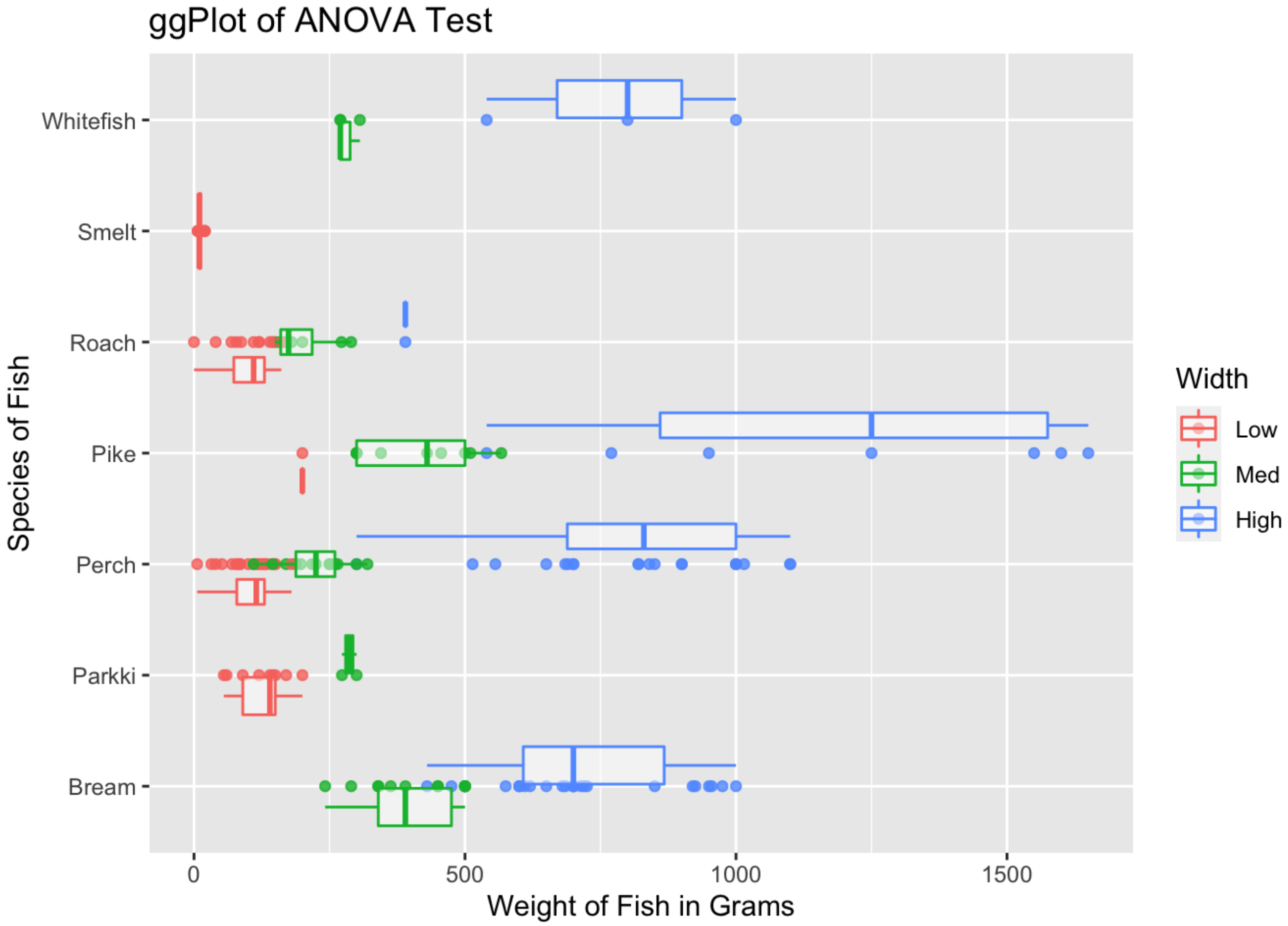


Figure 7. A ggPlot visualizing the results of the ANOVA test. The plot shows that lower widths of fish corresponds to a lower weight in fish. In addition, it can be seen that certain species of fish correlate to certain weights, indicating specific species of fish are genetically predisposed to grow to a certain amount.

Multivariate Regression

The Multivariate regression yields a p-value for the relationship between weight of fish and the interaction term of length and height of fish. The p-value is compared to an alpha level of 0.05. P-values below the alpha level result in a rejection of the null hypothesis, and p-values above the alpha level result in a failure to reject the null hypothesis.

The Multivariate regression model for weight a p-value of 1.21e-10. Due to the small nature of the p-value, I reject the null hypothesis and conclude that the interaction of height and lenght can be used to predict the weight of numerous fishes.

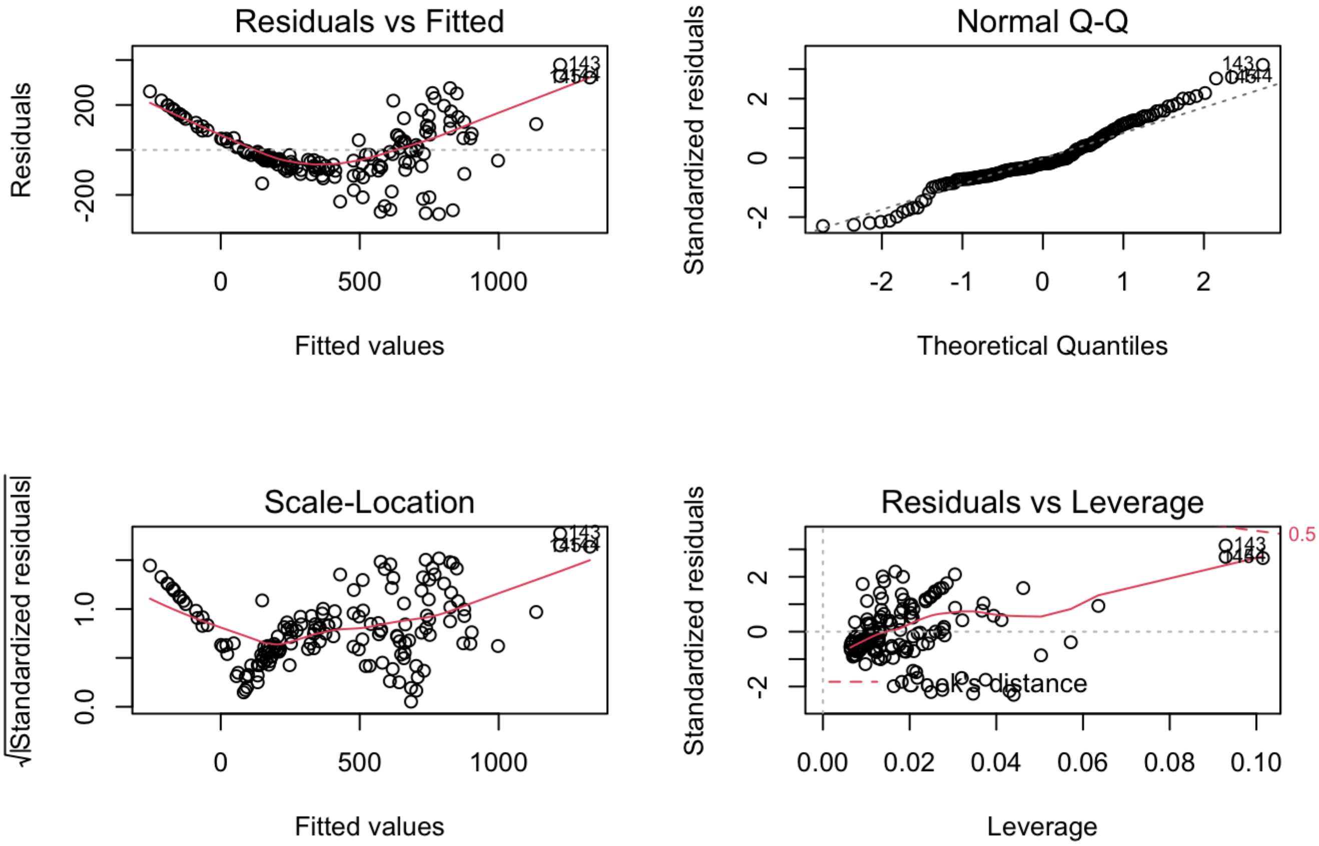


Figure 8. Residual plots for the linear regression. The residuals vs. fitted plot exhibits slight heteroskedasicity due to the slight cone shape of the distribution of data points. The data were not made more homoskedastic by transformations, and the dataset is large, so the data can be considered normal despite the minor anomaly. The normal Q-Q plot exhibits the normality required to assume normality of residuals via the close alignment of the datapoints to the generated trendline.

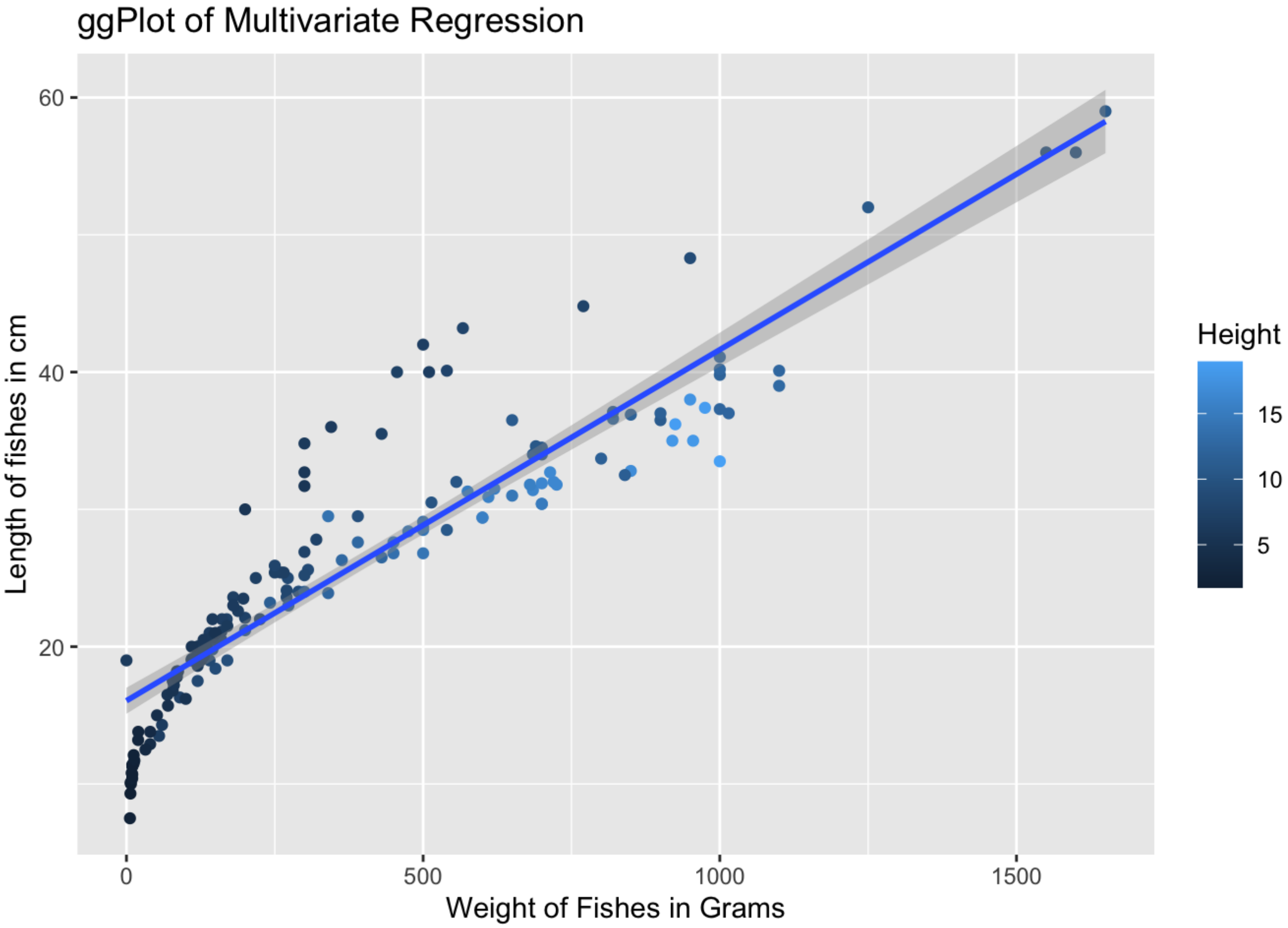


Figure 9. A ggPlot visualizing the results of the multivariate regression. The plot shows that the datapoints and trendline for weight very closely align with the datapoints and trendline for height and lenght, which evidences the idea that height and length are both good predictors in determine the weights of fish.

Discussion

The results of the ANOVA test demonstrate that the width of fishes as well as the species of fishes both contribute significantly to the weight of fishes. Naturally, this supports the idea that the larger the length and height of the frame of the fish is, the more body mass there is. There is a need for more scales, larger organs, larger bones, and many more body parts. A fishes growth be explained as either allometric versus isometric growth patterns and differs in body shapes of per respective species (Froese 2006). Isometric growth is associated with no change of body shape as an organism grows. Negative allometric growth implies the fish becomes more slender as it increase in weight while positive allometric growth implies the fish becomes relatively stouter or deeper-bodied as it increases in length (Riedel et al. 2007).

The ANOVA test for the interaction between Species and Width as it affects weight of fish was insignificant. However, this is expected, as in the natural world the interaction of the two variables has no physical effect on the weight of a fish. Rather the two individual factors contribute greatly to the weight of a fish.

The ANOVA test to determine the statistical significance between the species of fish and weight was significant. As expected, varying species of fish are genetically limited to how wide, tall, and heavy they can grow to. A fish weight can be limited by what species they fall into, thus the species is a significant factor in determining the weight of a fish in that given species.

The results of the multivariate regression demonstrate that the change in height and length can be used to predict the weight of a fish. This supports the idea that as the height of a fish increases, weight increases as well, but stagnant or decreasing width could mean no, negative, or positive change in weight. Naturally, as height or length of any organism increases, the internal aspect of that organism must fill the exterior shell. As nature takes its course and increases these factors of a fish, it is perfectly explainable that the weight of any fish will grow with its length and height. Subsequently, we could expect to see that the lack of change or negative change in height or width would predict and justify the same change in weight.

Limitations to this analysis are present. Given the varying conditions of the Chinese fish market, the lack of variety of fish may skew the conclusions as these findings may not apply to other species of fish native to other locations of the world. The conditions of the waters are vital to the growth of the fish as well. Freshwater and saltwater produce differing environments for growth because freshwater has low buffering effect, Carbon dioxide can accumulate in the water, thus creating a toxic aquatic locale. (Tucker et al., 1984). A lake in Michigan may yield completely different results than a lake or ocean in China. Given more time and data, I would run similar assessments on fish in different parts of the world and examine an additional variable, diet, competition, and get to allow the results to be more broadly applicable to worldwide species of fish.

The final takeaway presented by this study is that the weight of fish within a Chinese market can be influenced by the height, Width, and particular species of fish. The overall correlation of the variables can be used to investigate other organisms to determine whether the results are consistent with all living organisms.

References

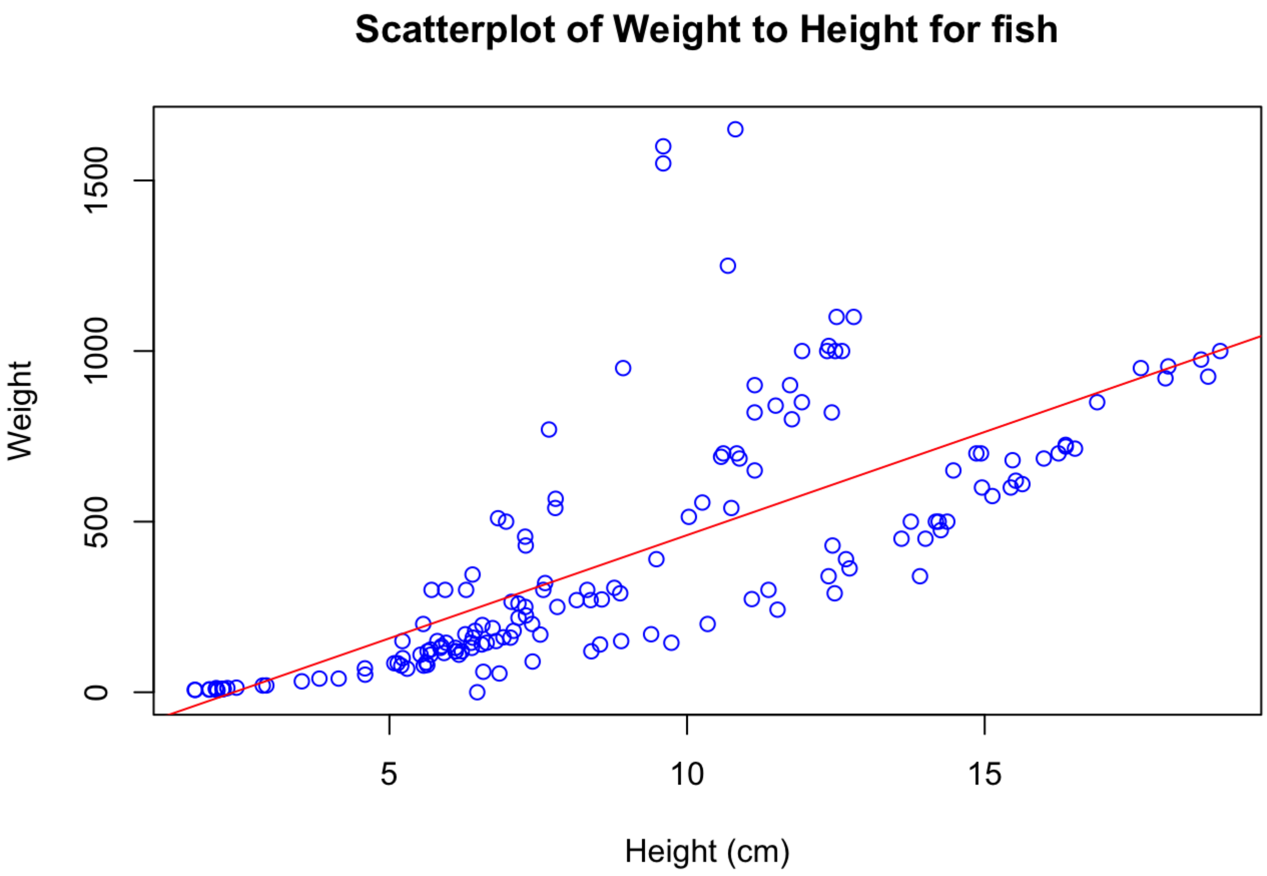
Froese R. Cube law, condition factor and weight–length relationships: history, meta-analysis and recommendations. J Appl Ichthyol 2006; 22: 241-253.

Riedel, R., Caskey, L.M., Hurlbert, S.H (2007). Lengthweight relations and growth rates of dominant fishes of the Salton Sea: implications for predation by fish-eating birds. Lake and Reservoir Management. 23:528-535.

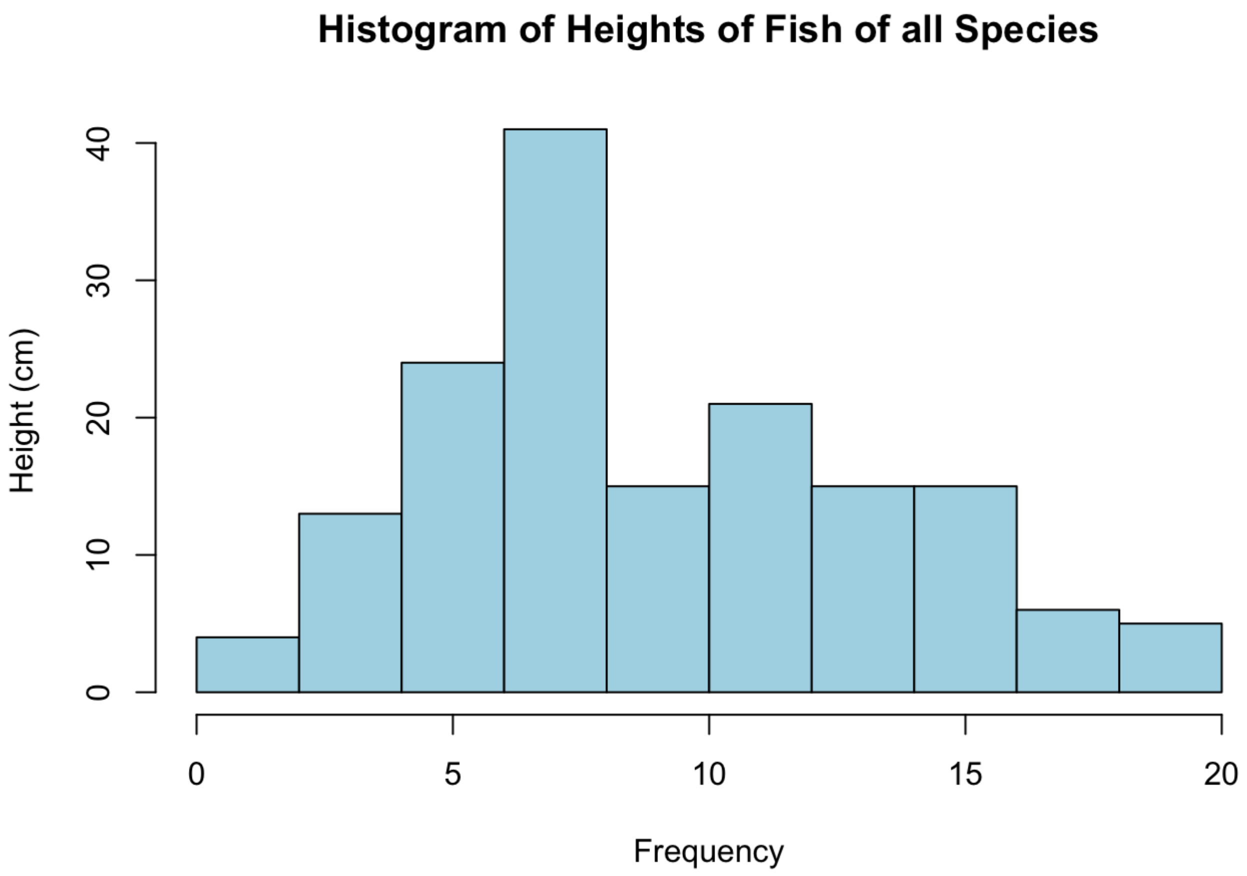
Tucker, C. S., Lloyd, S. W., Busch, R. L(1984). Relationship between phytoplankton periodicity and the concentrations of total and un-ionized ammonia in channel catfish ponds. Hydrobiologia. 111:75–79.

Appendix

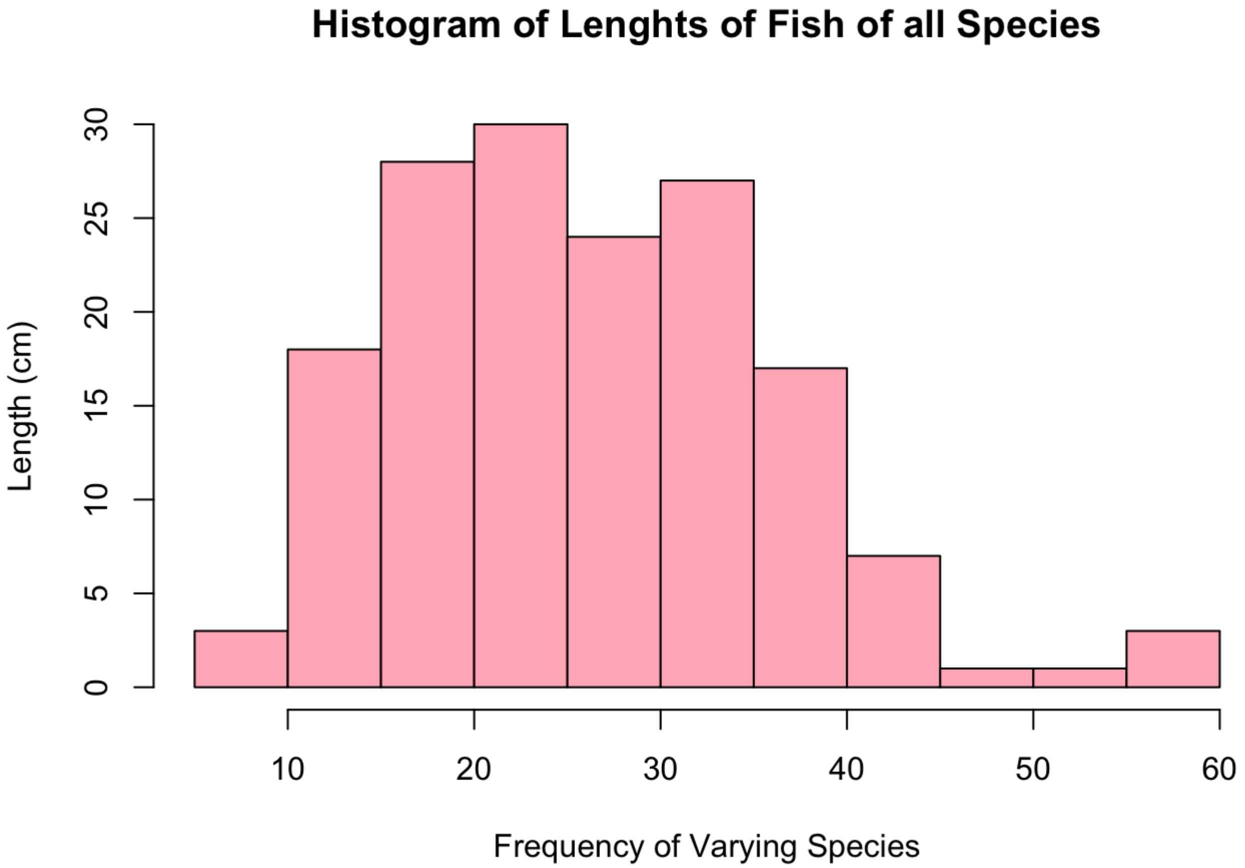
```
# Load Fish Data
fish <- read_csv("/Users/kankshatpatel/kjp/UCSB/COURSES/Spring2022/EEMB 146/project/fis
h.csv")

# Explore the data
plot(fish$Height, fish$Weight,
     main = "Scatterplot of Weight to Height for fish",
     xlab = "Height (cm)",
     ylab = "Weight",
     col = "blue")
abline(lm(fish$Weight ~ fish$Height), col = 'red')
```



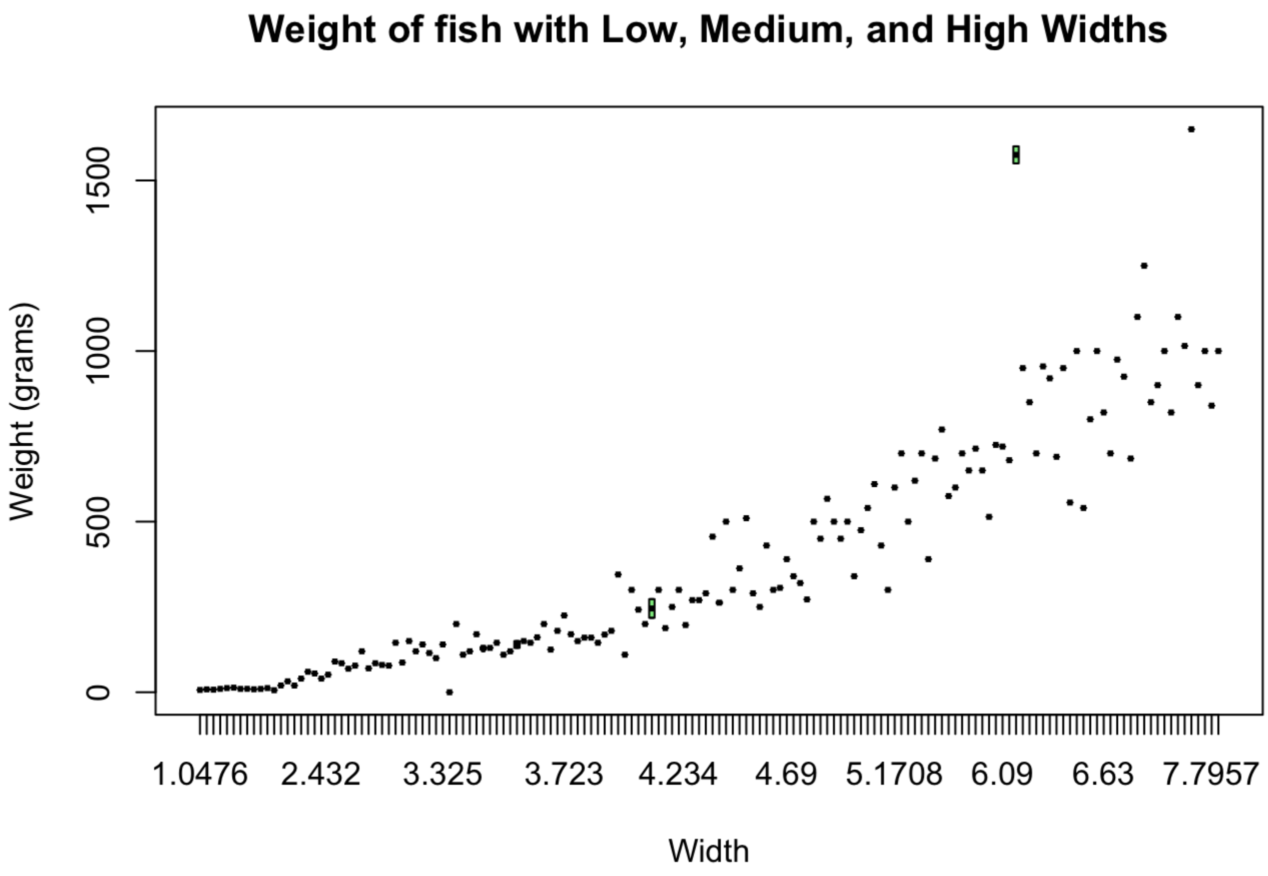
```
hist(fish$Height,
     main = "Histogram of Heights of Fish of all Species",
     xlab = "Frequency",
     ylab = "Height (cm)",
     col = "light blue")
```



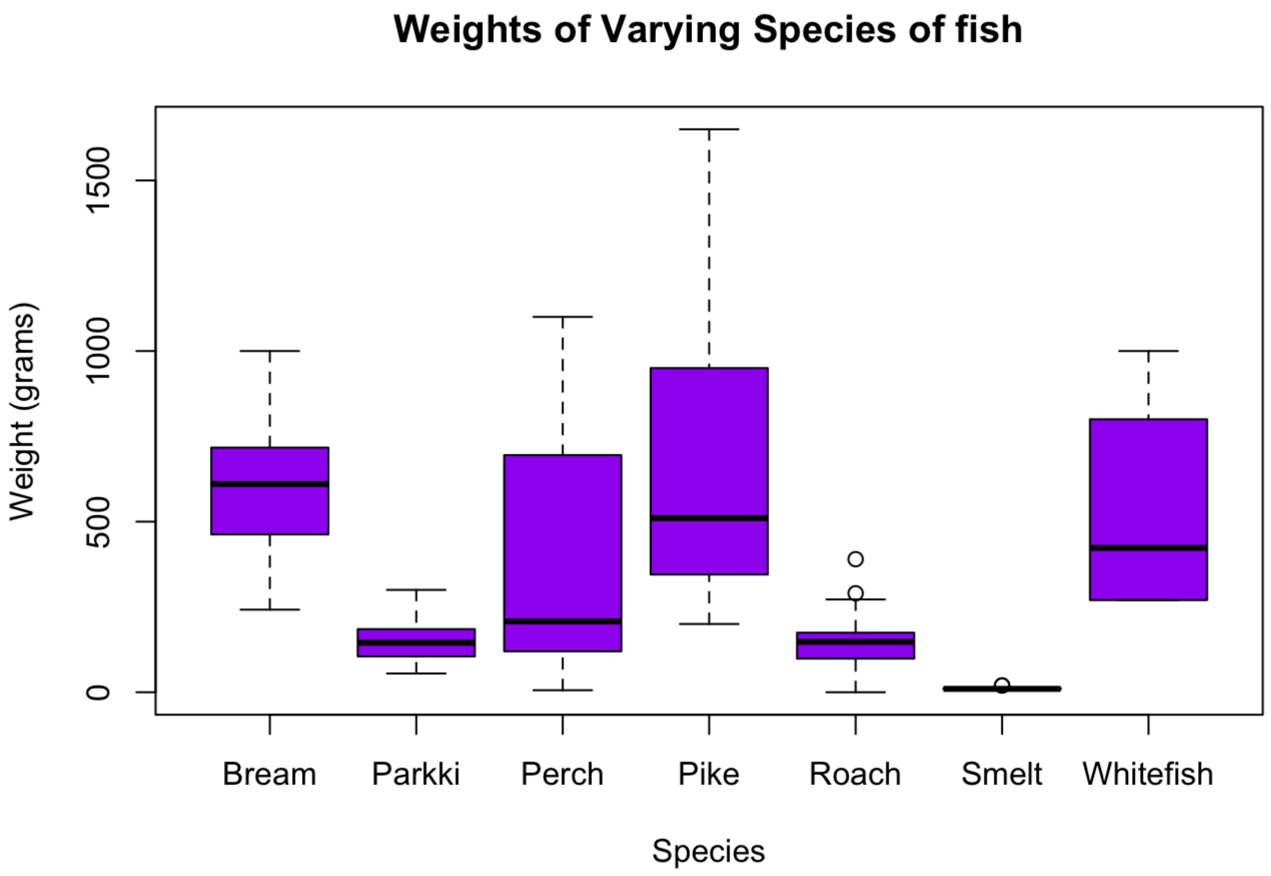
```
hist(fish$Length1,
     main = "Histogram of Lenghts of Fish of all Species",
     xlab = "Frequency of Varying Species",
     ylab = "Length (cm)",
     col = "light pink")
```



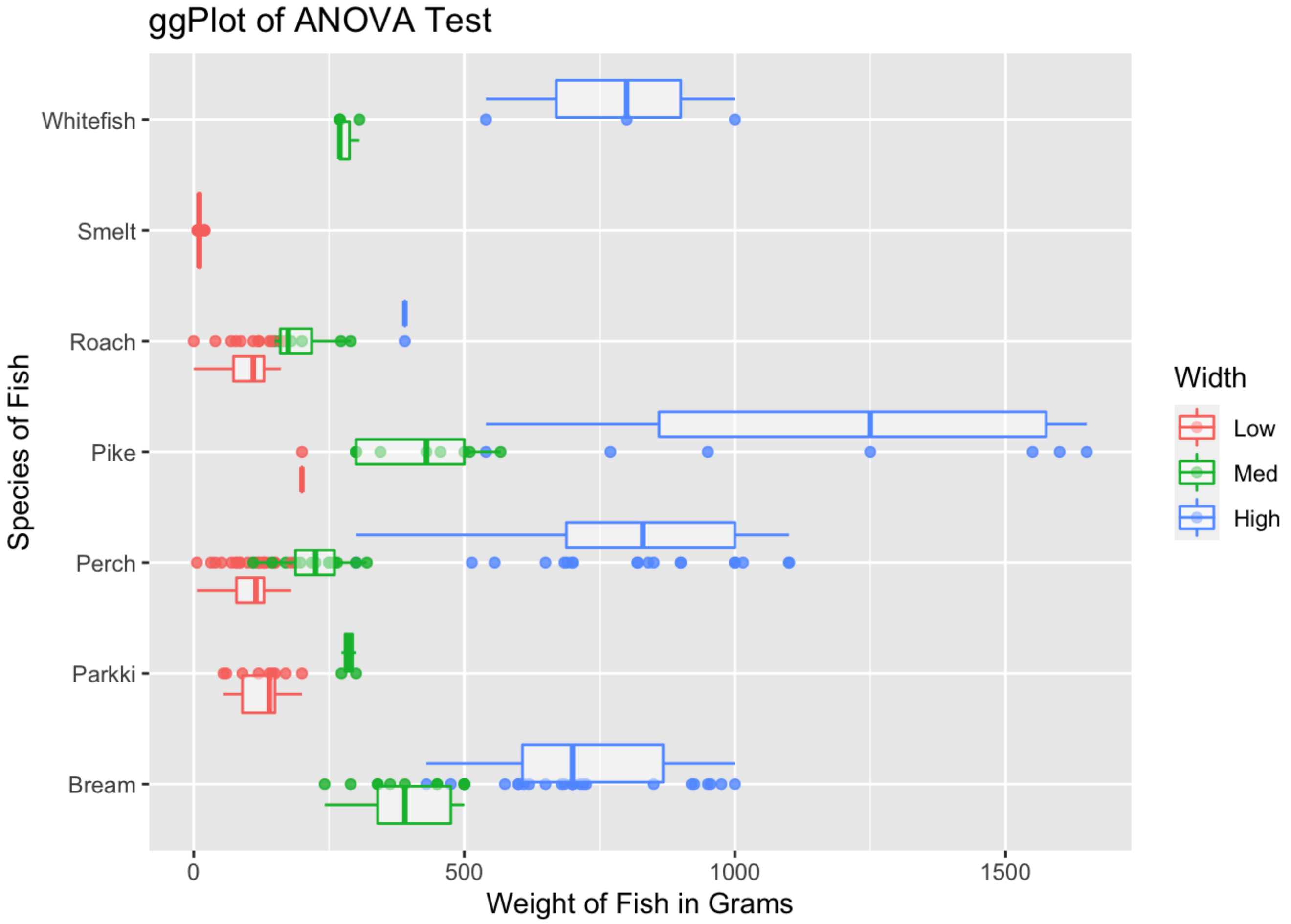
```
boxplot(Weight~Width,
        data=fish,
        xlab= "Width",
        ylab= "Weight (grams)",
        col = "light green",
        main = "Weight of fish with Low, Medium, and High Widths")
```



```
boxplot(Weight~Species,
        data=fish,
        xlab= "Species",
        ylab= "Weight (grams)",
        col = "purple",
        main = "Weights of Varying Species of fish")
```



```
#Graphically Represent ANOVA Results
ggplot(f1, aes(x=Weight, y=Species, color=Width))+
geom_point(alpha=0.8)+
geom_boxplot(alpha=0.5)+
labs(title="ggPlot of ANOVA Test", x="Weight of Fish in Grams", y="Species of Fish")
```

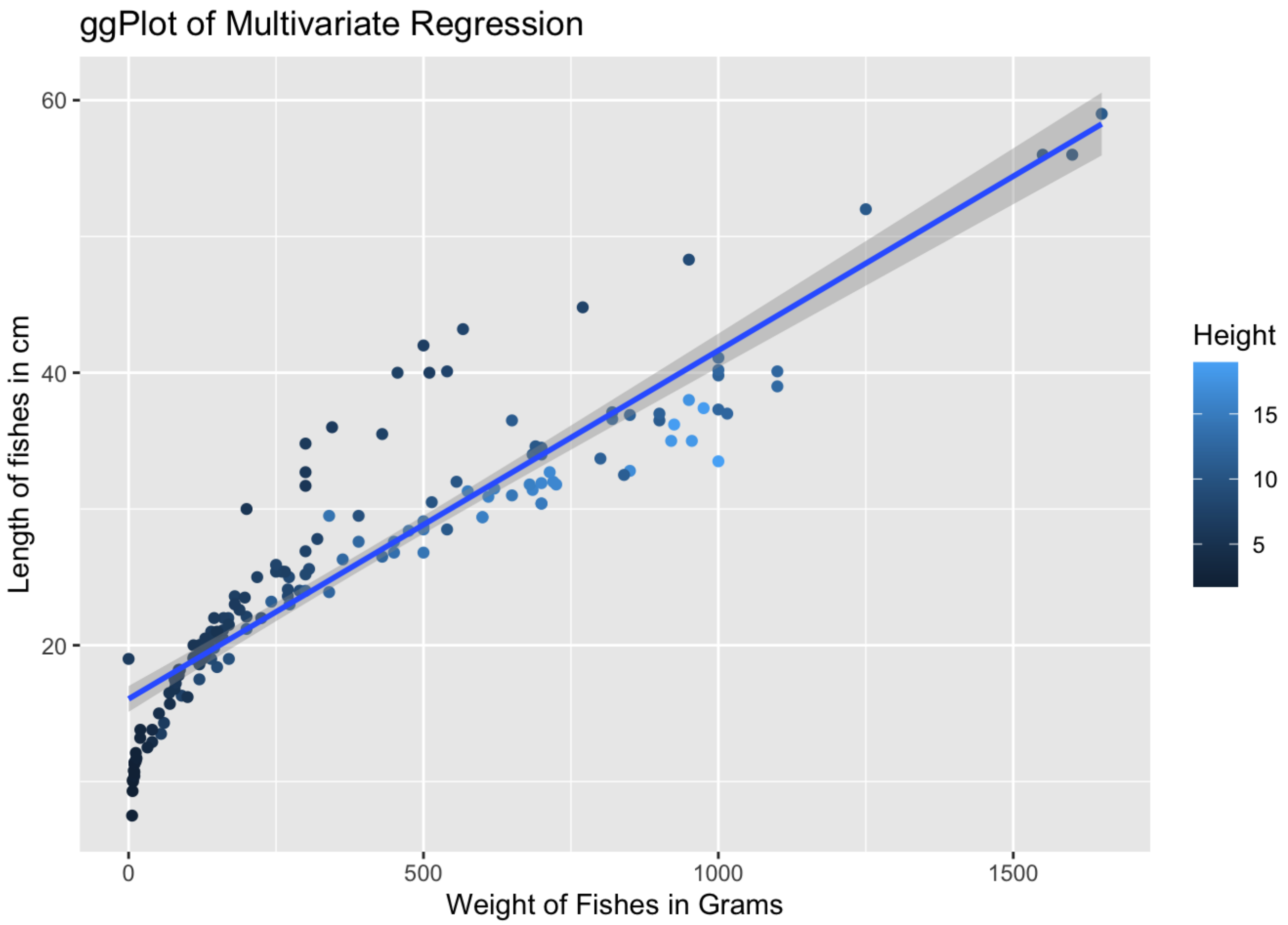


```
summary(mod)
```

```
##
## Call:
## lm(formula = Weight ~ Height + Length1, data = f1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -285.18  -76.00  -23.33   70.02  378.91
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -502.589     28.903 -17.389  < 2e-16 ***
## Height        20.805      3.014   6.903 1.21e-10 ***
## Length1       27.213      1.292  21.059  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 126.7 on 156 degrees of freedom
## Multiple R-squared:  0.8763, Adjusted R-squared:  0.8747
## F-statistic: 552.6 on 2 and 156 DF,  p-value: < 2.2e-16
```

```
#Graphically Represent Linear Regression Results
ggplot(f1, aes(x=Weight, y=Length1, color=Height))+
geom_point()+
geom_smooth(method="lm")+
labs(title="ggPlot of Multivariate Regression", x="Weight of Fishes in Grams", y="Length of fishes in cm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```


ggPlot of Multivariate Regression
Length of fishes in cm
Weight of Fishes in Grams
Height