Multiple Linear Regression on fish Dataset

Utkarsh

2022-09-24

## Introduction

A fish market is a marketplace for selling fish and fish products. It can be dedicated to wholesale trade between fishermen and fish merchants, or to the sale of seafood to individual consumers, or to both. Retail fish markets, a type of wet market, often sell street food as well.

## 1. Aim of the project

Our Aim is to Exploratory Data analysis on the 7 species of fish data for market sale and explain the linear dependence of Weight of fish on the different variables on the basis of Linear Regression.

### Meanings of the Columns:

* Species: Species name of fish
* Weight: Weight of fish in gram
* Length1: Vertical length in cm
* Length2: Diagonal length in cm
* Length3: Cross length in cm
* Height: Height in cm
* Width: Diagonal width in cm

Our dependent variable is ‘Weight’. Independent variables are ‘species’, different lengths, ‘height’ and ‘width’.

I will use independent variables (measurements of the fish) to estimate dependent variable (weight of the fish).

### Steps to follow:

* Download and preliminary analysis of data
* Linear Regression
* Multiple Regression
* Verifying Assumptions of Multiple Linear Regression

## 2. Importing Libraries

The following Libraries would be used

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)  
library(GGally)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

library(MLmetrics)

##   
## Attaching package: 'MLmetrics'

## The following object is masked from 'package:base':  
##   
## Recall

library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

## 3. Importing and cleaning Dataset

# Loading the dataset File  
library(readr)  
fish\_data<- read.csv("C:/Users/utkar/OneDrive/Desktop/22N0063 R/Fish.csv")  
View(fish\_data)  
head(fish\_data)

## Species Weight Length1 Length2 Length3 Height Width  
## 1 Bream 242 23.2 25.4 30.0 11.5200 4.0200  
## 2 Bream 290 24.0 26.3 31.2 12.4800 4.3056  
## 3 Bream 340 23.9 26.5 31.1 12.3778 4.6961  
## 4 Bream 363 26.3 29.0 33.5 12.7300 4.4555  
## 5 Bream 430 26.5 29.0 34.0 12.4440 5.1340  
## 6 Bream 450 26.8 29.7 34.7 13.6024 4.9274

# Removing the Outliers  
fish\_data=fish\_data[-c(142,143,144,145),]  
View(fish\_data)  
nrow(fish\_data)

## [1] 155

### Duplicate and Missing values

anyDuplicated(fish\_data)

## [1] 0

sum(is.na(fish\_data))

## [1] 0

# Since our data doesn't contain any duplicate or na values we are done.

## 4. Exploring the dataset

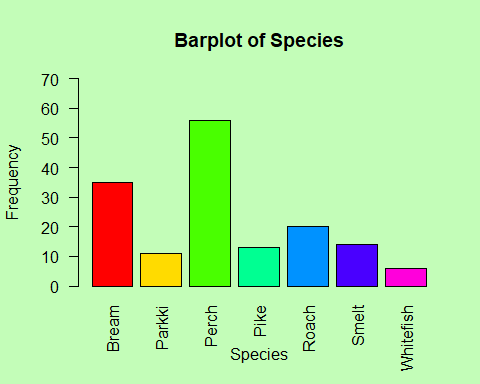
### 4.1 Summary of the dataset

summary(fish\_data)

## Species Weight Length1 Length2   
## Length:155 Min. : 0.0 Min. : 7.50 Min. : 8.4   
## Class :character 1st Qu.: 120.0 1st Qu.:19.00 1st Qu.:21.0   
## Mode :character Median : 270.0 Median :25.00 Median :26.5   
## Mean : 369.6 Mean :25.49 Mean :27.6   
## 3rd Qu.: 605.0 3rd Qu.:32.00 3rd Qu.:35.0   
## Max. :1100.0 Max. :48.30 Max. :51.7   
## Length3 Height Width   
## Min. : 8.80 Min. : 1.728 Min. :1.048   
## 1st Qu.:22.95 1st Qu.: 5.927 1st Qu.:3.364   
## Median :29.20 Median : 7.616 Median :4.239   
## Mean :30.38 Mean : 8.940 Mean :4.359   
## 3rd Qu.:39.35 3rd Qu.:12.379 3rd Qu.:5.373   
## Max. :55.10 Max. :18.957 Max. :8.142

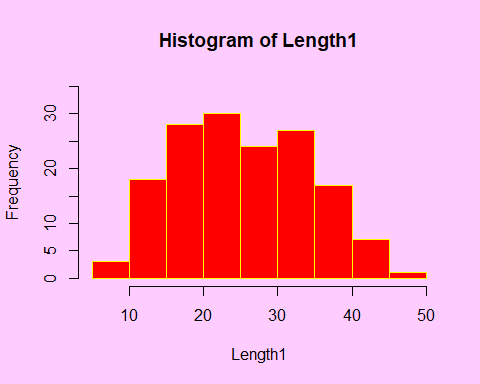
## Barplot of the discrete Variable

Species=fish\_data$Species  
par(bg="#C3FDB8")  
T=table(Species)  
barplot(T,xlab="Species",ylab="Frequency",main = "Barplot of Species",col=rainbow(7),ylim=c(0,70),las=2)

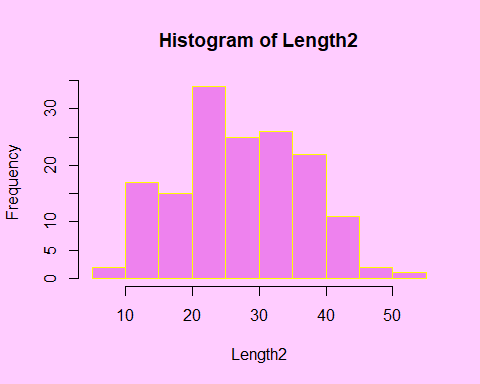


## Histogram and barplot of the Continuous Variable

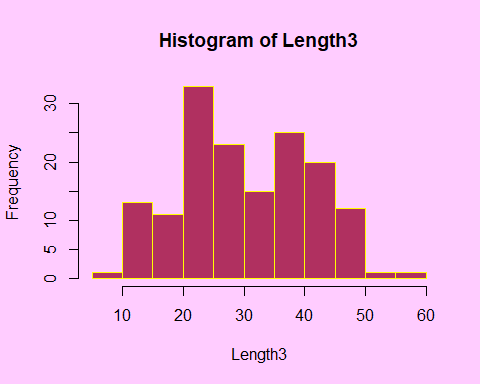
par(bg="#FFCCFF")  
Length1=fish\_data$Length1  
hist(Length1,col="red",border="yellow",ylim=c(0,35))



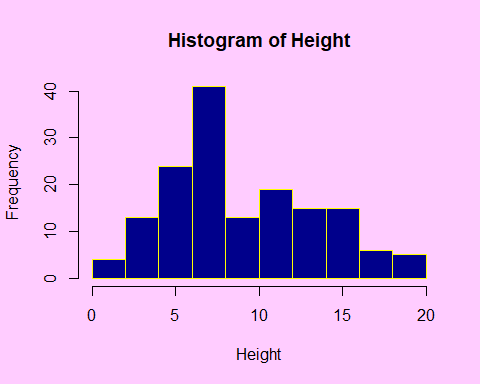
Length2=fish\_data$Length2  
hist(Length2,col="violet",border="yellow")



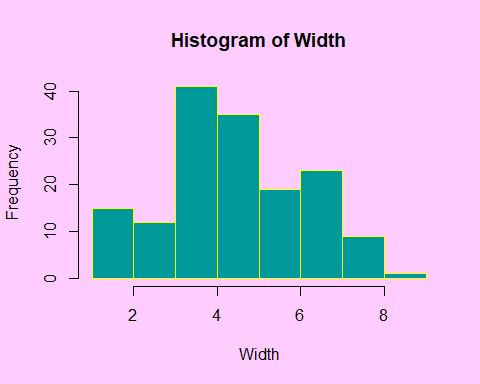
Length3=fish\_data$Length3  
hist(Length3,col="maroon",border="yellow")



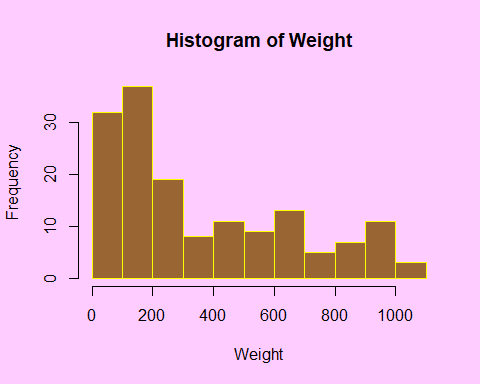
Height=fish\_data$Height  
hist(Height,col="#00008B",border="yellow")



Width=fish\_data$Width  
hist(Width,col="#009999",border="yellow")

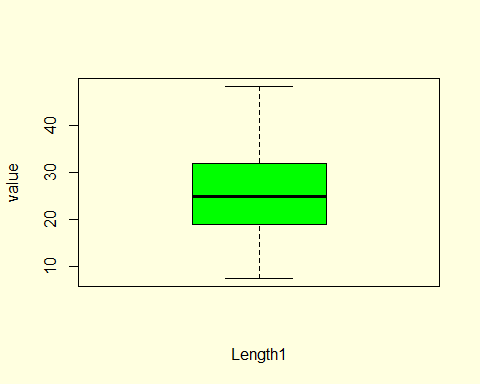


Weight=fish\_data$Weight  
hist(Weight,col="#996633",border="yellow")

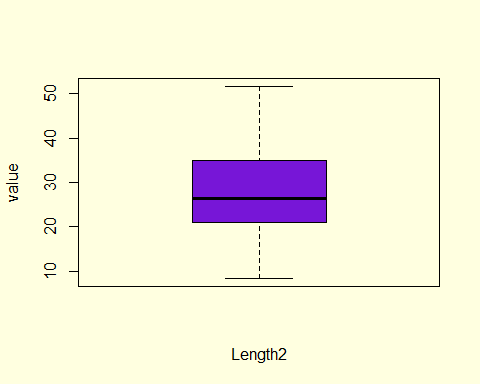


## Boxplot of Continuous Variable

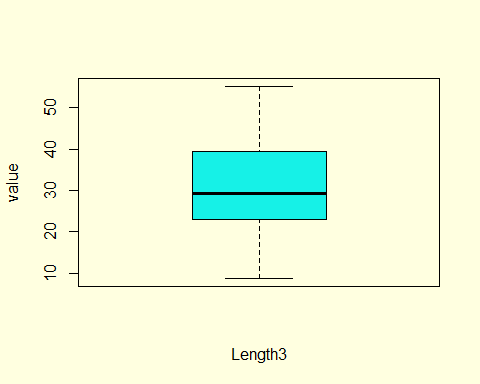
par(bg="#FFFFE0")  
boxplot(Length1,xlab="Length1",ylab="value",col="green")



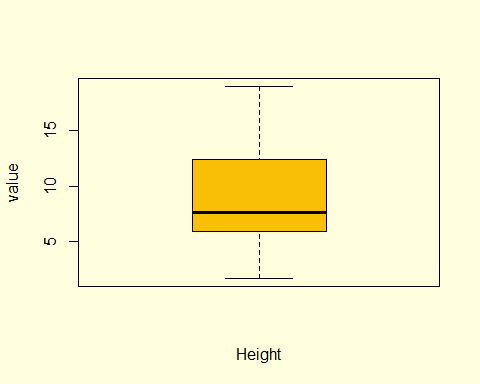
boxplot(Length2,xlab="Length2",ylab="value",col="#7716D7")



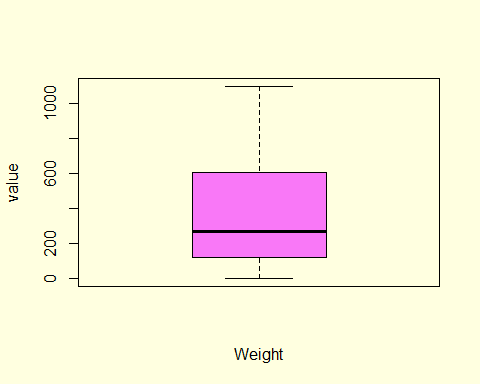
boxplot(Length3,xlab="Length3",ylab="value",col="#16F1E7")



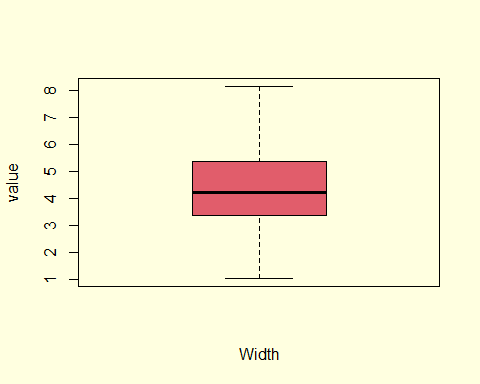
boxplot(Height,xlab="Height",ylab="value",col="#F9BE06")



boxplot(Weight,xlab="Weight",ylab="value",col="#F978F7")



boxplot(Width,xlab="Width",ylab="value",col="#E15D6B")

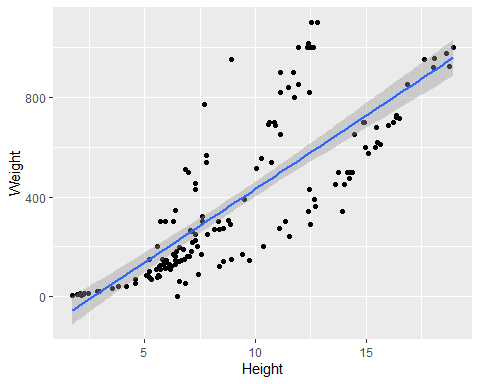


We can See from the Boxplot of the Variables there are no significant outliers in our data

## Scatter Plot

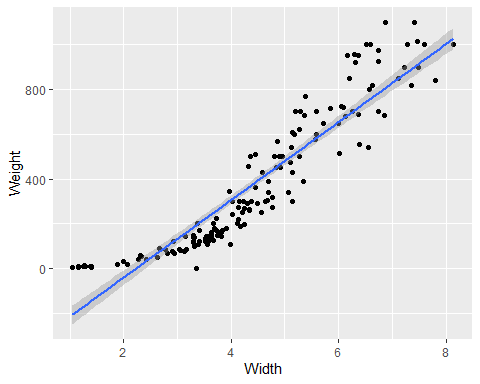
ggplot(fish\_data, aes(x=Height, y=Weight)) + geom\_point()+geom\_smooth(method = "lm")

## `geom\_smooth()` using formula 'y ~ x'



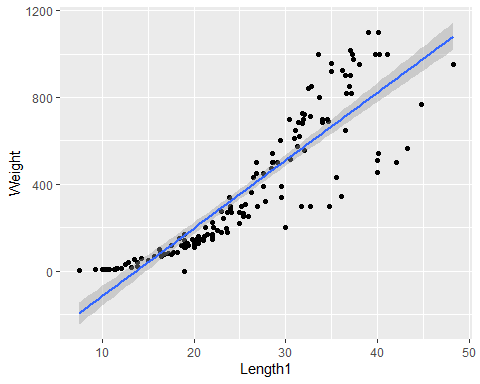
ggplot(fish\_data, aes(x=Width, y=Weight)) + geom\_point()+geom\_smooth(method = "lm")

## `geom\_smooth()` using formula 'y ~ x'



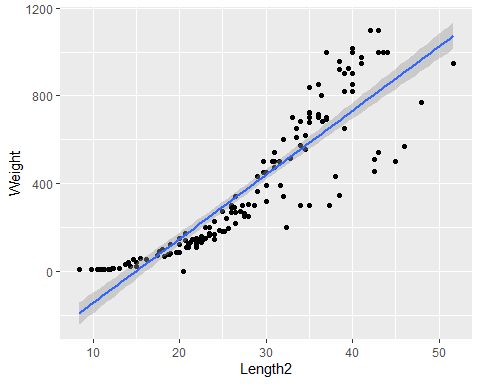
ggplot(fish\_data, aes(x=Length1, y=Weight)) + geom\_point()+geom\_smooth(method = "lm")

## `geom\_smooth()` using formula 'y ~ x'



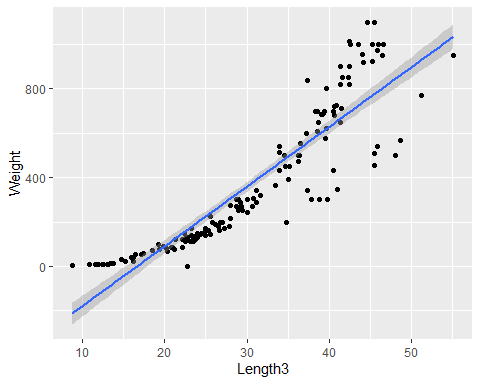
ggplot(fish\_data, aes(x=Length2, y=Weight)) + geom\_point()+geom\_smooth(method = "lm")

## `geom\_smooth()` using formula 'y ~ x'



ggplot(fish\_data, aes(x=Length3, y=Weight)) + geom\_point()+geom\_smooth(method = "lm")

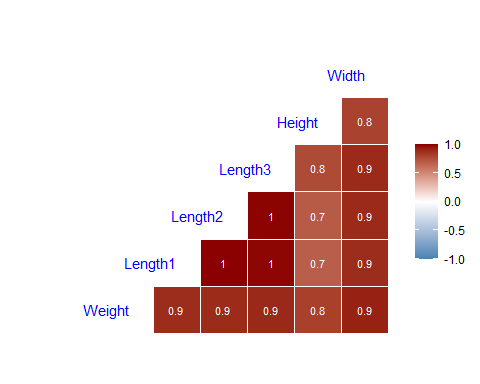
## `geom\_smooth()` using formula 'y ~ x'



As we can see from Scatter plot all independent variables are positively related with the dependent variable (i.e. Weight).

## Checking the Correlation between each variable

par(bg="#E9C9F3")  
ggcorr(fish\_data %>% select(-c(Species)) ,  
 label = TRUE, label\_size = 2.9, hjust =1, layout.exp = 2, low = "steelblue", mid = "white", high = "darkred",label\_color = "white", palette = "RdGy",col="blue")

 We can see from correlation matrix that independent variables length1,length2,length3 are highly correlated and two of them need to be eliminated from our regression model and width is also highly correlated with all the lengths so it should also be eliminated from regression model.

## Modelling

### Test-Train Split

Dividing the data into two parts training and Testing dataset randomly.

fish\_reg <- fish\_data %>%   
 select(-Species)  
set.seed(123)  
samplesize <- round(0.8 \* nrow(fish\_reg), 0)  
index <- sample(seq\_len(nrow(fish\_reg)), size = samplesize)  
data\_train <- fish\_reg[index, ]  
data\_test <- fish\_reg[-index, ]

## 5.Linear Regression

fish\_lm=lm(Weight~.,data=data\_train)  
summary(fish\_lm)

##   
## Call:  
## lm(formula = Weight ~ ., data = data\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -209.04 -65.08 -22.69 73.83 245.10   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -401.703 28.014 -14.339 < 2e-16 \*\*\*  
## Length1 121.243 37.305 3.250 0.00150 \*\*   
## Length2 -55.923 38.415 -1.456 0.14811   
## Length3 -45.200 16.022 -2.821 0.00562 \*\*   
## Height 42.648 8.098 5.266 6.34e-07 \*\*\*  
## Width 50.508 18.873 2.676 0.00851 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 99.36 on 118 degrees of freedom  
## Multiple R-squared: 0.9087, Adjusted R-squared: 0.9048   
## F-statistic: 234.9 on 5 and 118 DF, p-value: < 2.2e-16

Looking on Pr(>|t|) column we will take significance level of 0.05. It means if the value Pr(>|t|) is below 0.05, than we can asume that the variable has significant effect toward the model. The summary of fish\_lm shows four variables(Length1,Length3,Height and Width) are significant towards the model

Weight = -401.703+121.243(length1)-45.200 (Length3)+42.648(Height)+50.508(Width)

## 6.Step-Wise Regression

using Step-wise regression to eliminate the redundant variables(using backward elimination).

fish\_lm\_step <-step(fish\_lm, direction = "backward")

## Start: AIC=1146.34  
## Weight ~ Length1 + Length2 + Length3 + Height + Width  
##   
## Df Sum of Sq RSS AIC  
## <none> 1164924 1146.3  
## - Length2 1 20922 1185846 1146.5  
## - Width 1 70704 1235628 1151.6  
## - Length3 1 78575 1243499 1152.4  
## - Length1 1 104278 1269202 1155.0  
## - Height 1 273793 1438717 1170.5

summary(fish\_lm\_step)

##   
## Call:  
## lm(formula = Weight ~ Length1 + Length2 + Length3 + Height +   
## Width, data = data\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -209.04 -65.08 -22.69 73.83 245.10   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -401.703 28.014 -14.339 < 2e-16 \*\*\*  
## Length1 121.243 37.305 3.250 0.00150 \*\*   
## Length2 -55.923 38.415 -1.456 0.14811   
## Length3 -45.200 16.022 -2.821 0.00562 \*\*   
## Height 42.648 8.098 5.266 6.34e-07 \*\*\*  
## Width 50.508 18.873 2.676 0.00851 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 99.36 on 118 degrees of freedom  
## Multiple R-squared: 0.9087, Adjusted R-squared: 0.9048   
## F-statistic: 234.9 on 5 and 118 DF, p-value: < 2.2e-16

This step-wise regression method will produce an optimum formula based on the lowest AIC value. We can see that the step wise regression eliminates the Length2 and Width variables to produce the smallest AIC. The selected variables are Length1, Length2, Height. Length1 and Height have a significant effect to our model fish\_lm\_step (Pr(>|t|) is below 0.05). We can check the Adjusted R-Squared value from fish\_lm and fish\_lm\_step. The first model with complete variables has adjusted R-squared of 0.9048 or fish\_lm model can explain 90.48% of variance in Fish Weight (independent variable). While the step-wise regression has adjusted R-squared of 0.948. There’s no difference with fish\_lm and fish\_lm\_step. Looking at the p-value of Length3 we we can see that it is above 0.05 so it is not significant towards the model so dropping the Length3 variable from the Regression model

fish\_lm\_step2=lm(Weight~Length1+Height,data=data\_train)  
summary(fish\_lm\_step2)

##   
## Call:  
## lm(formula = Weight ~ Length1 + Height, data = data\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -261.62 -70.44 -30.40 74.97 318.25   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -439.000 31.081 -14.124 <2e-16 \*\*\*  
## Length1 22.206 1.631 13.616 <2e-16 \*\*\*  
## Height 27.711 3.381 8.195 3e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 116.2 on 121 degrees of freedom  
## Multiple R-squared: 0.872, Adjusted R-squared: 0.8699   
## F-statistic: 412.1 on 2 and 121 DF, p-value: < 2.2e-16

## 7. Checking the Assumptions

In Linear regression Method, there are a few assumption that need to be fulfilled. Suppose the assumptions are not satisfied, the result of the model may be inaccurate or misleading.

Assumptions to be checked

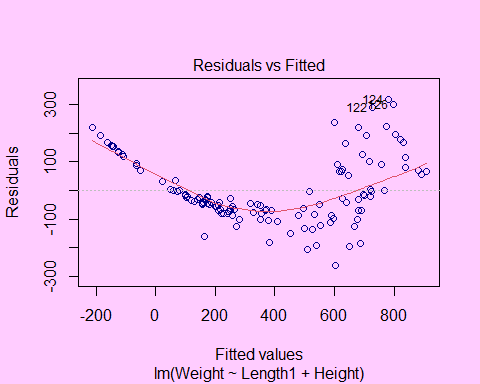
1.Linearity

2.Normality

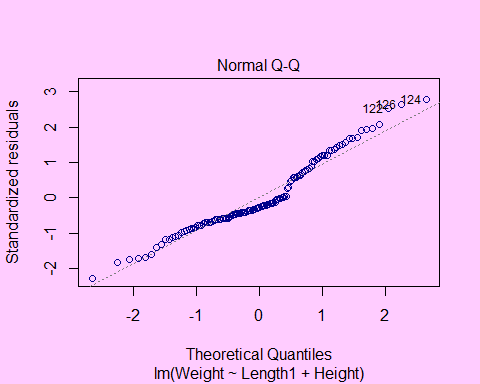
3.Homoscedasticity

4.Multicolinearity

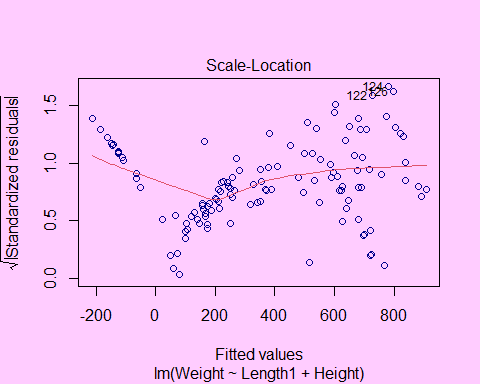
par(bg="#FFCCFF")  
plot(fish\_lm\_step2,col="#00008B",1)



# To verify Linearity assumptions Residual vs Fitted should be linear but we are getting Quadratic curve so linearity assumption is violated.  
plot(fish\_lm\_step2,col="#00008B",2)



# Normal Q-Q plot is used to examine whether the residuals are normally distributed. the residual do not exactly follow the dashed line so the normality assumption is violated.  
plot(fish\_lm\_step2,col="#00008B",3)



# Used to check the homogeneity of variance of the residuals (homoscedasticity). Horizontal line with equally spread points is a good indication of homoscedasticity. Residuals are not uniformly distributed along predictors so it violates Homoscedasticity.  
vif(fish\_lm\_step2)

## Length1 Height   
## 2.007122 2.007122

# Multicollinearity is the occurrence of high intercorrelations among two or more independent variables in a multiple regression model. To check the multicollinearity, we can measure the variance inflation factor (VIF). There’s no multicollinearity when VIF<10 so our model doesn't have multicollinearity.

## 8.Improved Model

fish\_new <- fish\_data %>%   
 mutate\_if(~is.numeric(.), ~sqrt(.)) %>%   
 select(Weight,Length1, Height)  
set.seed(123)  
data\_train2 <- fish\_new[index, ]  
data\_test2 <- fish\_new[-index, ]

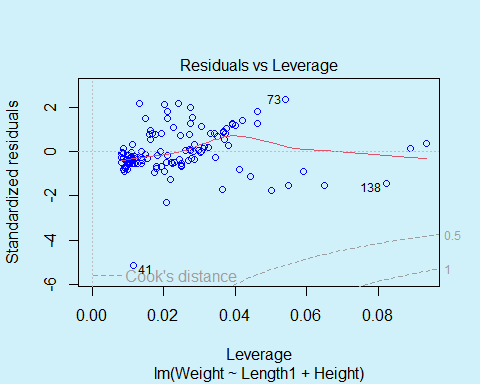
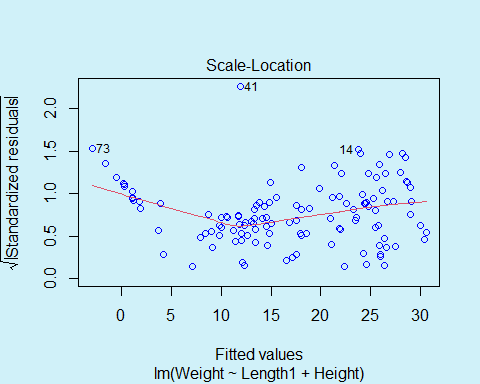
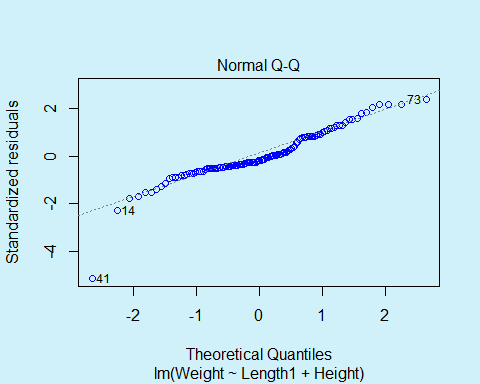
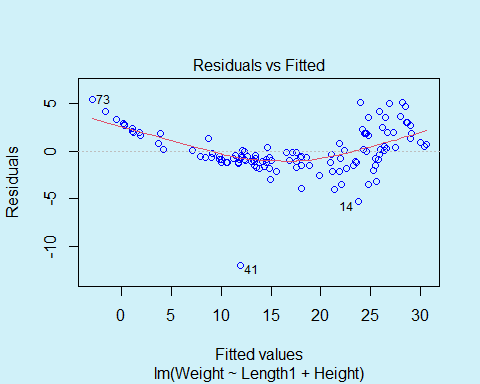
## 9.Regression on Improved Model

fish\_lm\_new <- lm(Weight ~ Length1+Height, data = data\_train2)  
summary(fish\_lm\_new)

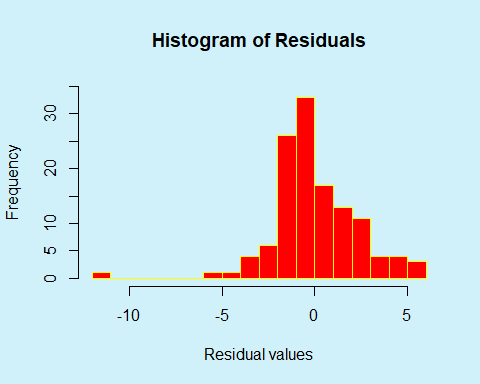
##   
## Call:  
## lm(formula = Weight ~ Length1 + Height, data = data\_train2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.9086 -1.1708 -0.4555 1.6839 5.3606   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -26.3095 1.1453 -22.971 <2e-16 \*\*\*  
## Length1 6.2247 0.3603 17.277 <2e-16 \*\*\*  
## Height 4.3564 0.4399 9.903 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.333 on 121 degrees of freedom  
## Multiple R-squared: 0.9332, Adjusted R-squared: 0.9321   
## F-statistic: 844.7 on 2 and 121 DF, p-value: < 2.2e-16

## 10. Verifying Assumptions of New Model

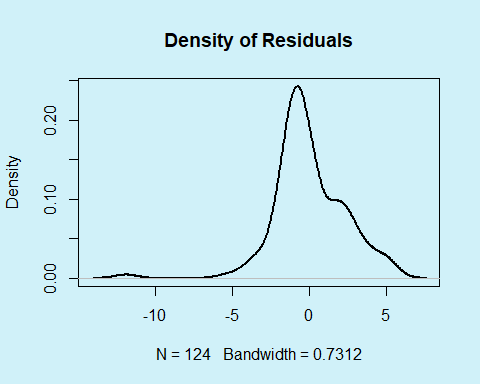
par(bg="#D0F1F9")  
plot(fish\_lm\_new,col="blue")



hist(fish\_lm\_new$residuals,breaks=20,main="Histogram of Residuals",col="red",ylim=c(0,35),border = "yellow",xlab="Residual values")



plot(density(fish\_lm\_new$residuals),main="Density of Residuals",lwd=2)



vif(fish\_lm\_new)

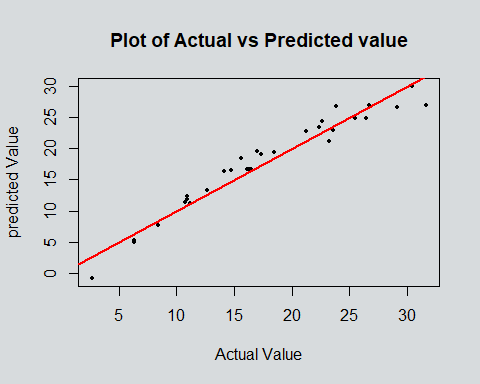
## Length1 Height   
## 2.547134 2.547134

As we can see our all assumptions are satisfied

1. Residual vs Fitted plot gives a straight line so Linearity assumption is satisfied
2. Normal Q-Q plot of the residuals lies along the dashed line so the normality assum[ption is satisfied Histogram and Density plot also verifies normality assumption.
3. We can see a horizontal line with equally spread points in scale location plot so Homoscedasticity is also verified.
4. The Variance Inflation factor (VIF) shows a value less than 10 so data is not multicollinear.

## 11. Prediction on the basis of New Model

lm\_pred2 <- predict(fish\_lm\_new, newdata = data\_test2 %>% select(-Weight))  
par(bg="#D7DBDD")  
plot(data\_test2$Weight,lm\_pred2,xlab="Actual Value",ylab="predicted Value",main = "Plot of Actual vs Predicted value",cex=0.6,pch=19)  
abline(a=0, b=1,col="red",lwd=2)



As we can see from the above scatter plot almost all of the predicted value lies close to the actual value we can conform the same from the red line as most of the value lies in the vicinity of the y=x line.

## 12.Conclusion

Variables that have a significant effect on weight of fish are length1 and height. we need to eliminate the variables Length2,Length3 and Width because these variables have a high intercorrelation. We have transformed the data to square root and fulfilled all the assumptions of the linear regression. The final linear regression model has a Adjusted R-squared of 0.9321,it means that 93.21% of the variables can explain the variance in fish weight. For profitable business purposes,to increase the Weight of fish, fisherman needs to increase the Length1 and Height of each fish.