SECTION 1 Problem 1

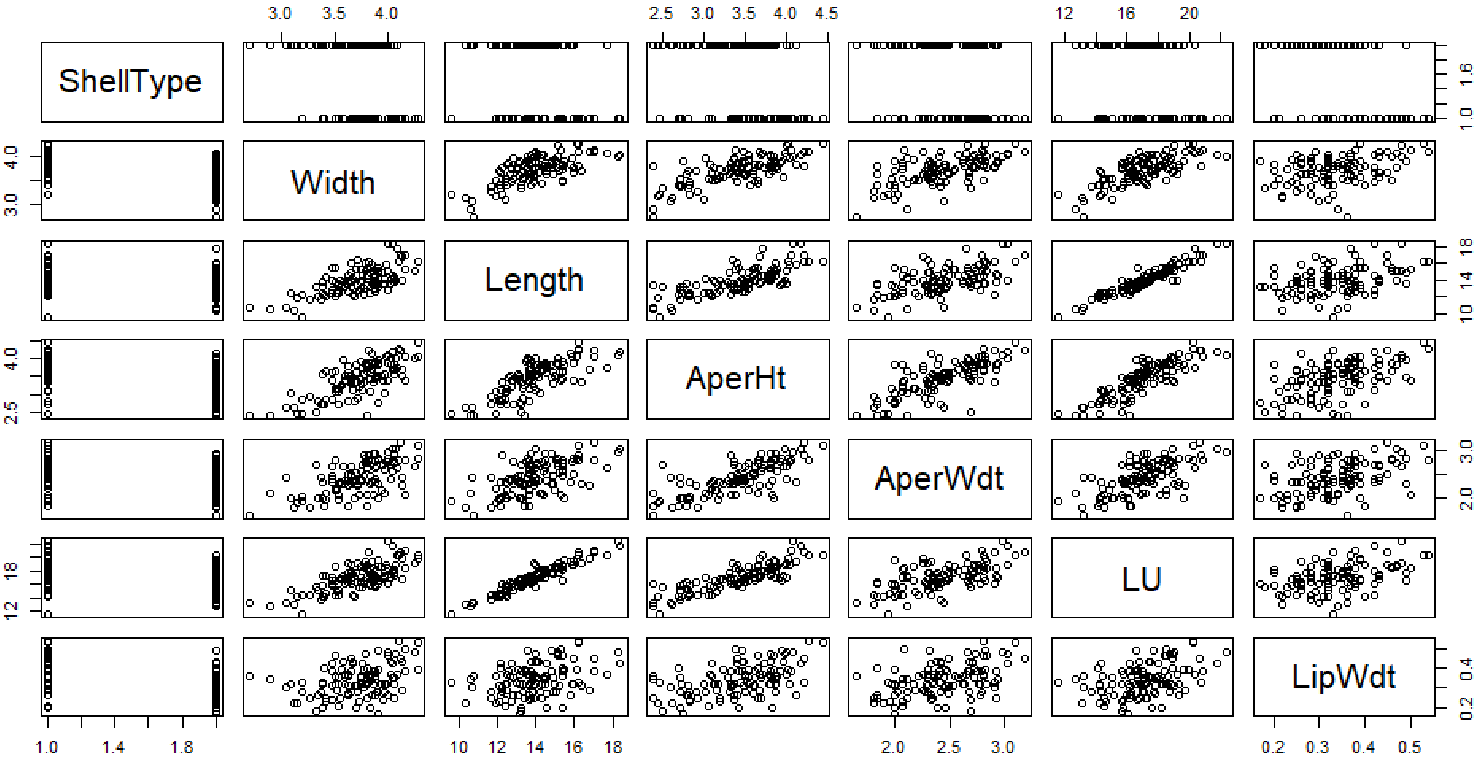
a) First for the exploratory analysis, the structure of the data set was observed. Following is an outline of the data set.

```
'data.frame':110 obs. of  7 variables:
 $ ShellType: Factor w/ 2 levels "Type1","Type2": 1 1 1 1 1 1 1 1 1 1 ...
 $ Width    : num  4.04 3.77 3.86 4.13 4.25 4.16 3.87 3.91 3.99 4.02 ...
 $ Length   : num  14.4 13.1 13.6 13.9 15.1 ...
 $ AperHt   : num  3.7 3.54 3.33 3.49 3.99 3.38 3.41 3.95 3.36 3.81 ...
 $ AperWdt  : num  2.58 2.46 2.06 2.42 2.66 2.08 2.32 2.82 2.78 2.7 ...
 $ LU       : num  17.9 16.6 16.9 16.8 19 ...
 $ LipWdt   : num  0.35 0.39 0.32 0.43 0.48 0.5 0.26 0.49 0.26 0.43 ...
```

This tells us that the dataset has 110 observations and 7 variables. First variable is categorical and the other 6 variables are numerical. Below shows summaries of each numeric variable. To visualize each of them we can use box plot and to see the relationships between them we can use scatterplots.

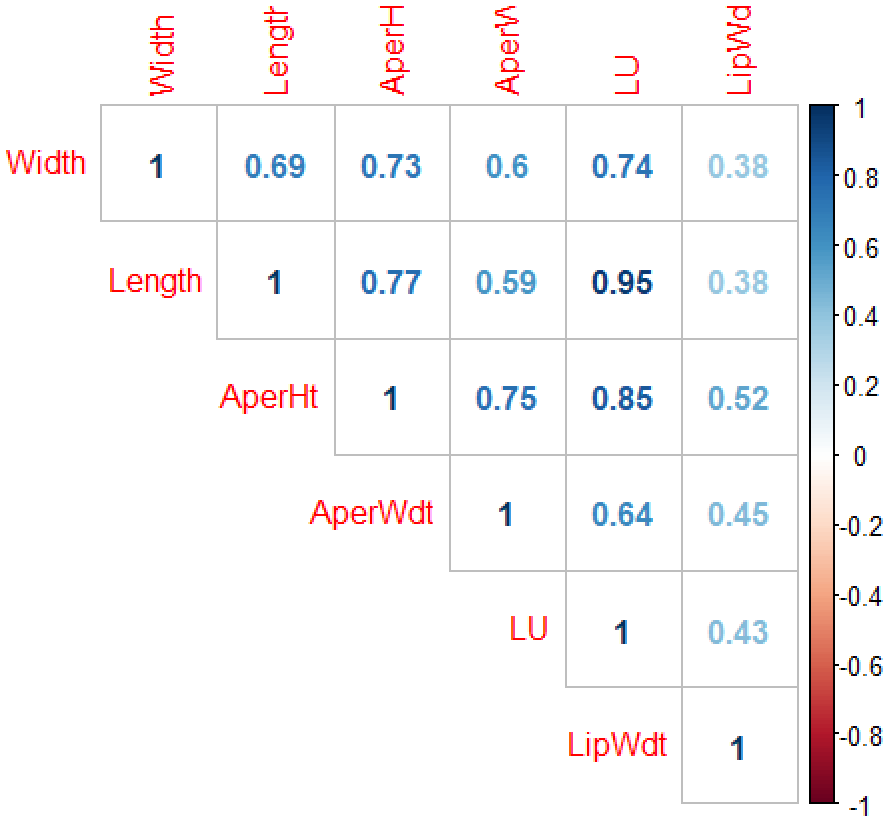
```
     Width           Length          AperHt         AperWdt     
 Min.   :2.700   Min.   : 9.63   Min.   :2.400   Min.   :1.640  
 1st Qu.:3.530   1st Qu.:12.88   1st Qu.:3.103   1st Qu.:2.165  
 Median :3.760   Median :13.79   Median :3.490   Median :2.410  
 Mean   :3.718   Mean   :13.94   Mean   :3.426   Mean   :2.428  
 3rd Qu.:3.908   3rd Qu.:14.96   3rd Qu.:3.800   3rd Qu.:2.700  
 Max.   :4.280   Max.   :18.39   Max.   :4.430   Max.   :3.180  
       LU            LipWdt      
 Min.   :11.59   Min.   :0.1700  
 1st Qu.:15.96   1st Qu.:0.2800  
 Median :17.02   Median :0.3300  
 Mean   :17.03   Mean   :0.3397  
 3rd Qu.:18.33   3rd Qu.:0.3800  
 Max.   :22.36   Max.   :0.5400  
```

Simple Scatterplot Matrix



By looking at the Scatterplot, some of the variables clearly shows a linear relation. Namely, Length, AperHt, AperWdt and LU shows strong linear relation.

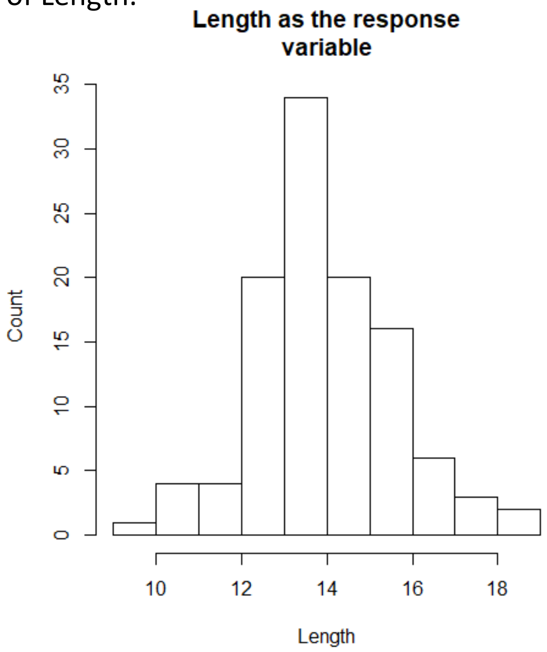
Another way to see the relationship between variables is to have the Correlation matrix.

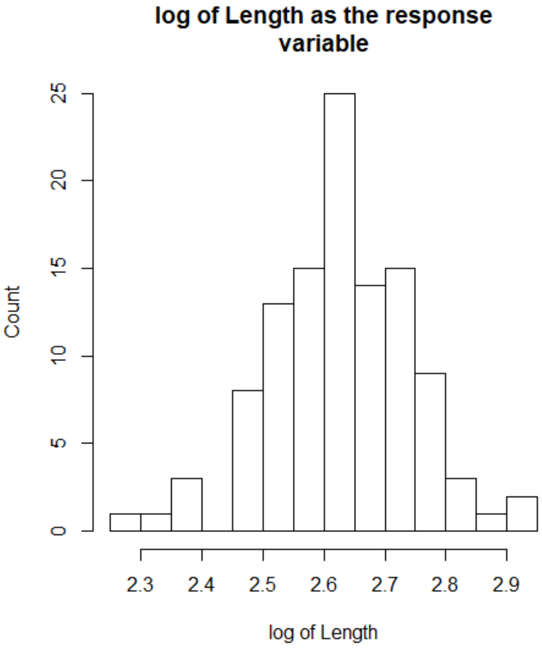


	Width	Lengtł	AperH	AperW	LU	LipWd
Width	1	0.69	0.73	0.6	0.74	0.38
Length		1	0.77	0.59	0.95	0.38
AperHt			1	0.75	0.85	0.52
AperWdt				1	0.64	0.45
LU					1	0.43
LipWdt						1

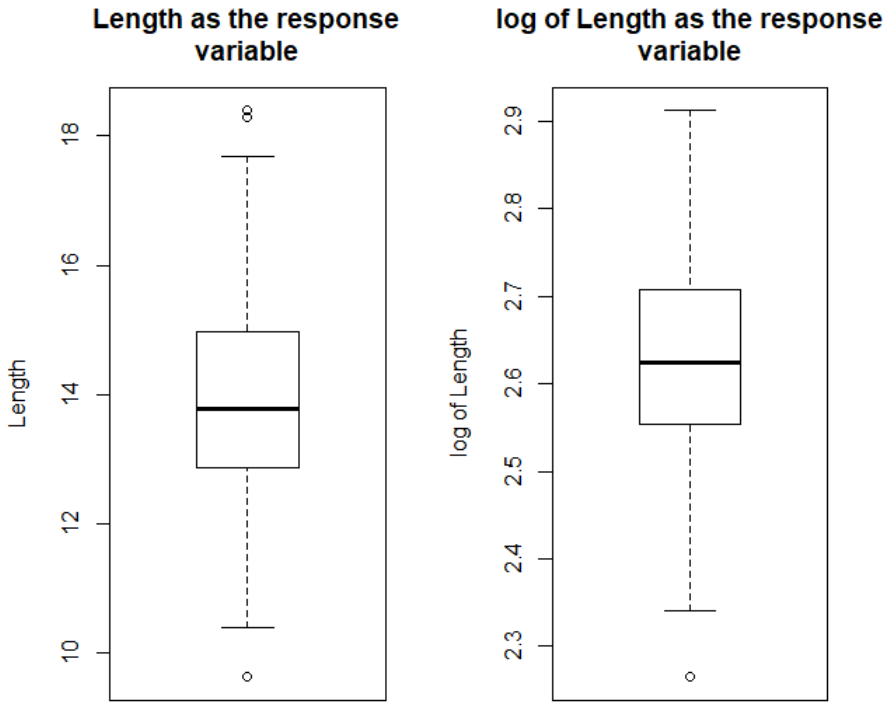
With the Correlation matrix, it is easy to identify which variables have the strongest correlations. Length and LU have the strongest of them and AperHt and LU, Length and AperHt, AperHt and AperWdt, Length and AperHt also seem to have some strong relations.

b) To check if Length is appropriate as a response variable, we check whether it fits to a normal distribution. To do that, we look at the histogram of Length and another possible transformation, Log of Length.





As both plots are reasonably symmetric and uni-model, they seem to have normal distribution shape. And since the range of Length variable is from 9.63 to 18.39, we don't need to have log transformation for Length. Below shows another way to visualize this bell shape using box plots.



This can be further confirmed by the Shapiro Wilk Test.

$$H_0 = Variable\ is\ normally\ distributed$$
$$H_1 = Variable\ is\ not\ normally\ distributed$$

```
Shapiro-Wilk normality test
data:  Length
W = 0.98854, p-value = 0.4784


Shapiro-Wilk normality test
data:  log.Length
W = 0.98825, p-value = 0.456
```

P-value for Length and log of Length are both close to 0.5. Hence, both variables can be assumed normally distributed and since P-value of Length is more close to 0.5, best out of these two is to choose Length as the respond variable.

c) Simple Linear models were constructed for each of the variables in the data set.
For each of the models the significance of the variable was tested using the t test for variable significance for the following hypothesis.

$$H_0\ = The\ variable\ is\ significant$$
$$H_1 = The\ variable\ is\ not\ significant$$

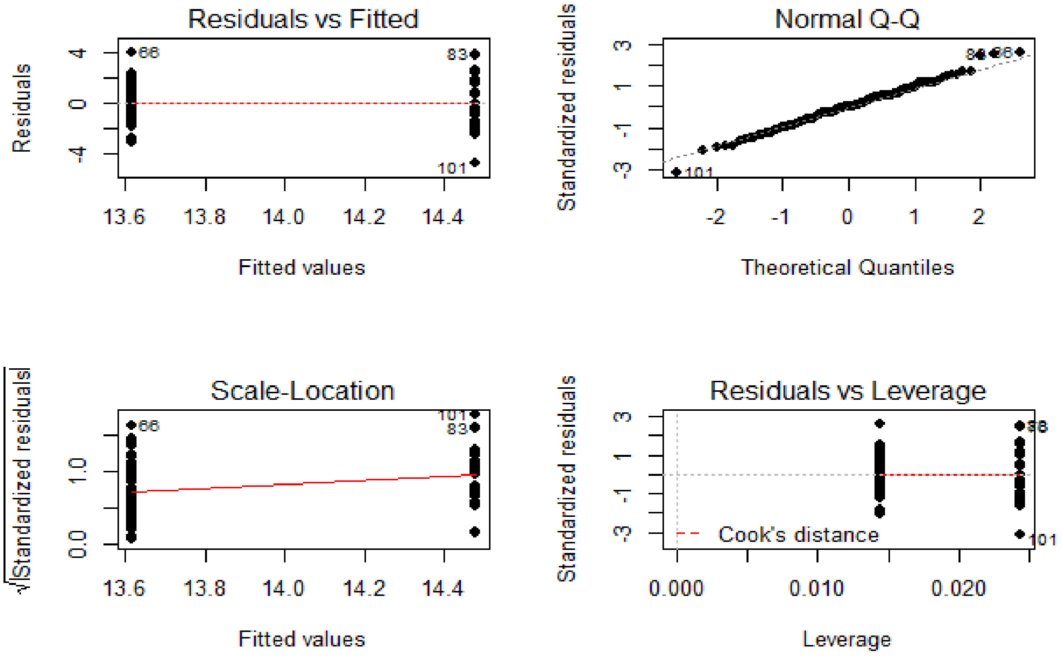
```
Model 1 (Length ~ ShellType)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.8505 -0.9865  0.0055  0.9195  4.0655 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     14.4805     0.2452  59.059   <2e-16 ***
ShellTypeType2  -0.8660     0.3096  -2.797   0.0061 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.57 on 108 degrees of freedom
Multiple R-squared:  0.06756,	Adjusted R-squared:  0.05893 
F-statistic: 7.825 on 1 and 108 DF,  p-value: 0.006101
```

The R^2 of the model is 0.06756 which states that 6.76% of the variation of fitted values of Length variable is explained by ShellType variable which is a categorical variable with two levels. The P value of significance for ShellType is less than 0.05. This concludes that at 0.05 confidence level the variable ShellType is significant for predicting Length Level.



The end points in the QQ plot do not fall on the straight line. Therefore, the plot do not justify the normality assumption.

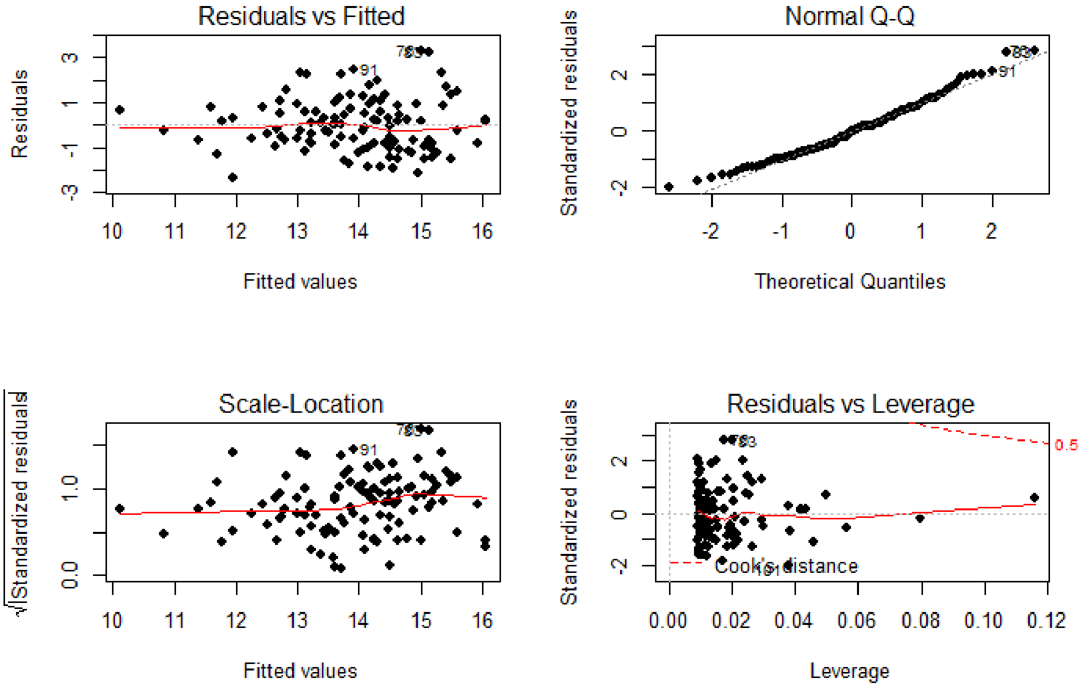
```
Model 2 (Length ~ Width)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.3282 -0.8405 -0.0127  0.7662  3.3068 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.005743   1.408795   0.004    0.997    
Width       3.746866   0.377692   9.920   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 1.176 on 108 degrees of freedom
Multiple R-squared:  0.4768,     Adjusted R-squared:  0.4719
F-statistic: 98.42 on 1 and 108 DF,  p-value: < 2.2e-16
```

The R^2 of the model is 0.4768 which states that 47.68% of the variation of fitted values of Length variable is explained by Width variable. The P value of significance for Width is less than 0.05. This concludes that at 0.05 confidence level the variable Width is significant for predicting Length Level.



For this model the plot of residuals Vs fitted values does not show any obvious pattern. Also, we can see that the residuals are evenly distributed around the center zero line. Therefore, we can assume that the assumption of Linearity holds for this model. However all the points in the QQ plot do not fall along the straight line. Therefore, we cannot make any assumption on normality.

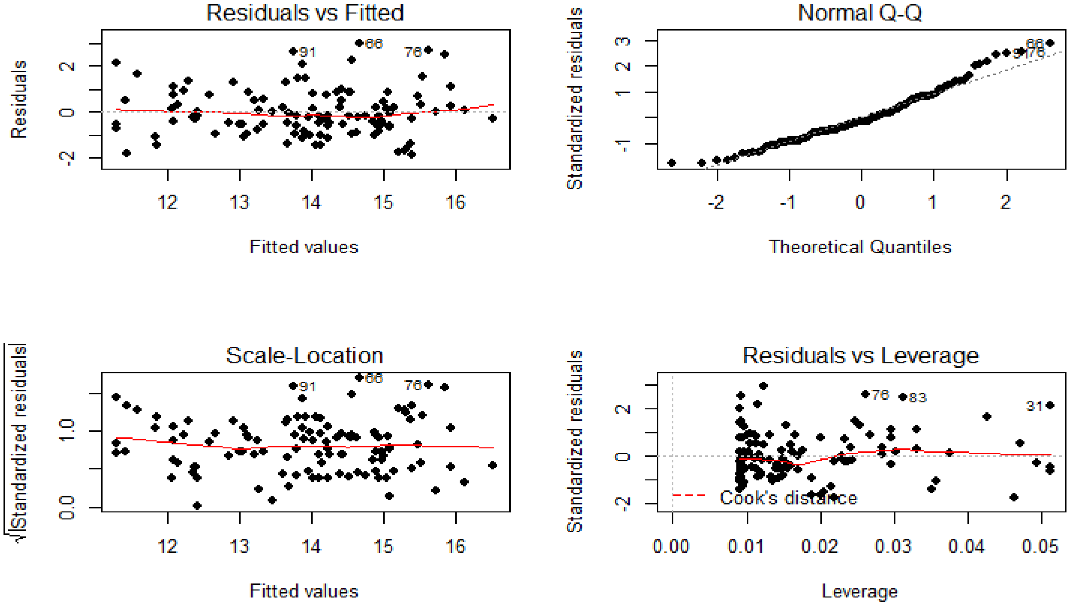
```
Model 3 (Length ~ AperHt)


Residuals:
    Min      1Q  Median      3Q     Max 
-1.8489 -0.6294 -0.1706  0.6583  3.0064 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   5.0634     0.7221   7.012 2.13e-10 ***
AperHt        2.5904     0.2088  12.407  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.044 on 108 degrees of freedom
Multiple R-squared:  0.5877,     Adjusted R-squared:  0.5839
F-statistic: 153.9 on 1 and 108 DF,  p-value: < 2.2e-16
```

The R^2 of the model is 0.5877 which states that 58.77% of the variation of fitted values of Length variable is explained by AperHt variable. The P value of significance for AperHt is less than 0.05. This concludes that at 0.05 confidence level the variable AperHt is significant for predicting Length Level.



Again, in this model, the plot of residuals Vs fitted values does not show any obvious pattern. Also, we can see that the residuals are evenly distributed around the center zero line. Therefore, we can assume that the assumption of Linearity holds for this model. However the end points in the QQ plot do not fall on the straight lin. Therefore, we cannot make any assumption on normality.

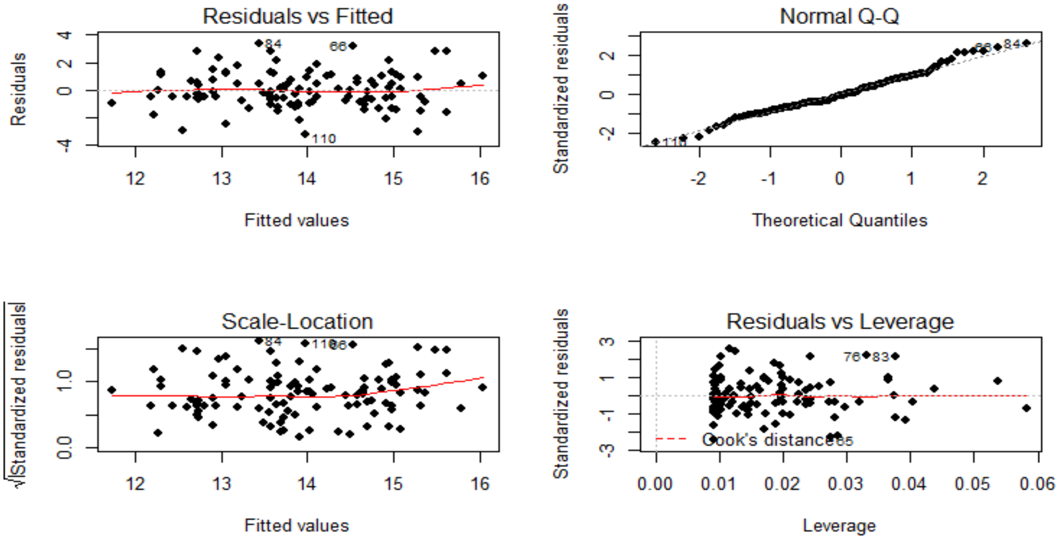
```
Model 4 (Length ~ AperWdt)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.2505 -0.8431 -0.1279  0.8592  3.3895 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   7.1638     0.9079   7.891 2.63e-12 ***
AperWdt       2.7896     0.3703   7.533 1.60e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.316 on 108 degrees of freedom
Multiple R-squared:  0.3445,	Adjusted R-squared:  0.3384 
F-statistic: 56.75 on 1 and 108 DF,  p-value: 1.602e-11
```

The R^2 of the model is 0.3445 which states that 34.45% of the variation of fitted values of Length variable is explained by AperWdt variable. The P value of significance for AperWdt is less than 0.05. This concludes that at 0.05 confidence level the variable AperWdt is significant for predicting Length Level.



We can see that the residuals are evenly distributed around the center zero line. Therefore, we can assume that the assumption of Linearity holds for this model. However the end points in the QQ plot do not fall on the straight lin. Therefore, we cannot make any assumption on normality.

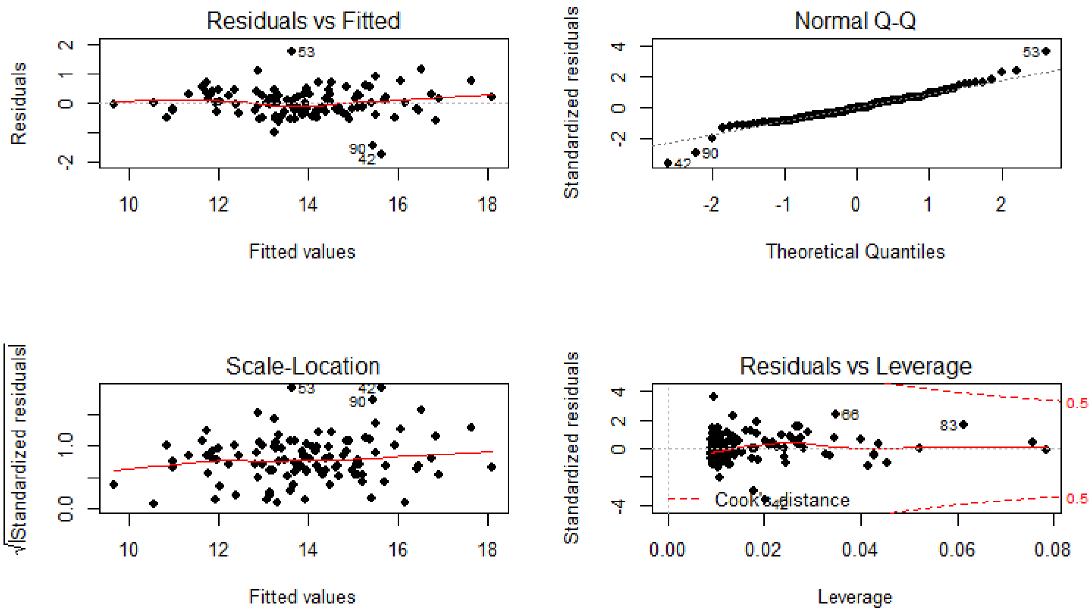
```
Model 5 (Length ~ LU)

Residuals:
     Min       1Q   Median       3Q      Max
-1.74502 -0.28542 -0.01718  0.27650  1.75611

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6430     0.4011   1.603    0.112
LU            0.7808     0.0234  33.369   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4834 on 108 degrees of freedom
Multiple R-squared:  0.9116,    Adjusted R-squared:  0.9108
F-statistic:  1114 on 1 and 108 DF,  p-value: < 2.2e-16
```

The R^2 of the model is 0.9116 which states that 91.16% of the variation of fitted values of Length variable is explained by LU variable. The P value of significance for LU is less than 0.05. This concludes that at 0.05 confidence level the variable LU is significant for predicting Length Level.



the plot of residuals Vs fitted values shows the points fall along the center zero line and evenly scattered around the center line. Therefore, we can assume that the assumption of Linearity holds for this model. However the end points in the QQ plot do not fall on the straight lin. Therefore, we cannot make any assumption on normality.

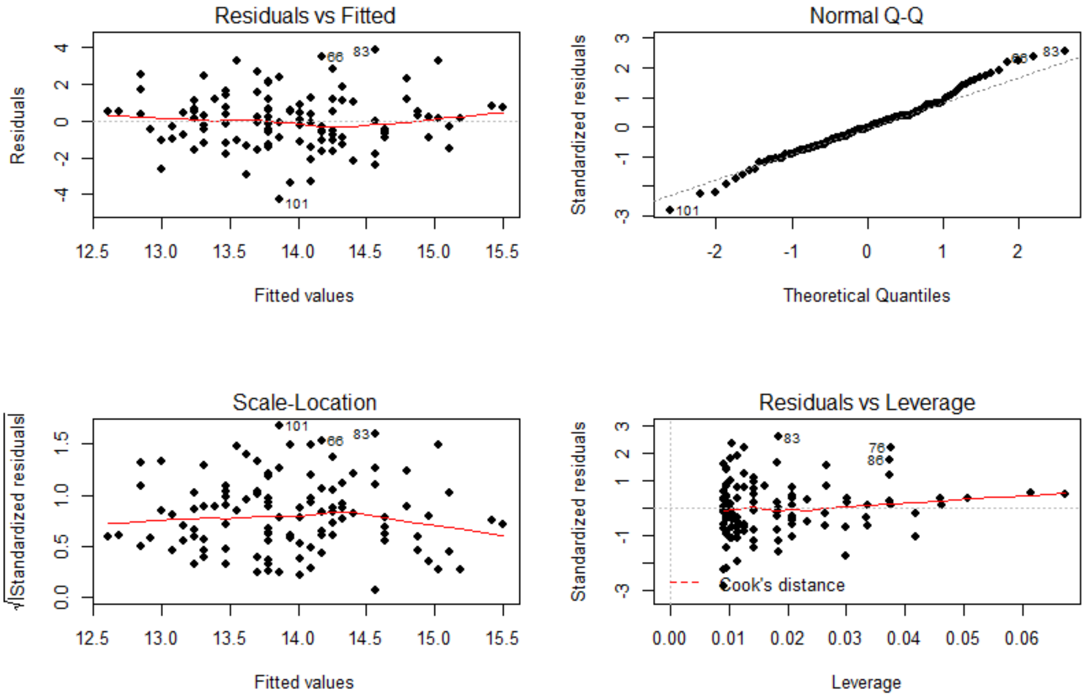
```
Model 6 (Length ~ LipWdt)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2314 -0.9618 -0.0883  0.7516  3.8264

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.2864     0.6298  17.922  < 2e-16 ***
LipWdt        7.8029     1.8052   4.322 3.45e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.501 on 108 degrees of freedom
Multiple R-squared:  0.1475,    Adjusted R-squared:  0.1396
F-statistic: 18.68 on 1 and 108 DF,  p-value: 3.448e-05
```

The R^2 of the model is 0.1475 which states that 14.75% of the variation of fitted values of Length variable is explained by LipWdt variable. The P value of significance for LipWdt is less than 0.05. This concludes that at 0.05 confidence level the variable LipWdt is significant for predicting Length Level.



For this model the plot of residuals Vs fitted values does not show any obvious pattern. Also, we can see that the residuals are evenly distributed around the center zero line. Therefore, we can assume that the assumption of Linearity holds for this model. However all the points in the QQ plot do not fall along the straight line. Therefore, we cannot make any assumption on normality.

d) A Multiple Regression model was obtained for all the variables.

```
Model.ALL

Residuals:
     Min       1Q   Median       3Q      Max 
-1.71415 -0.23723 -0.03966  0.22707  1.95250 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     1.08963    0.64828   1.681   0.0958 .  
ShellTypeType2 -0.18429    0.10589  -1.740   0.0848 .  
Width          -0.09897    0.24115  -0.410   0.6824    
AperHt         -0.52159    0.21601  -2.415   0.0175 *  
AperWdt         0.04369    0.20264   0.216   0.8297    
LU              0.89316    0.04562  19.578   <2e-16 ***
LipWdt         -0.57332    0.71363  -0.803   0.4236    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4687 on 103 degrees of freedom
Multiple R-squared:  0.9207,    Adjusted R-squared:  0.9161 
F-statistic: 199.4 on 6 and 103 DF,  p-value: < 2.2e-16
```

The adjusted R^2 of the model is 0.9161 which states that 91.61% of the variation of fitted values of Length variable is explained by the variables in the model.

Overall F Test

$$H_0 = All\ the\ variables\ in\ the\ model\ are\ not\ associated\ with\ Length$$
$$H_1 = At\ least\ one\ variable\ in\ the\ model\ is\ associated\ with\ Length$$

The P-value under the overall F test is essentially zero. Therefor we can conclude that at least one of the variables in the model is significant in predicting PSA Level. When the individual P values for the significance of variables are considered only the Variables AperHt and LU have P values less than 0.05 and hence are associated with Length while the other variables under the presence of these significant variables are not associated with Length. Following is the model fitted with these significant variables.

Model (Length ~ AperHt + LU)

```
Residuals:
    Min      1Q  Median      3Q     Max 
-1.7881 -0.2713 -0.0476  0.2748  1.8932 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.60596    0.38745   1.564  0.12078    
AperHt      -0.52429    0.17638  -2.972  0.00365 ** 
LU           0.88850    0.04269  20.813  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4668 on 107 degrees of freedom
Multiple R-squared:  0.9183,	Adjusted R-squared:  0.9168 
F-statistic: 601.6 on 2 and 107 DF,  p-value: < 2.2e-16
```

The P value of the overall F test is very low concluding the model significance. It also can be observed that individual P values of all the variables are less than 0.05 concluding all the variables are significant in predicting Length at less than 0.05 significance level.

The nest task is to check for the interactions between ShellType and significant variables. The following is the model with all the interaction terms.

e)

Model with interactions of ShellType (Length ~ AperHt + LU + AperHt:ShellType + LU:ShellType)

```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.65580 -0.25051 -0.01991  0.23192  1.84090 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)            1.03894    0.40384   2.573  0.01149 *  
AperHt                -1.15622    0.28510  -4.056 9.63e-05 ***
LU                     0.99726    0.05786  17.237  < 2e-16 ***
AperHt:ShellTypeType2  0.97063    0.36668   2.647  0.00937 ** 
LU:ShellTypeType2     -0.20506    0.07468  -2.746  0.00710 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.4519 on 105 degrees of freedom
Multiple R-squared:  0.9249,	Adjusted R-squared:  0.922 
F-statistic: 323.2 on 4 and 105 DF,  p-value: < 2.2e-16
```

The adjusted R^2 of the model is 0.922 which states that 92.2% of the variation of fitted values of Length variable is explained by the variables in the model. This percentage is bit higher than we observed without interactions. Thus, this is a better model to fit Length.

```
Analysis of Variance Table

Model 1: Length ~ AperHt + LU
Model 2: Length ~ AperHt + LU + AperHt:ShellType + LU:ShellType
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)  
1    107 23.315                              
2    105 21.445  2    1.8701 4.5783 0.01241 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

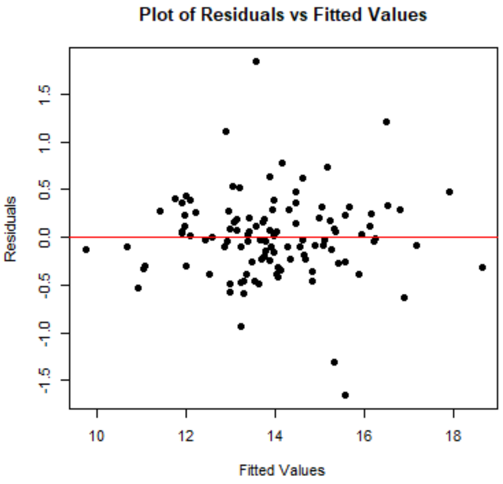
The P value is less than 0.05. Therefore at 5% significance level we reject the null hypothesis concluding that the model 2 is the better fit.

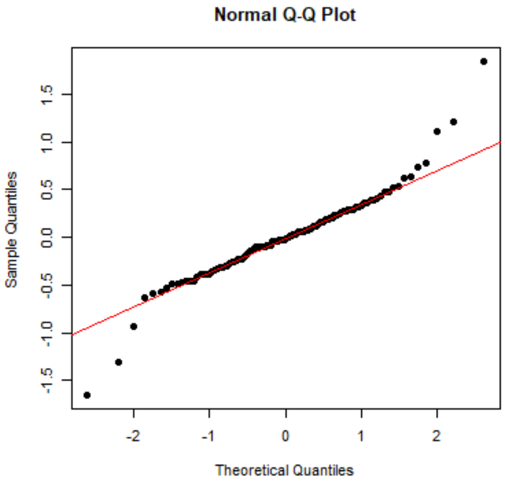
f) final model is,

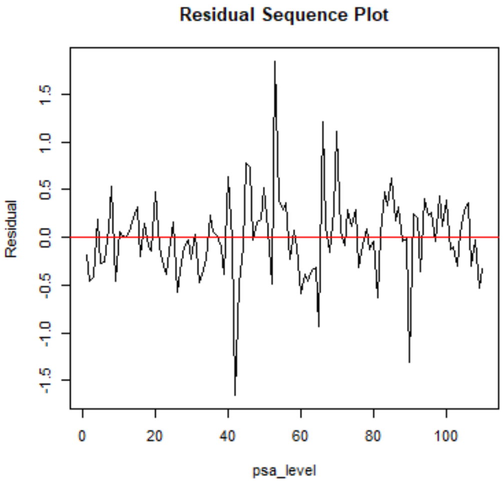
$$\mathbf{Length} = 1.03894 - 1.15622 \ast \mathbf{AperHt} + 0.99726 \ast \mathbf{LU} + 0.97063 \ast (\mathbf{AperHt} \ast \mathbf{ShellType}) - 0.20506 \ast (\mathbf{LU} \ast \mathbf{ShellType})$$

Considering the interactions, we can note these slops.

$$\text{slop of } \mathbf{AperHt} = -1.15622 + 0.97063 \ast \mathbf{ShellType}$$
$$\text{slop of } \mathbf{LU} = 0.99726 - 0.20506 \ast \mathbf{ShellType}$$







The residuals does not show any obvious pattern and are evenly distributed around the zero line in the plot of residuals Vs fitted values. This provides evidence that the assumptions of homoscedasticity and linearity holds. The normality of the model is questionable from the plot since the end points doesn't lie on the straight line. Also The Residual sequence plot do not seem to show a completely random pattern. It can be observed that in the middle we can see very sharp spikes to both directions. This indicates that the assumption of independence of error terms might not hold for the model. Apart from this Graphical diagnostics following tests were conducted to further establish these conclusions.

```
Shapiro-Wilk normality test

data:  residuals(Model.int)
W = 0.93694, p-value = 5.74e-05
```

H_0: *The errors are normally distributed.*
H_1: *The errors are not normally distributed.*

p-value = 5.74e-05 < 0.05, assumption of the normality of errors terms doesn't hold at 5% level of significance.

```
studentized Breusch-Pagan test

data:  Model.int
BP = 4.684, df = 4, p-value = 0.3213
```

H_0: *The errors have constant variance.*
H_1: *The errors do not have constant variance.*

p-value = 0.3213 > 0.05, the assumption of the constant variance of errors hold at 5% level of significance.

```
Durbin-Watson test

data:  Model.int
DW = 1.7817, p-value = 0.08951
alternative hypothesis: true autocorrelation is greater than 0
```

d = 1.7817. this value is closer to 2. This indicates that the residuals are independent.

g) The fitted value for Length when the quantitative variables are at their mean levels and the qualitative variable, ShellType, is Type1 and Type2 are as follows.

```
> Length.Type1
       1
14.05715

> Length.Type2
       1
13.89095
```

SECTION 2 Problem 1

```
# Importing the dataset
train.data <- read.csv(file.choose(), header = T)
attach(train.data)

a)
# Visualizing the structure of the dataset
str(train.data)
summary(train.data[,-1]) # removing categorical variable

# Scatterplot of the variables
pairs(~ + ShellType  + Width + Length  + AperHt + AperWdt + LU + LipWdt, data=train.data,
main="Simple Scatterplot Matrix")

# Plotting correlation plot between PSA Level and other variables
install.packages("corrplot")
library(corrplot)

M <- cor(train.data [,-1]) # removing categorical variable
corrplot(M, type="upper", method="number")


b)
#checking whether the response variable Length approximately follow normal
distribution
log.Length <- log(Length)
# Histograms of Length and log of Length
par(mfrow = c(1, 2))
hist(Length, main="Length as the response
variable", breaks=10, xlab="Length", ylab="Count")
hist(log.Length, main="log of Length as the response
variable", breaks=10, xlab="log of Length", ylab="Count")

# Box plots of Length and log of Length
par(mfrow = c(1, 2))
boxplot(train.data$Length, main="Length as the response
variable", ylab="Length")
boxplot(log.Length, main="log of Length as the response
variable", ylab="log of Length")


c)

# Fitting Linear regression models for the respond variable with others
Model.ShellType = lm(Length~ShellType, data=train.data)
summary(Model.ShellType)
par(mfrow = c(2, 2))
plot(Model.ShellType, pch = 19)

Model.Width = lm(Length~Width, data=train.data)
summary(Model.Width)
par(mfrow = c(2, 2))
plot(Model.Width, pch = 19)

Model.AperHt = lm(Length~AperHt, data=train.data)
```

```
summary(Model.AperHt)
par(mfrow = c(2, 2))
plot(Model.AperHt, pch = 19)

Model.AperWdt = lm(Length~AperWdt, data=train.data)
summary(Model.AperWdt)
par(mfrow = c(2, 2))
plot(Model.AperWdt, pch = 19)

Model.LU = lm(Length~LU, data=train.data)
summary(Model.LU)
par(mfrow = c(2, 2))
plot(Model.LU, pch = 19)

Model.LipWdt = lm(Length~LipWdt, data=train.data)
summary(Model.LipWdt)
par(mfrow = c(2, 2))
plot(Model.LipWdt, pch = 19)
```

d)

```
# Fitting a multiple linear regression model using all the predictor variables
Model.All = lm(Length ~. , data=train.data)
summary(Model.All)


# Fitting a multiple linear regression model using only the significant variables
Model.droped = lm(Length ~ + AperHt + LU , data=train.data)
summary(Model.droped)
```

e)

```
Model.int=lm(Length ~ AperHt + LU + AperHt:ShellType + LU:ShellType, data=train.data)
summary(Model.int)


anova(Model.droped, Model.int)
```

f)

```
par(mfrow = c(1, 3))
# Plot of residuals vs fitted for main effects only model
plot(fitted(Model.int), residuals(Model.int), pch = 19, xlab = "Fitted Values", ylab =
"Residuals", main = "Plot of Residuals vs Fitted Values")
abline(h = 0, col = "red")

# Normal Q-Q plot
qqnorm(residuals(Model.int), pch = 19, main = "Normal Q-Q Plot")
qqline(residuals(Model.int), col = "red")

# Residual sequential plot to check the independence of residuals
plot(residuals(Model.int), xlab = "psa_level", ylab = "Residual", main="Residual Sequence Plot",
type = "l")
abline(h = 0, col = "red")

# Normality test: Shapiro-Wilk normality test
shapiro.test(residuals(Model.int))
```

```
# Constant variance: Breusch-Pagan test
library(lmtest)
bptest(Model.int)

# Durbin-Watson test to check whether residuals are uncorrelated
dwtest(Model.int)
```

g)

```
# predicting Length with ShellType = ShellType1
newdata1 <- data.frame(ShellType=factor("Type1", levels=c("Type1", "Type2")),
AperHt=mean(AperHt), LU=mean(LU))
Length.Type1 = predict(Model.int, newdata1)

# predicting Length with ShellType = ShellType2
newdata2 <- data.frame(ShellType=factor("Type2", levels=c("Type1", "Type2")),
AperHt=mean(AperHt), LU=mean(LU))
Length.Type2 = predict(Model.int, newdata2)
```