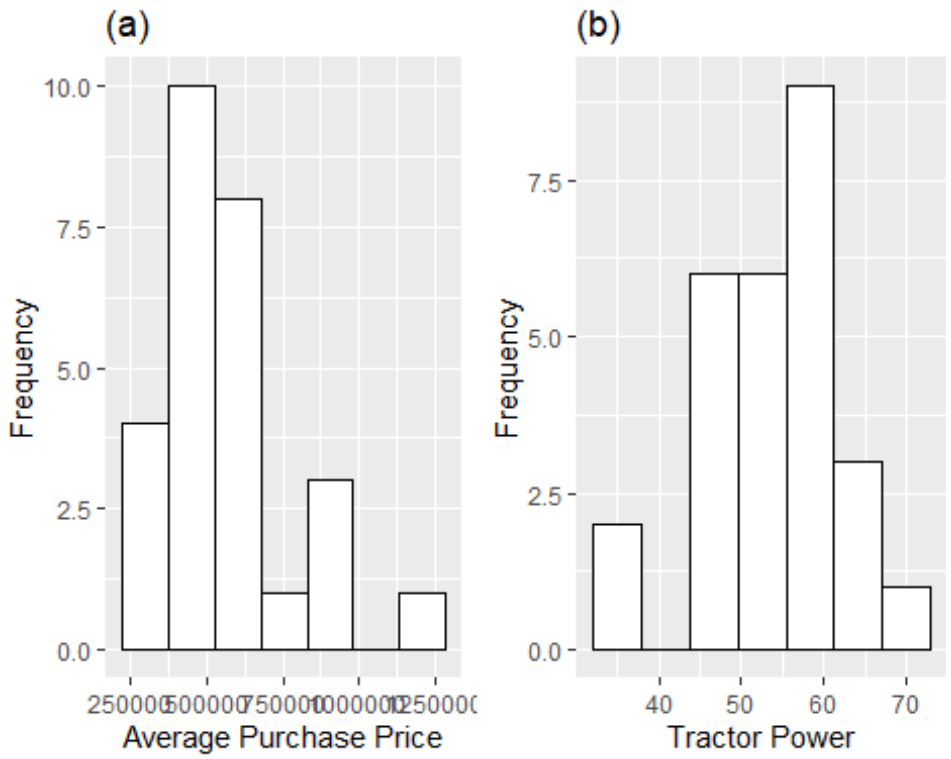
STA4813 Assignment 1

Neo Deane 66931185

2024-04-12

Question A

```
Machinery <- read.csv("C:\\Users\\Deane R.H\\Documents\\Machinery.csv",header
=TRUE,dec=",",sep=";")
Machinery$Type_Tractor <- as.factor(Machinery$Type_Tractor)



library(gridExtra)
library(ggplot2)
Ave_Purchase_price <- ggplot(data = Machinery, aes(x = Average_Purchase_Price
)) +
  geom_histogram(colour = "black", fill = "white", bins = 7) +
  xlab("Average Purchase Price") + ylab("Frequency") + labs(title="(a)")
Trac_Power <- ggplot(data = Machinery, aes(x = Tractor_Power)) +
  geom_histogram(colour ="black", fill = "white", bins = 7) +
  xlab("Tractor Power") + ylab("Frequency") + labs(title="(b)")
grid.arrange(Ave_Purchase_price,Trac_Power , ncol = 2)
```



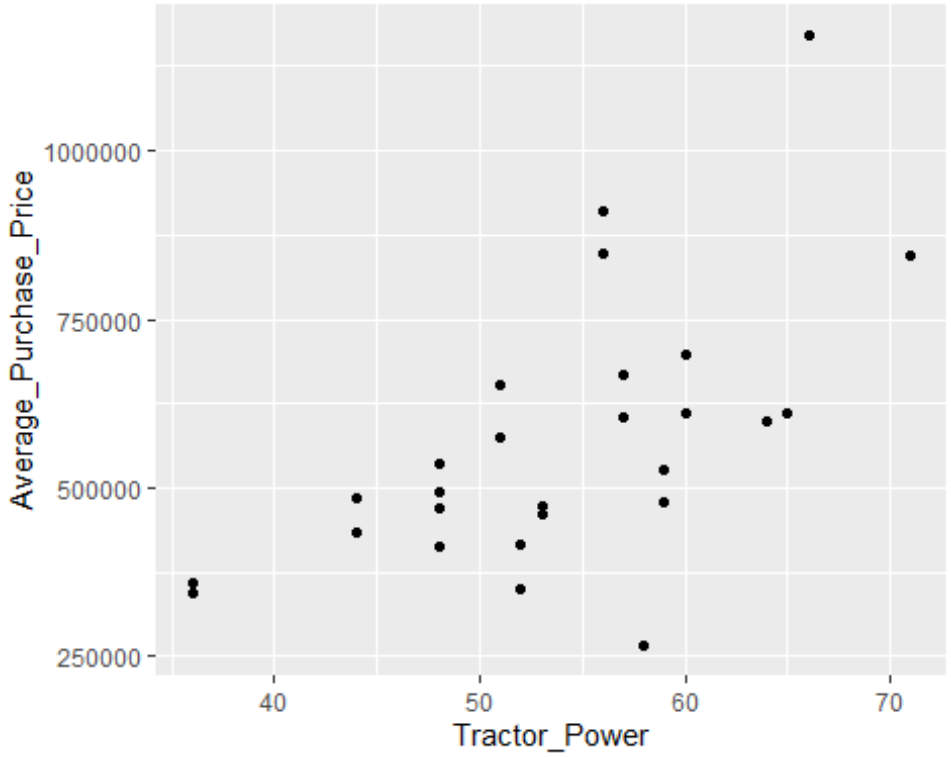
Histograms of Average Purchase Price and Tractor Power.

```
# Summary statistics

# Average Purchase Price
summary(Machinery$Average_Purchase_Price)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  266650  446975  526400  566553  631840 1171456

# Tractor Power
summary(Machinery$Tractor_Power)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   36.00   48.00   53.00   53.78   59.00   71.00
```

The Average Purchase Price has a distribution which is skewed to the right as seen in figure (a), which is supported by the symmetry statistics that show that the mean value is greater than the median value.The Average Purchase Price is mainly distributed between R266650 and R700000.

The distribution of the Tractor Power is approximately symmetrical as cann be seen in figure (b) and supported by symmetry statistics with the median and the mean tractor power having a difference of less than 1 Kilowatts. Mainly the Tractor Power lies around 53 Kilowatts.

```
library(ggplot2)
ggplot(Machinery, aes(Tractor_Power, Average_Purchase_Price)) + geom_point()
```




```
cor(Machinery$Tractor_Power, Machinery$Average_Purchase_Price)
```

```
## [1] 0.6061933
```

There is a positive, moderate linear relationship between the Average Purchase Price and the Tractor, with the lowest Tractor Power having the least Average Purchase Price and generally as the Tractor Power increases, the Average Selling Price increases too. This analysis of the relationship between Tractor Power and Average Purchase Price is supported by a correlation coefficient of 0.61.

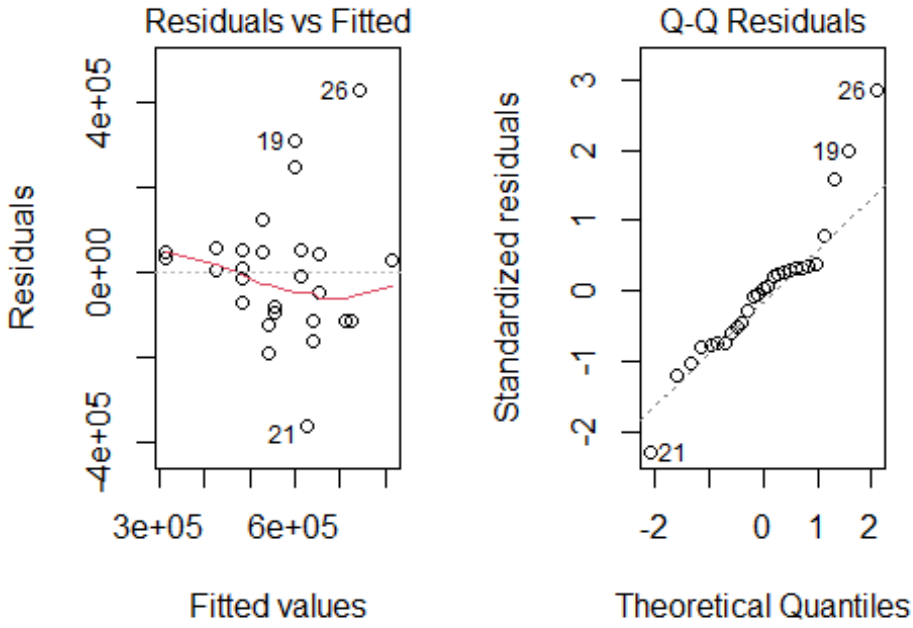
Question B

```
#Model fitting
model1 <- lm(Average_Purchase_Price ~ Tractor_Power, data = Machinery)

#Coefficients

coef(summary(model1))
```

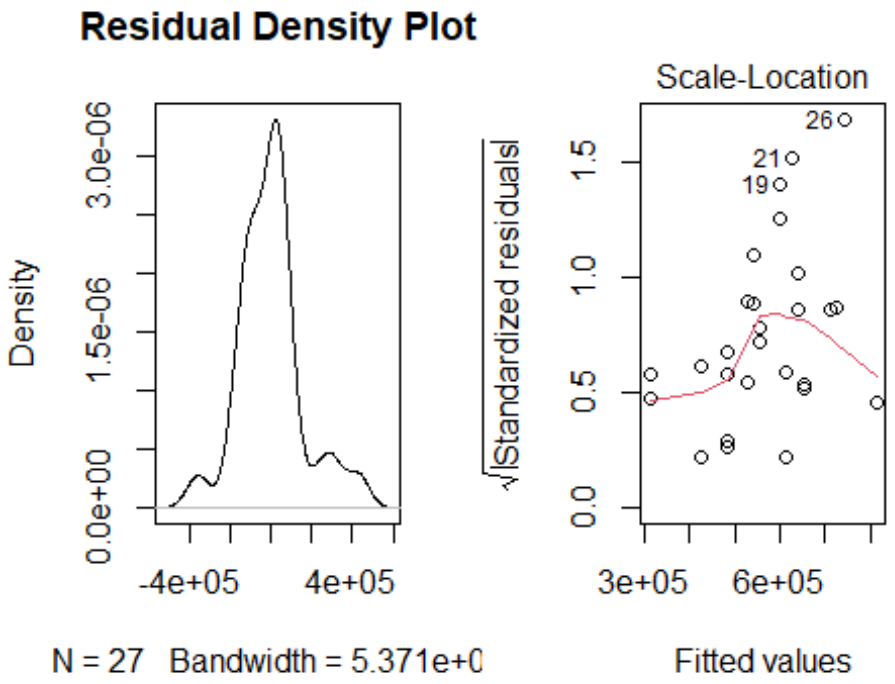
```
##                 Estimate Std. Error    t value    Pr(>|t|)
## (Intercept)   -203952.56  204542.59 -0.9971154 0.328262408
## Tractor_Power   14327.59    3759.52  3.8110150 0.000803672
```

The slope coefficient, $\hat{\beta}_1$, tells us that for every unit Kilowatt increase of Tractor Power, the Average Purchase Price increases by R14327.59.

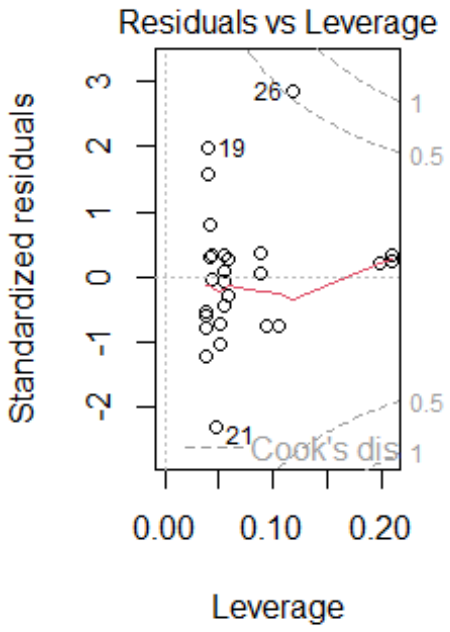
```r
par(mfrow=c(1,2))
plot(model1,which=1)
plot(model1,which=2)
```




```r
plot(density(residuals(model1)), main = "Residual Density Plot")
plot(model1,which=3)
```




```
plot(model1,which=5)
```



The assumption that the residuals of the linear regression are of equal variance is not met. It can be seen in the "Residuals vs Fitted" plot that the variance of the residuals increases as the fitted values increase, even though they are centered around mean 0.The Q-Q plot suggests that the residuals are not normally distributed as most points do not lie on the Q-Q line. A plot of the density of the residuals supports that the normal distribution assumption is not met as the shape is not smooth and generally bell shaped.There is a single potential influential point which can be observed in the northeast section of the "Residuals vs Leverage" plot.

Question C

```
Machinery$log_Average_Purchase_Price <- log(Machinery$Average_Purchase_Price)
model2 <- lm(log_Average_Purchase_Price ~ Tractor_Power, data = Machinery)

coef(summary(model2))
```

```
##                  Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)   11.91884732 0.339158430 35.142418 8.362159e-23
## Tractor_Power  0.02370773 0.006233778  3.803109 8.200380e-04
```

The estimated regression line equation is:

$$log(Y)_i = 11.9188 + 0.0237\text{Tractor\_Power}_i.$$

Question D

The log purchase price increases by R0.0237 for every 1 Kilowatt Tractor Power increase.

Question E

$$log(Y)_i = 11.9188 + 0.0237(1) = 11.9425$$

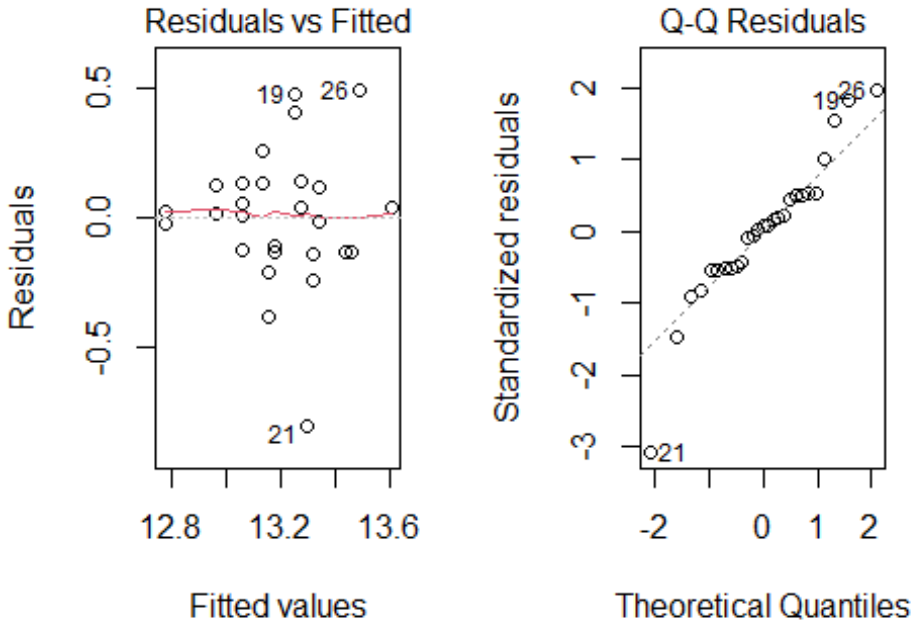
To get the change in Average_Purchase_Price for a unit change in Tractor_Power we need to compute the exponential of 11.9425. Average_Purchase_Price increases by R153660.4.

```
exp(11.9425)
```

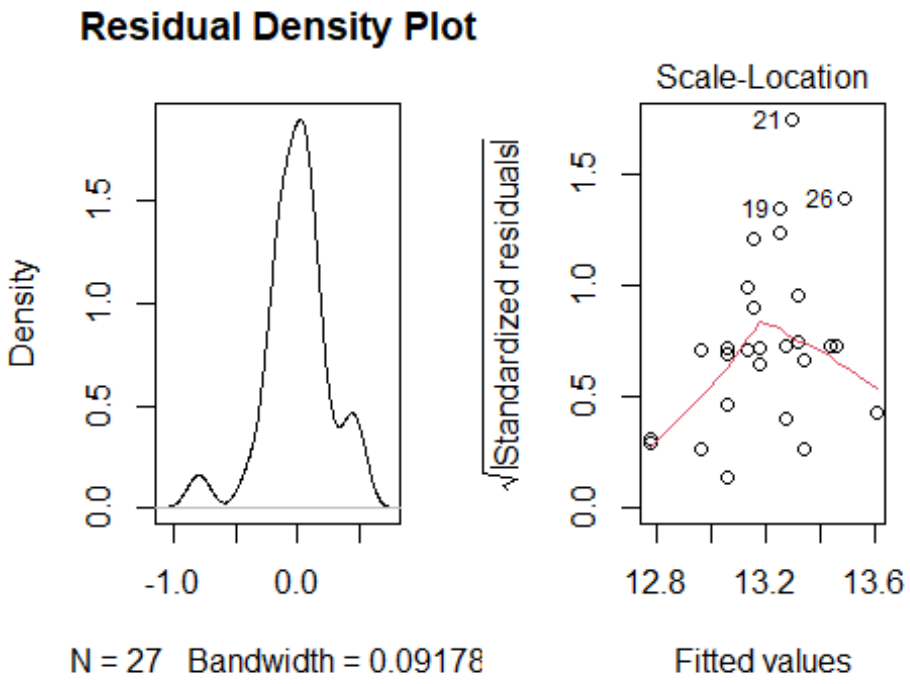
```
## [1] 153660.4
```

Question F

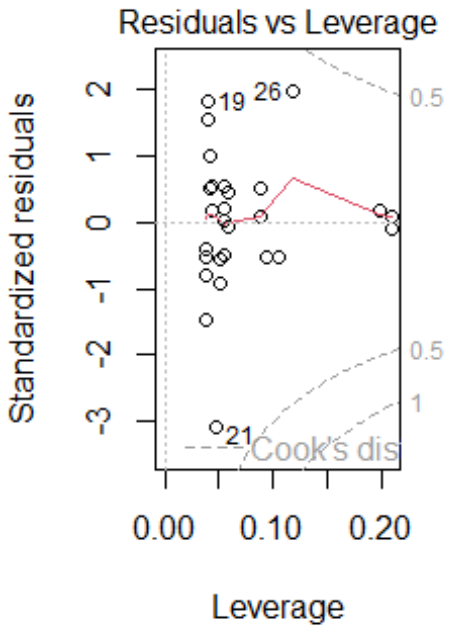
```r
par(mfrow=c(1,2))
plot(model2,which=1)
plot(model2,which=2)
```




```r
plot(density(residuals(model2)), main = "Residual Density Plot")
plot(model2,which=3)
```




```
plot(model2,which=5)
```



From the “Residuals vs Fitted” plot we can observe that the constant variance of residuals assumptions is met as there is no apparent pattern on the plot. The fitted red line on the same plot is approximately horizontal with no trends and so it can be interpreted that the linearity assumption of the model has been met. On the Q-Q plot, most residuals lie on the Q-Q line indicating that the residuals are approximately normally distributed. This assumption is cemented by the density plot of the residuals which has a general bell shape. The “Residuals vs Leverage” plot shows an absence of potential influential points..

Question G

```
model3 <- lm(Average_Purchase_Price ~ Tractor_Power + Type_Tractor, data = Ma
chinery)
coef(summary(model3))
```

```
##                     Estimate Std. Error    t value    Pr(>|t|)
## (Intercept)      -196060.95 207906.993 -0.9430224 0.355066303
## Tractor_Power      13774.94   3939.447  3.4966684 0.001857609
## Type_Tractor4 WD   36835.91  66126.564  0.5570517 0.582650358
```

The estimated regression line equation is:

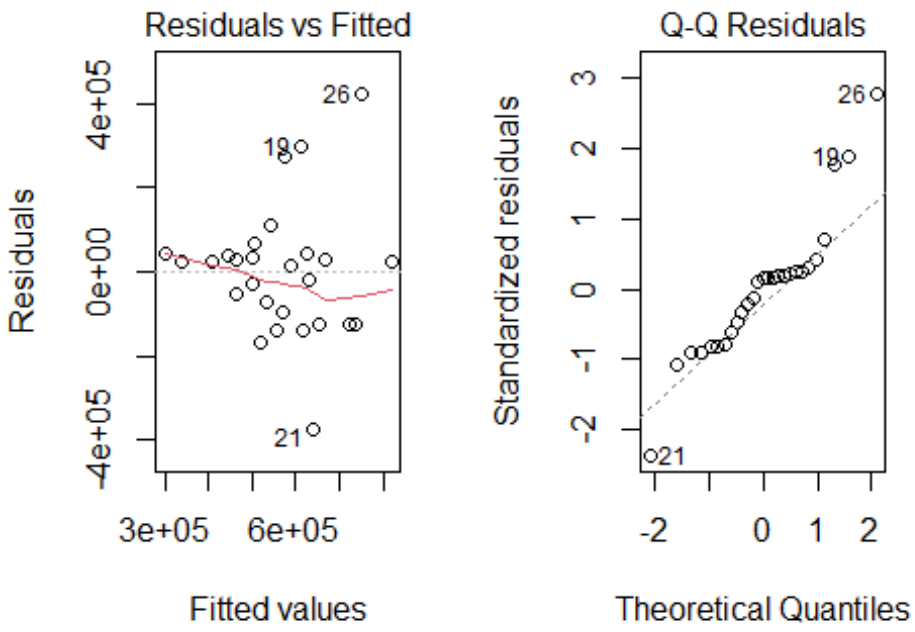
$$(Y)_i = -196060.95 + 13774.94\text{Tractor\_Power}_i + 36835.91\text{Type\_Tractor}_i$$

with $\text{Type\_Tractor}_i = 1$ when Type_Tractor = 4 WD, else $\text{Type\_Tractor}_i = 0$

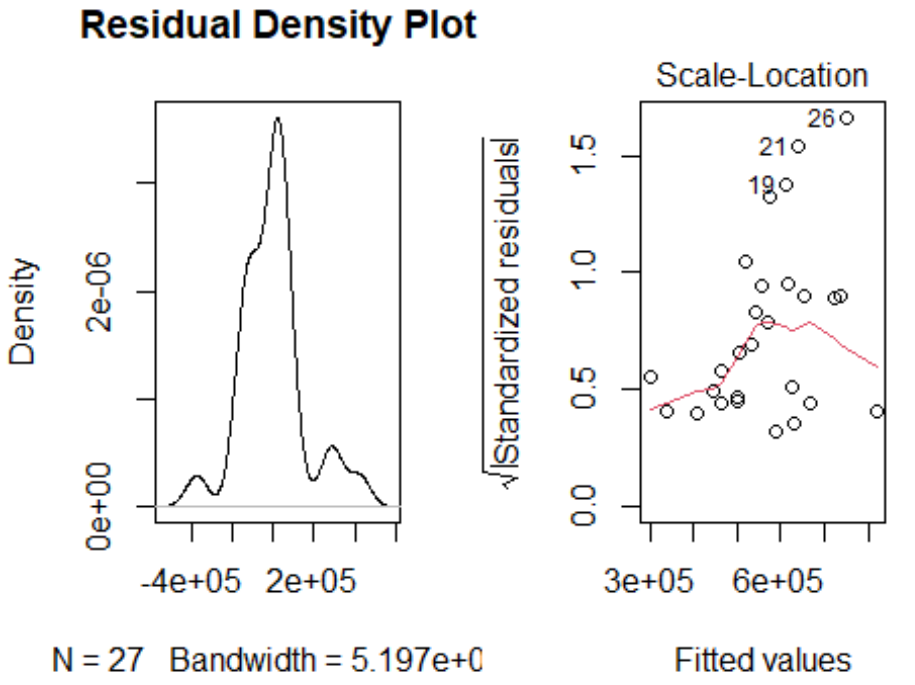
The coefficient for Tractor Power,$\hat{\beta}_1$ tells us that when other coefficients in the model are kept constant and Tractor Power increases by 1 Kilowatt then the Average Purchase price increases by R13774.94. The coefficient for Type Tractor, $\hat{\beta}_2$, tells us that when other coefficients in the model are kept constant, Average Selling Price increase by R36835.91 if and only if Type Tractor is 4D.

Question H

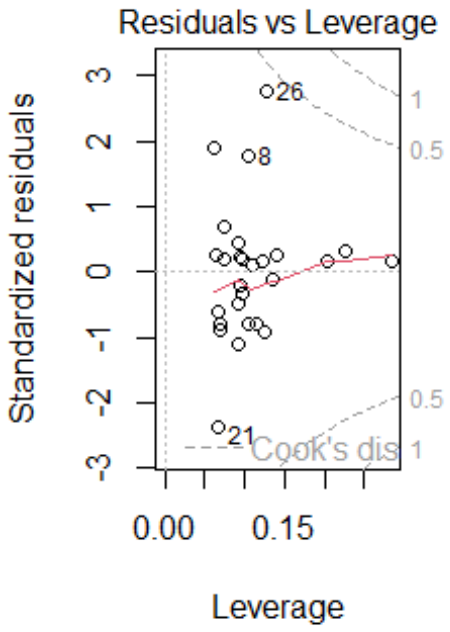
```
par(mfrow=c(1,2))
plot(model3,which=1)
plot(model3,which=2)
```




```
plot(density(residuals(model3)), main = "Residual Density Plot")
plot(model3,which=3)
```



```
plot(model3,which=5)
```



From the Residual vs Fitted plot it can be observed that the variance of te plotted residuals increase as the Fitted Values increase and so the constant variance of residuals assumption is not met. From observing the Q-Q plot it can be said that the assumption that the residulsare normally distributed is not met as most of the residuals do not lie on the diagonal line. No potential influential points are detected in the Residual vs Leverage plot.

Question I

```
col <- c(4,13,5,6,7,8,9,11)
cor(Machinery[,col])
```

```
##                          Average_Purchase_Price Fuel_Usage Salvage_Value
## Average_Purchase_Price                 1.0000000  0.6061933     1.0000000
## Fuel_Usage                             0.6061933  1.0000000     0.6061928
## Salvage_Value                          1.0000000  0.6061928     1.0000000
## Average_Investment                     1.0000000  0.6061932     1.0000000
## Depreciation_Costs                     1.0000000  0.6061943     1.0000000
## Insurance_Licence_Costs                0.9999970  0.6055121     0.9999970
## Interest_Costs                         1.0000000  0.6061642     1.0000000
## Repair_Maintenance_Costs               1.0000000  0.6061968     1.0000000
##                          Average_Investment Depreciation_Costs
## Average_Purchase_Price            1.0000000          1.0000000
## Fuel_Usage                        0.6061932          0.6061943
```

```
## Salvage_Value                      1.0000000          1.0000000
## Average_Investment                 1.0000000          1.0000000
## Depreciation_Costs                 1.0000000          1.0000000
## Insurance_Licence_Costs            0.9999970          0.9999969
## Interest_Costs                     1.0000000          1.0000000
## Repair_Maintenance_Costs           1.0000000          1.0000000
##                           Insurance_Licence_Costs Interest_Costs
## Average_Purchase_Price                  0.9999970      1.0000000
## Fuel_Usage                              0.6055121      0.6061642
## Salvage_Value                           0.9999970      1.0000000
## Average_Investment                      0.9999970      1.0000000
## Depreciation_Costs                      0.9999969      1.0000000
## Insurance_Licence_Costs                 1.0000000      0.9999971
## Interest_Costs                          0.9999971      1.0000000
## Repair_Maintenance_Costs                0.9999971      1.0000000
##                           Repair_Maintenance_Costs
## Average_Purchase_Price                   1.0000000
## Fuel_Usage                               0.6061968
## Salvage_Value                            1.0000000
## Average_Investment                       1.0000000
## Depreciation_Costs                       1.0000000
## Insurance_Licence_Costs                  0.9999971
## Interest_Costs                           1.0000000
## Repair_Maintenance_Costs                 1.0000000
```

From the correlation matrix we can see that there is a correlation of > 0.6 between all variables. To the model I would add the variable Fuel Uage as it has the smallest correlation to Average Purchase Price in the correlation matrix.