Stat201 Assignment 2

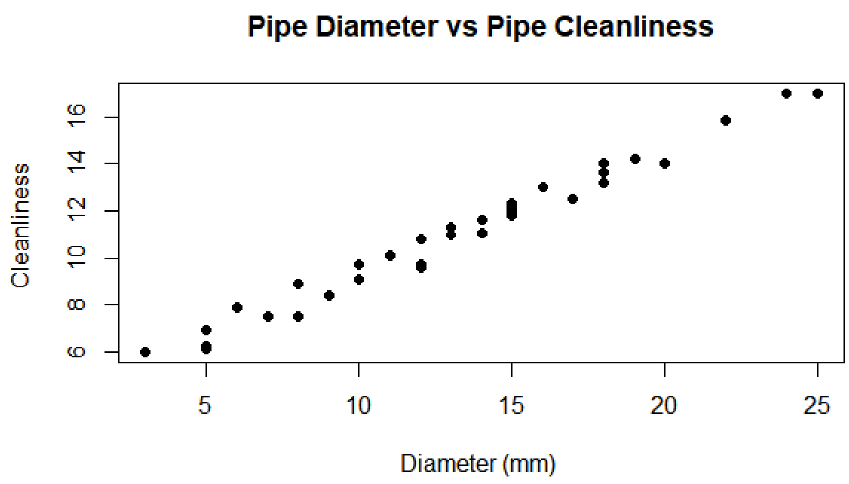
Robert Ivill 46012819

To begin this assignment, the 'pipe.csv' dataset was imported into RStudio and was confirmed to be read in correctly by using the head() and tail() functions to see the first and last 6 values. To discover the relationship between the pipe clean score and the diameter, a scatter plot was created using the plot() function and a histogram of pipe clean scores was done with the hist() function. The functions and plots of this assignment can be seen in the appendix of this report. In the scatter plot, we can see that there is a very strong positive relationship between the pipe diameter and clean score. In the histogram, it appears to be normally distributed with a heavy right skew. The next step was to create a linear regression model, which was done using the lm() function. Using summary(), the model was printed, and we can see the coefficients of the model. This shows that there is a moderately significant positive relationship between the diameter and cleanliness score, having a coefficient of 0.528 (3dp). Viewing the residual plots was done using the plot() function with the which arguments. The residuals vs predicted values model shows that it is a good model as there is a random pattern with no clear trend. The normal probability plot (Normal Q-Q) shows that the residuals fall along a straight line apart from the ends of the plot. The final plot is a plot of square roots of the residuals against the fitted values. We see that there is a constant spread of points with no clear pattern. To predict the pipe clean score for a pipe with a diameter of 10, we use the formula of the linear regression model. This gives us the formula Clean = 4.010 + 0.528 * Diameter. Substituting 10 for diameter gives us Clean = 4.01 + 0.528*10 = 4.01+5.28 = 9.29 approximately. To predict the scores for diameters of 10, 15 and 20 we use the predict() function. This gives fits of 9.29, 11.94 and 14.58 respectively and bounds of 8.39-10.19, 11.04-12.83 and 13.66-15.49.

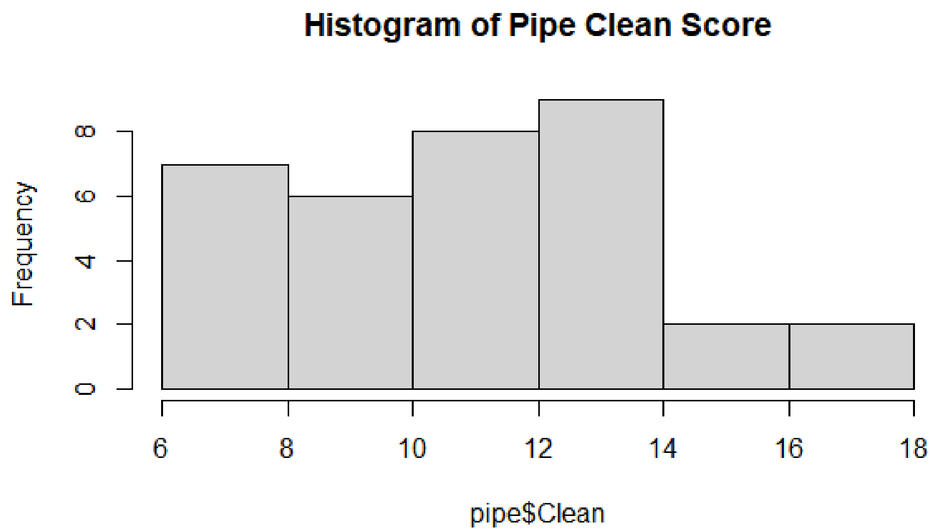
To check if the fish dataset was read in correctly, we use the head and tail functions again. Using the plot function, we create a scatter plot of the fish weight vs length and the hist function creates a histogram of the fish weight measurements. There is a clear positive correlation between fish length and weight, showing that larger fish tend to weigh more. The histogram shows that there is a large right skew in the data, with much more lighter fish being measured then larger ones. We fit a linear regression model with the lm() function and the summary function to show the results. The summary shows that the coefficient estimate for length is 35, showing that weight increases 35 grams for every cm increase in length of fish. The p-value for the coefficient is <2e-16, showing that fish length is a very meaningful predictor for fish weight as it is less that 0.05. The three residual plots were created using plot function using the which argument. The residuals vs fitted model for the most part has no clear trend, which is good for the model. The normal probability plot has the residuals along a straight line which is strong for the model. The cook's distance model shows that there are a couple of outliers at the ends of the model. We then fit a quadratic regression model to predict the fish weight vs fish length, again using the lm() function with weight being compared to length + I(length^2). The summary output has a coefficient for length that is negative and positive for the quadratic term, showing that the model is not simply linear. The p-value for the quadratic term is much less than 0.05, showing that the quadratic model is a pretty good fit. This suggests that we cannot reduce the model down to the linear model, so the quadratic term is important for predicting the model. We can look at the three residual plots for the quadratic model again using the plot function with which argument. The residual plots appear like the linear models plots. Although the new cook's distance model has a greater number of high data points, which may be affecting the quadratic model. After creating a plot with the linear and quadratic model using the lines() function. We can see that the quadratic model is preferred as it appears to fit the scatterplot better than the linear model. We can confirm this by comparing the R-squared values of both models, linear having a value of 0.92 versus quadratic's higher value of 0.97.

Appendix:

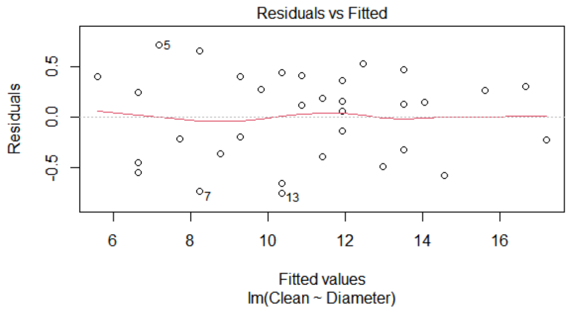
```
> head(pipe)
  Diameter Clean
1        3   6.0
2        5   6.1
3        5   6.9
4        5   6.2
5        6   7.9
6        7   7.5
> tail(pipe)
   Diameter Clean
29       18  13.2
30       19  14.2
31       20  14.0
32       22  15.9
33       24  17.0
34       25  17.0
> plot(pipe$Diameter, pipe$Clean, main="Pipe Diameter vs Pipe Cleanliness"
,
+      xlab="Diameter (mm) ", ylab="Cleanliness ", pch=19)
```

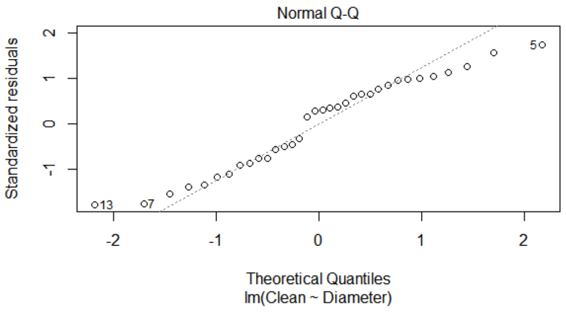


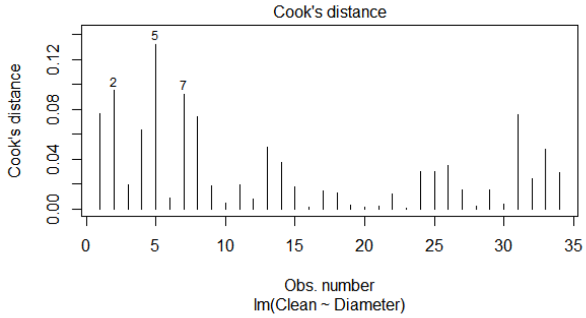

```
> hist(pipe$Clean, main = "Histogram of Pipe Clean Score")
```



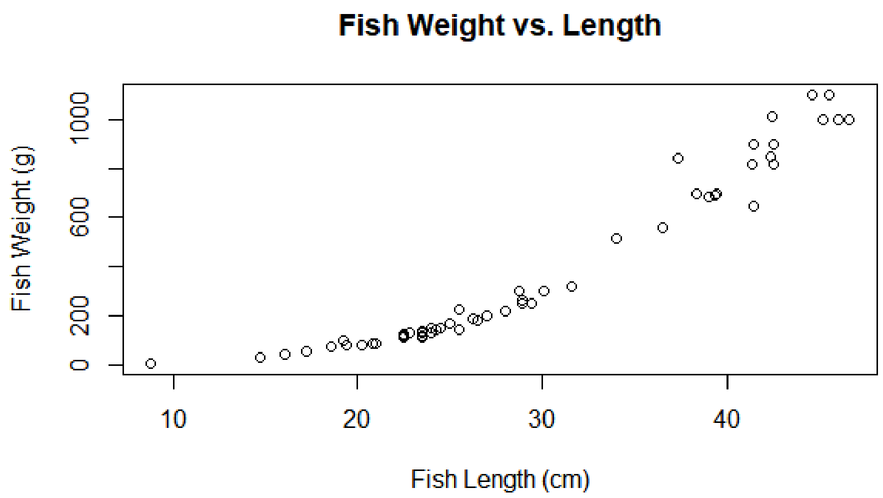

```
> pipe.lm1<-lm(Clean ~ Diameter, data = pipe)

> summary(pipe.lm1)

Call:
lm(formula = Clean ~ Diameter, data = pipe)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.7506 -0.3543  0.1250  0.3503  0.7197 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.00996    0.19331   20.74   <2e-16 ***
Diameter     0.52839    0.01343   39.35   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4328 on 32 degrees of freedom
Multiple R-squared:  0.9798,	Adjusted R-squared:  0.9791 
F-statistic:  1549 on 1 and 32 DF,  p-value: < 2.2e-16
> plot(pipe.lm1, which = 1) # Residuals vs Fitted Model
> plot(pipe.lm1, which = 2) # Normal Q-Q model
> plot(pipe.lm1, which = 4) # Cook's distance model
```



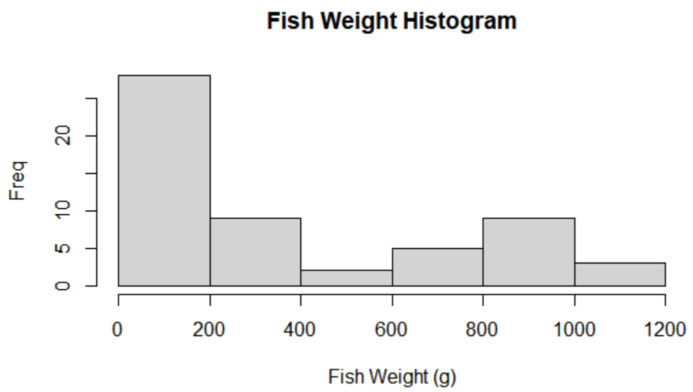




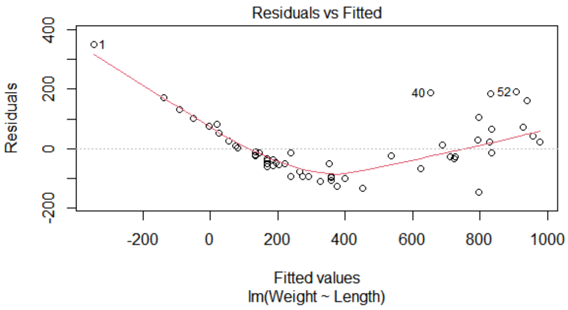

```
> predict(pipe.lm1, pred, interval = "prediction")
        fit       lwr      upr
1  9.293841  8.394858 10.19282
2 11.935779 11.040106 12.83145
3 14.577717 13.664652 15.49078


> head(fish)
  Weight Length
1    5.9    8.8
2   32.0   14.7
3   40.0   16.0
4   51.5   17.2
5   70.0   18.5
6  100.0   19.2
> tail(fish)
   Weight Length
51    820   42.5
52   1100   44.6
53   1000   45.2
54   1100   45.5
55   1000   46.0
56   1000   46.6
> plot(Weight ~ Length, data = fish, main = "Fish Weight vs. Length",
+      xlab = "Fish Length (cm)", ylab = "Fish Weight (g)")
```

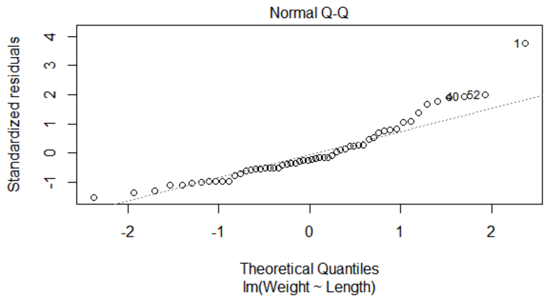


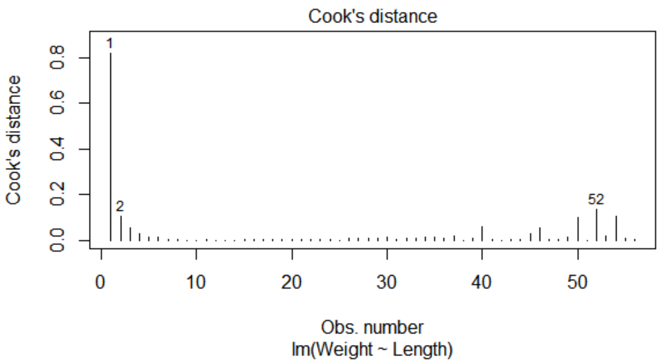

```
> hist(fish$Weight, main = "Fish Weight Histogram", xlab = "Fish Weight (g
)",
+      ylab = "Freq")
```



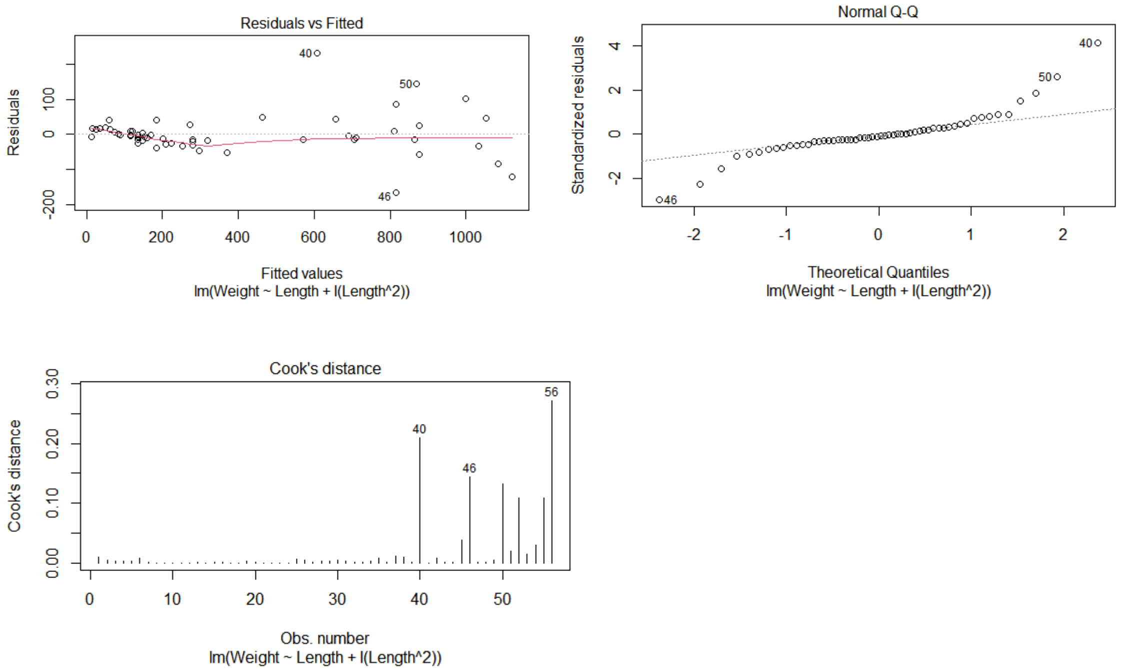

```
> fish.lm1 <- lm(Weight ~ Length, data = fish)
> summary(fish.lm1)

Call:
lm(formula = Weight ~ Length, data = fish)

Residuals:
    Min      1Q  Median      3Q     Max 
-146.25  -57.86  -23.99   45.00  350.68 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -652.787     43.407  -15.04   <2e-16 ***
Length        35.001      1.398   25.03   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 98.82 on 54 degrees of freedom
Multiple R-squared:  0.9207,	Adjusted R-squared:  0.9192 
F-statistic: 626.5 on 1 and 54 DF,  p-value: < 2.2e-16
> plot(fish.lm1, which = 1)
> plot(fish.lm1, which = 2)
> plot(fish.lm1, which = 4)
```








```
> fish.lm2<-lm(Weight~Length+I(Length^2), data = fish)
> summary(fish.lm2)

Call:
lm(formula = Weight ~ Length + I(Length^2), data = fish)

Residuals:
     Min       1Q   Median       3Q      Max 
-165.295  -18.970   -5.747   15.652  231.691 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 128.34533   78.77870   1.629  0.10920    
Length      -21.02388    5.41770  -3.881  0.00029 ***
I(Length^2)   0.90862    0.08689  10.458 1.72e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 56.99 on 53 degrees of freedom
Multiple R-squared:  0.9741,	Adjusted R-squared:  0.9731 
F-statistic: 996.6 on 2 and 53 DF,  p-value: < 2.2e-16


> plot(fish.lm2, which = 1)
> plot(fish.lm2, which = 2)
> plot(fish.lm2, which = 4)
```




```
> x<-order(fish$Length)
> plot(Weight~Length, data = fish)
> lines(fish$Length[x], fitted(fish.lm1)[x], lwd=2, col="green")
> lines(fish$Length[x], fitted(fish.lm2)[x], lwd=2, col="blue")
```