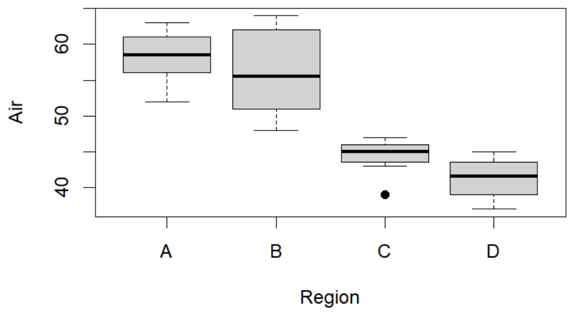
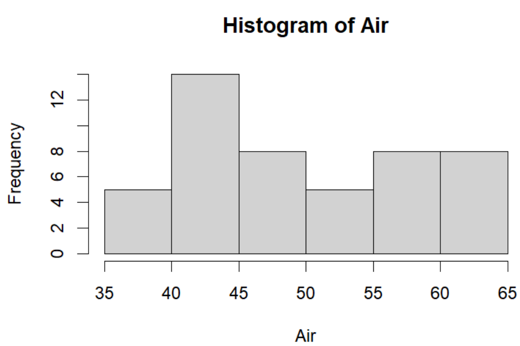
Statistics201 Assignment 4

1a)



This boxplot graph shows clear differences between the explanatory variables “A, B, C, D” and the response variable “Air”. For variables A and B, the median sits higher on the air pollution scale in comparison to the variables C and D, because of this I predict that the C and D regions will have less air pollution than the A and B variables. This boxplot graph also shows that B has more spread than A, C, and D seen by the length of the box and lower upper quartile range sticks. A highlighted point below the box for the C variable shows an outlier.



The histogram of the response variables gives an approximately normal distribution.

1b)

Assuming there’s no statistical differences in the region variables, the null and alternative hypothesis is as follows

$H0{:}\ u1 = u2 = u3$

$H1{:}\ One\ or\ more\ region\ means\ are\ different$

```
Analysis of Variance Table

Response: Air
          Df  Sum Sq Mean Sq F value    Pr(>F)    
Region     3 2512.50   837.5  60.692 1.118e-15 ***
Residuals 44  607.17    13.8                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

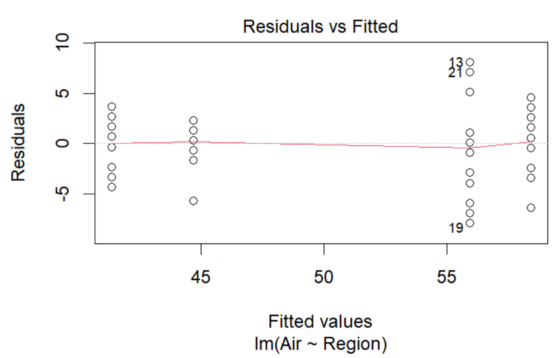
The ANOVA table for the constructed linear model outputs these results. With an observed significance level of 1.118e-15 in the Pr(>F) and an F value of 60.7 which is rather high, I can confidently reject the Null hypothesis and conclude that there’s a difference between the means of the region variables and that there’s a statistically significant difference between the region variables.

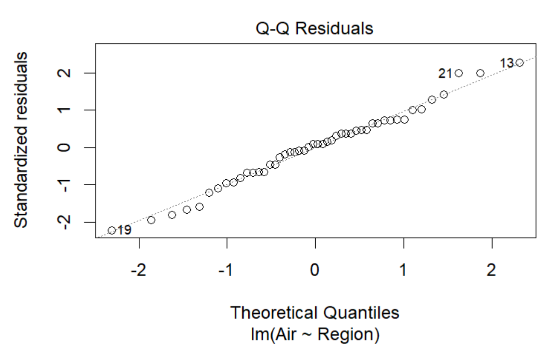
```
Call:
lm(formula = Air ~ Region, data = pollution)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.9167 -2.3542  0.3333  2.3333  8.0833 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   58.417      1.072  54.475  < 2e-16 ***
RegionB       -2.500      1.517  -1.648    0.106    
RegionC      -13.750      1.517  -9.067 1.25e-11 ***
RegionD      -17.083      1.517 -11.265 1.50e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.715 on 44 degrees of freedom
Multiple R-squared:  0.8054,	Adjusted R-squared:  0.7921 
F-statistic: 60.69 on 3 and 44 DF,  p-value: 1.118e-15
```

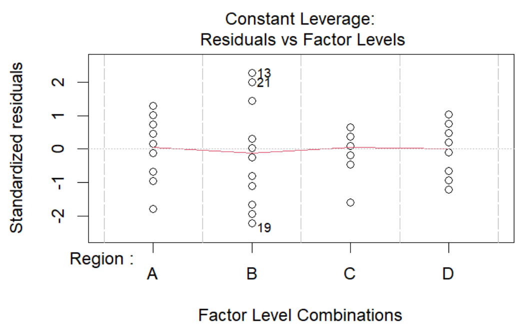
The summary data for these variables is interesting as every Region variable is significant in the linear model apart from Region B, this means that Region B isn’t statistically different from Region A.

1c)

Below are the residual graphs for the model.







The residuals vs fitted plot provides scatter randomly around 0 with no clear pattern and the trend line is a straight line meaning there's linearity and that the linear model is a good fit.

The Q-Q residual graph shows that the majority of the residuals are normally distributed with some minor or heavy outlier deviation at the extremes (13, 19, 21).

The residuals vs factor levels graph shows that for all variables, the residuals are roughly evenly spread allowing the assumption of constant variance for every Region.

1d)

```
> summary(pollution.lm2)
            Df Sum Sq Mean Sq F value   Pr(>F)    
Region       3 2512.5   837.5   60.69 1.12e-15 ***
Residuals   44  607.2    13.8                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(pollution.lm1)
Analysis of Variance Table

Response: Air
          Df  Sum Sq Mean Sq F value    Pr(>F)    
Region     3 2512.50   837.5  60.692 1.118e-15 ***
Residuals 44  607.17    13.8                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Comparing the models, the numbers have changed but despite the minor difference, it's still okay to conduct a Tukey's honest significant difference test on this data.

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Air ~ Region, data = pollution)

$Region
          diff        lwr         upr     p adj
B-A  -2.500000  -6.549153   1.5491529 0.3628245
C-A -13.750000 -17.799153  -9.7008471 0.0000000
D-A -17.083333 -21.132486 -13.0341804 0.0000000
C-B -11.250000 -15.299153  -7.2008471 0.0000000
D-B -14.583333 -18.632486 -10.5341804 0.0000000
D-C  -3.333333  -7.382486   0.7158196 0.1397001
```

Looking at the Tukey Honest Test output, the variables (B, A) and (D, C) aren't significantly different meaning there's no statistical evidence to suggest differences between variables B and A and variables D and C. All other variables have a significant p-value.

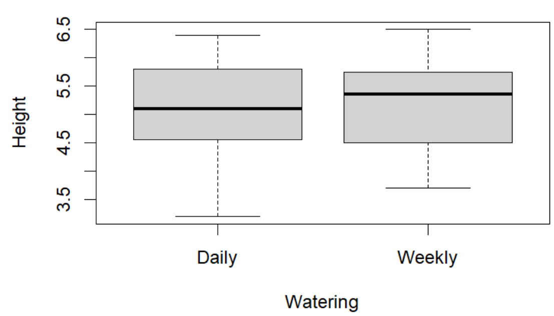
Regions C and D both have significantly lower air pollution levels than A and B.
Region C has a 13.75 unit lower air pollution level than Region A
Region D has a 17.08 unit decrease in air pollution level than Region A
Region C has a 11.25 unit lower air pollution level than Region B
Region D has a 14.58 unit lower air pollution than Region B

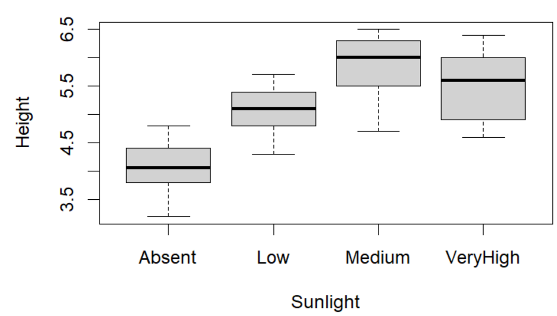
2a)

When analyzing the data these boxplots are useful for examining whether sunlight and water frequency influence grass height. The median height varies slightly when grass is watered daily or frequently but the difference is not substantial.

In the second graph, another boxplot helps determine whether sunlight affects grass height. Based on the graph it appears that a moderate amount of sunlight results in the tallest grass.

The ordering of the boxplot variables is important as it distinguishes between explanatory and response variables. For example, plotting water frequency against grass height makes sense as it shows how watering affects growth. However reversing the variables would be illogical as grass height does not determine how frequently it's watered.





2b)

```
Call:
lm(formula = Height ~ Sunlight * Watering, data = plants)

Residuals:
   Min     1Q Median     3Q    Max 
 -1.02  -0.27   0.08   0.31   0.68 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     4.1400     0.2330  17.771  < 2e-16 ***
SunlightLow                     0.8400     0.3295   2.550   0.0158 *  
SunlightMedium                  1.5800     0.3295   4.796 3.60e-05 ***
SunlightVeryHigh                1.6400     0.3295   4.978 2.12e-05 ***
WateringWeekly                 -0.1400     0.3295  -0.425   0.6737    
SunlightLow:WateringWeekly      0.3800     0.4659   0.816   0.4208    
SunlightMedium:WateringWeekly   0.4800     0.4659   1.030   0.3107    
SunlightVeryHigh:WateringWeekly -0.3200     0.4659  -0.687   0.4972    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5209 on 32 degrees of freedom
Multiple R-squared:  0.6949,	Adjusted R-squared:  0.6281 
F-statistic: 10.41 on 7 and 32 DF,  p-value: 9.669e-07
```

The summary of the linear model provides insight. It indicates that when the grass receives low, medium, or very high sunlight and is watered weekly, the linear equation is not statistically significant. Overall the p-value of the linear model is 9.6693-07 meaning that the model is statistically significant.

```
Analysis of Variance Table

Response: Height
                  Df  Sum Sq Mean Sq F value    Pr(>F)    
Sunlight           3 18.7648  6.2549 23.0490 3.898e-08 ***
Watering           1  0.0002  0.0002  0.0009    0.9760    
Sunlight:Watering  3  1.0107  0.3369  1.2415    0.3109    
Residuals         32  8.6840  0.2714                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This ANOVA table for the linear model indicates that the interaction between sunlight and watering frequency isn’t significant, regardless of whether the grass is watered weekly or daily. As a result the model can be simplified by removing this interaction term.

‘

2c)

```
Call:
lm(formula = Height ~ Sunlight + Watering, data = plants)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.1925 -0.2988  0.0500  0.3287  0.8475 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)        4.0725     0.1861  21.886  < 2e-16 ***
SunlightLow        1.0300     0.2354   4.376 0.000104 ***
SunlightMedium     1.8200     0.2354   7.733 4.45e-09 ***
SunlightVeryHigh   1.4800     0.2354   6.288 3.23e-07 ***
WateringWeekly    -0.0050     0.1664  -0.030 0.976204    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5263 on 35 degrees of freedom
Multiple R-squared:  0.6594,	Adjusted R-squared:  0.6204 
F-statistic: 16.94 on 4 and 35 DF,  p-value: 8.196e-08
```

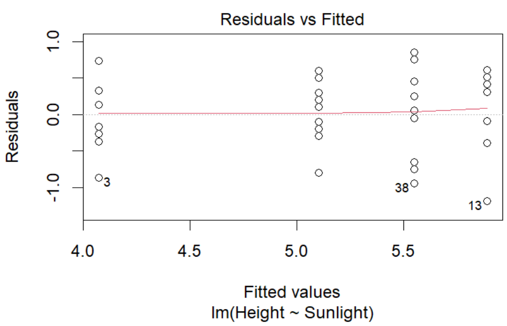
After simplifying the model, all variables except watering weekly are significant. Therefore the model can be further refined by removing the watering weekly variable.

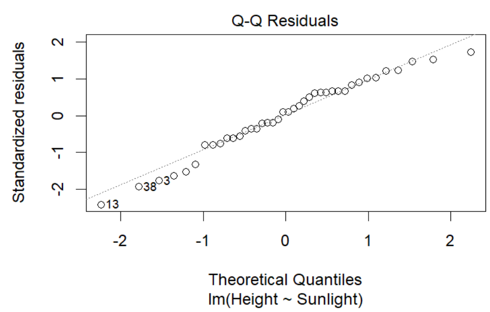
2d)

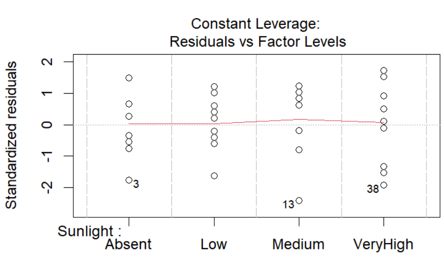
```
Call:
lm(formula = Height ~ Sunlight, data = plants)

Residuals:
   Min     1Q Median     3Q    Max 
 -1.19  -0.30   0.05   0.33   0.85 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)        4.0700     0.1641  24.801  < 2e-16 ***
SunlightLow        1.0300     0.2321   4.438 8.24e-05 ***
SunlightMedium     1.8200     0.2321   7.842 2.67e-09 ***
SunlightVeryHigh   1.4800     0.2321   6.377 2.18e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5189 on 36 degrees of freedom
Multiple R-squared:  0.6593,	Adjusted R-squared:  0.631 
F-statistic: 23.23 on 3 and 36 DF,  p-value: 1.535e-08
```

In this reduced model without the interaction between watering and sunlight and taking out water frequency, I have achieved a parsimonious, simplified model with a significance value of 1.535e-08.

2e)







These diagnostic plots assess the assumptions of the linear model Height ~ Sunlight. The residuals vs. fitted plot shows that residuals are randomly scattered around zero with no clear pattern indicating that the linearity assumption is met. The red trend line remains relatively flat suggesting no major violation though some potential outliers (such as points 3, 38, and 13) stand out. The Q-Q plot shows that most residuals align well with the diagonal reference line confirming approximate normality though slight deviations at the extremes suggest minor outliers. The residuals vs. factor levels plot indicates that the spread of residuals is fairly consistent across sunlight categories.

From the model I can conclude that the grass height isn't affected by how regularly it's watered, instead the sunlight exposure is a significant factor in the height of the grass.

2f)

```
Analysis of Variance Table

Response: Height
          Df Sum Sq Mean Sq F value    Pr(>F)
Sunlight   3 18.765  6.2549  23.226 1.535e-08 ***
Residuals 36  9.695  0.2693
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Height ~ Sunlight, data = plants)

$Sunlight
                 diff        lwr       upr     p adj
Low-Absent       1.03  0.4049559 1.6550441 0.0004607
Medium-Absent    1.82  1.1949559 2.4450441 0.0000000
VeryHigh-Absent  1.48  0.8549559 2.1050441 0.0000013
Medium-Low       0.79  0.1649559 1.4150441 0.0085384
VeryHigh-Low     0.45 -0.1750441 1.0750441 0.2302229
VeryHigh-Medium -0.34 -0.9650441 0.2850441 0.4685780
```

The Tukey Honest Significant Difference (HSD) test results indicate that grass height varies significantly across different sunlight levels, but not all levels are distinct from each other. Grass grown in absent light conditions is significantly shorter than grass grown under low, medium, and very high light levels. Similarly, grass in medium light conditions grows significantly taller than those in low light. However, the difference in grass height between very high and low light, as well as between very high and medium light, is not statistically significant. This suggests that while some light levels have a clear impact on grass height, very high light does not result in significantly different growth compared to medium or low light conditions.