1) Examine the data for ***sulfate_reduction.xlsx*** from tutorial 4 and consider your scatterplot. Is it reasonable to use a linear correlation model? Or is it necessary to do a transformation? Show your graph to justify your answer with an appropriate caption (**1 point**).

Q1 0.8

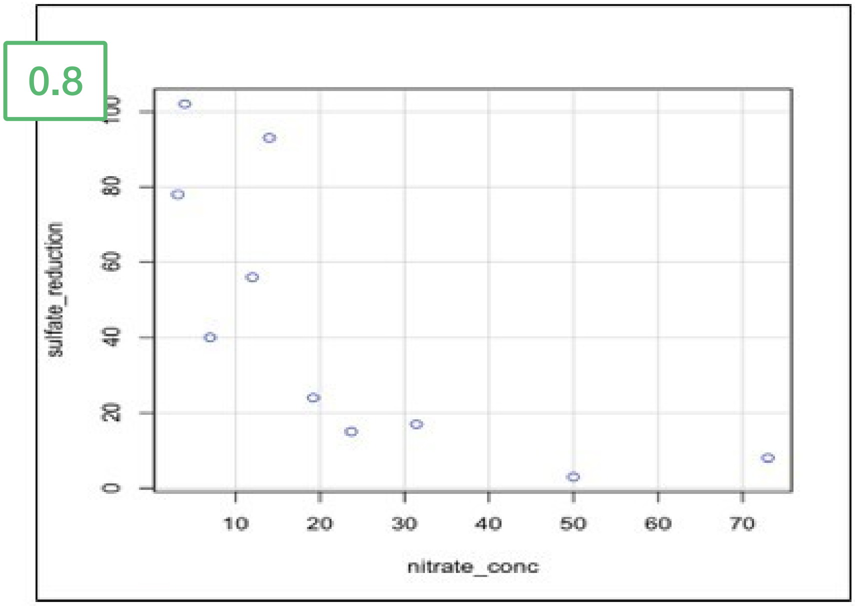
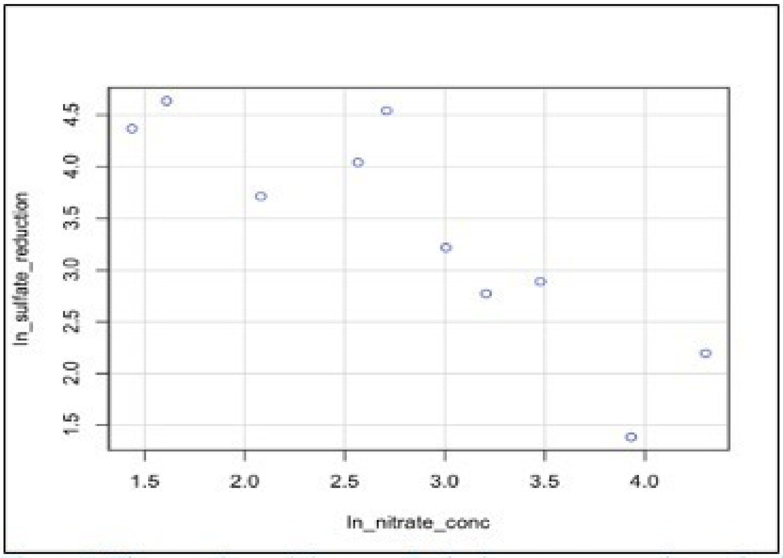


Figure 1. The scatterplot model of sulfate reduction as nitrate concentration increases. The model shows a negative curvilinear relationship between the sulfate reduction and nitrate concentration. Logarithmic transformation is necessary to make this distribution less skewed.

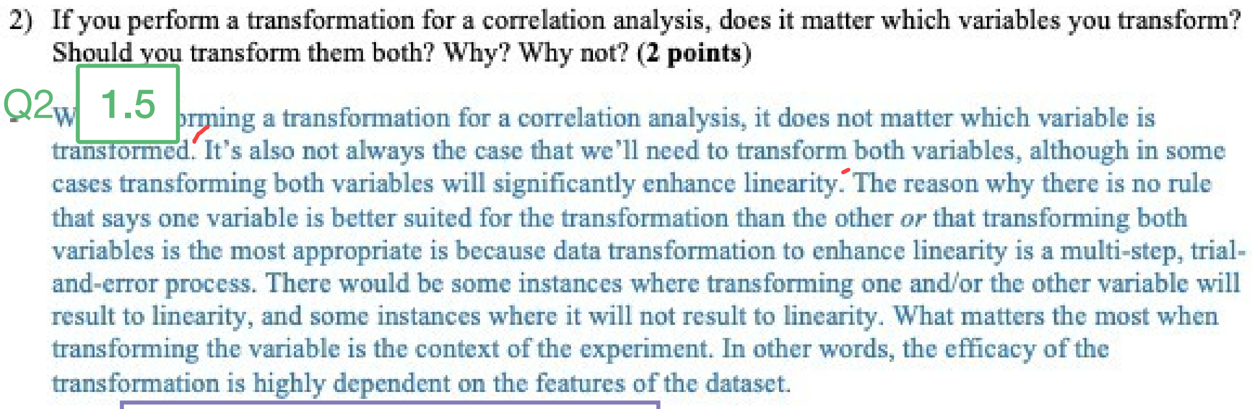
It's necessary to do a logarithmic transformation of both variables because of the negative curvilinear relationship between the sulfate reduction and the nitrate concentration. By doing the transformation, it helps make the data meet the assumptions of inferential statistics.



Remember to also try transforming one variable at a time (eg. ln(sulfate) and untransformed nitrate)

Try tiling the graphs to make a single figure

Figure 2. The transformed datasets. Both nitrate concentration and sulfate reduction were transformed using logarithmic transformation to enhance linearity of the graph.

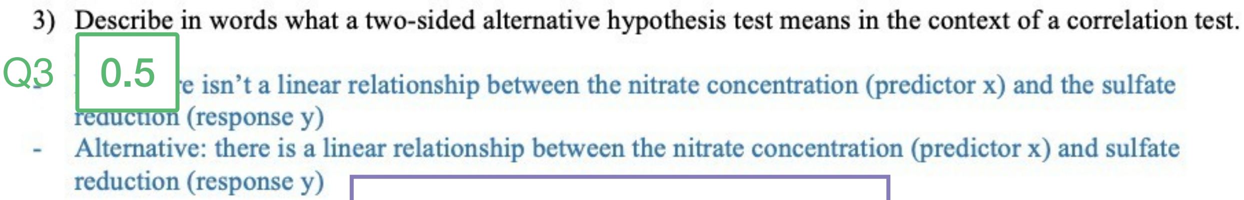


2) If you perform a transformation for a correlation analysis, does it matter which variables you transform? Should you transform them both? Why? Why not? (**2 points**)

Q2 1.5

W[illegible]orming a transformation for a correlation analysis, it does not matter which variable is transformed. It's also not always the case that we'll need to transform both variables, although in some cases transforming both variables will significantly enhance linearity. The reason why there is no rule that says one variable is better suited for the transformation than the other *or* that transforming both variables is the most appropriate is because data transformation to enhance linearity is a multi-step, trial-and-error process. There would be some instances where transforming one and/or the other variable will result to linearity, and some instances where it will not result to linearity. What matters the most when transforming the variable is the context of the experiment. In other words, the efficacy of the transformation is highly dependent on the features of the dataset.

What are some downsides of transforming both variables?



3) Describe in words what a two-sided alternative hypothesis test means in the context of a correlation test.

Q3 0.5

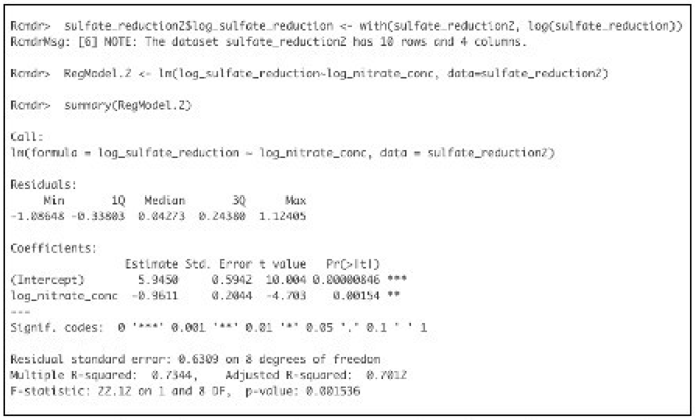
- [illegible]re isn’t a linear relationship between the nitrate concentration (predictor x) and the sulfate reduction (response y)
- Alternative: there is a linear relationship between the nitrate concentration (predictor x) and sulfate reduction (response y)

What does the alternative hypothesis mean for Pearson's correlation coefficient?

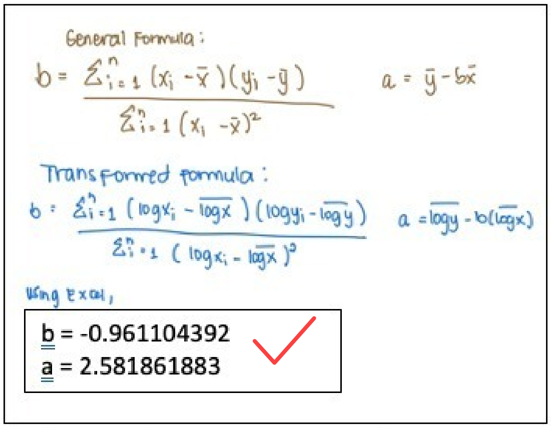
After doing transformation trial-and-error, it was identified that transforming both variables via logarithmic transformation will result to the most linear graph. This transformation is also identified to be the most ideal when looking at the normality of the distribution when looking at the linear regression table. It's this transformation that had the almost equal min and max linear regression (left pic).

Q4 2

So, the formula to calculate the regression intercept will be what is shown on the right picture.
After calculation, the intercept of a regression to predict sulfate reduction is found to be 2.58. The R^2 on Excel is 0.7344, and it matches the R^2 value in R. The slope from the manual calculation is -0.9611, and it matches the log(nitrate concentration) on R. So we know that the transformed formula that was used is accurate.



```
Rcmdr>  sulfate_reduction2$log_sulfate_reduction <- with(sulfate_reduction2, log(sulfate_reduction))
RcmdrMsg: [6] NOTE: The dataset sulfate_reduction2 has 10 rows and 4 columns.

Rcmdr>  RegModel.2 <- lm(log_sulfate_reduction~log_nitrate_conc, data=sulfate_reduction2)

Rcmdr>  summary(RegModel.2)

Call:
lm(formula = log_sulfate_reduction ~ log_nitrate_conc, data = sulfate_reduction2)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.08648 -0.33803  0.04273  0.24380  1.12405 

Coefficients:
                 Estimate Std. Error t value   Pr(>|t|)    
(Intercept)        5.9450     0.5942  10.004 0.00000846 ***
log_nitrate_conc  -0.9611     0.2044  -4.703    0.00154 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6309 on 8 degrees of freedom
Multiple R-squared:  0.7344,	Adjusted R-squared:  0.7012 
F-statistic: 22.12 on 1 and 8 DF,  p-value: 0.001536
```



General Formula :

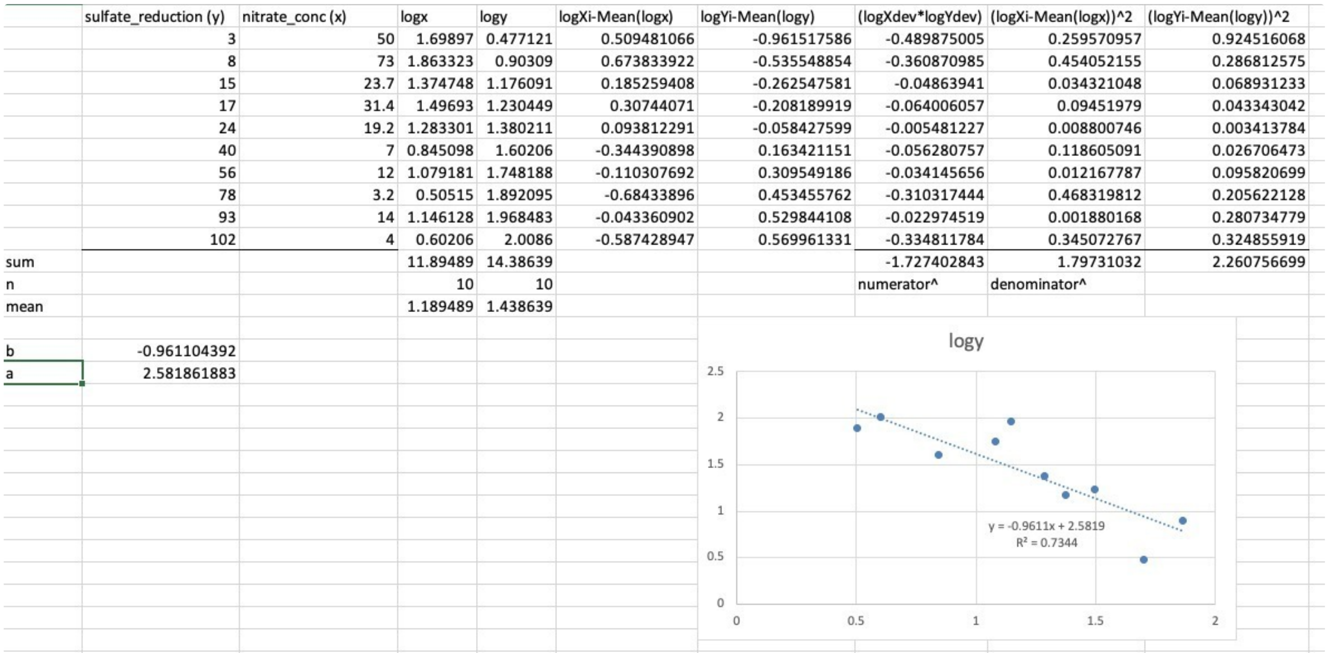
$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad a = \bar{y} - b\bar{x}$$

Transformed formula :

$$b = \frac{\sum_{i=1}^{n} (\log x_i - \overline{\log x})(\log y_i - \overline{\log y})}{\sum_{i=1}^{n} (\log x_i - \overline{\log x})^2} \qquad a = \overline{\log y} - b(\overline{\log x})$$

Using Excel,

b = -0.961104392
a = 2.581861883 ✓

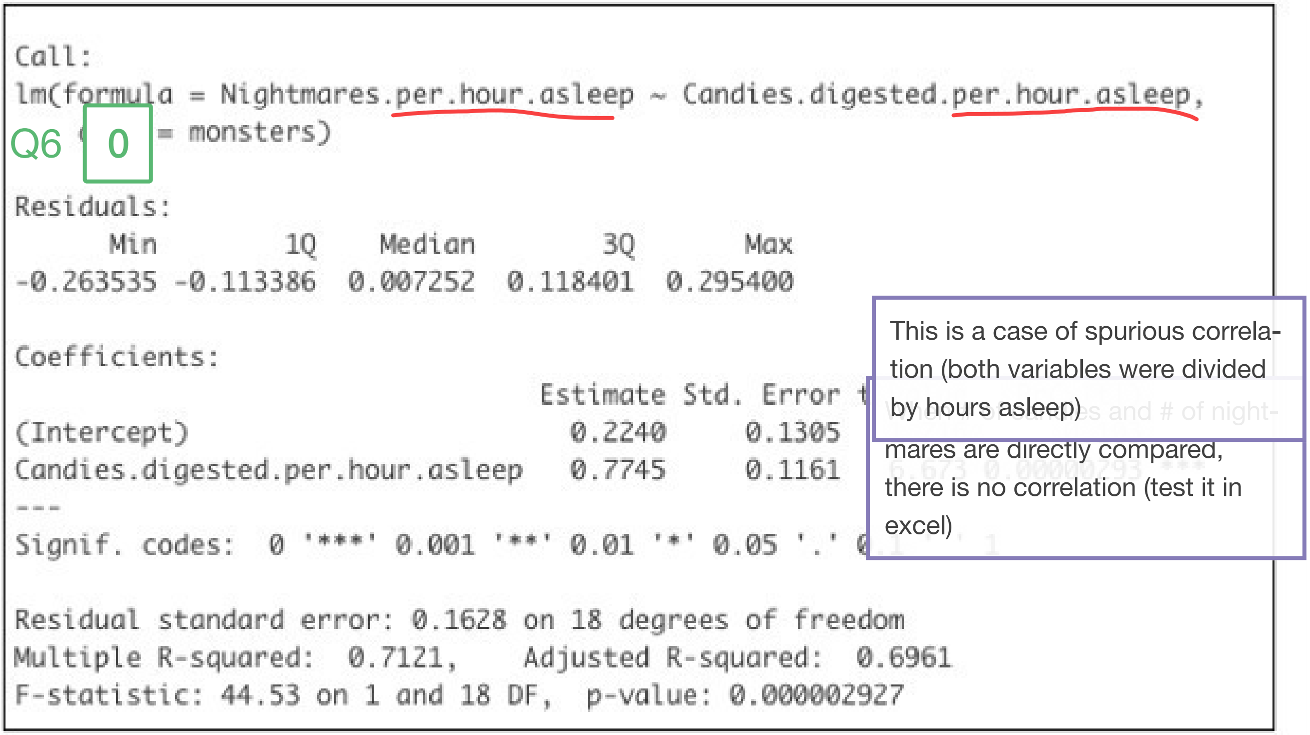


	sulfate_reduction (y)	nitrate_conc (x)	logx	logy	logXi-Mean(logx)	logYi-Mean(logy)	(logXdev*logYdev)	(logXi-Mean(logx))^2	(logYi-Mean(logy))^2
	3	50	1.69897	0.477121	0.509481066	-0.961517586	-0.489875005	0.259570957	0.924516068
	8	73	1.863323	0.90309	0.673833922	-0.535548854	-0.360870985	0.454052155	0.286812575
	15	23.7	1.374748	1.176091	0.185259408	-0.262547581	-0.04863941	0.034321048	0.068931233
	17	31.4	1.49693	1.230449	0.30744071	-0.208189919	-0.064006057	0.09451979	0.043343042
	24	19.2	1.283301	1.380211	0.093812291	-0.058427599	-0.005481227	0.008800746	0.003413784
	40	7	0.845098	1.60206	-0.344390898	0.163421151	-0.056280757	0.118605091	0.026706473
	56	12	1.079181	1.748188	-0.110307692	0.309549186	-0.034145656	0.012167787	0.095820699
	78	3.2	0.50515	1.892095	-0.68433896	0.453455762	-0.310317444	0.468319812	0.205622128
	93	14	1.146128	1.968483	-0.043360902	0.529844108	-0.022974519	0.001880168	0.280734779
	102	4	0.60206	2.0086	-0.587428947	0.569961331	-0.334811784	0.345072767	0.324855919
sum			11.89489	14.38639			-1.727402843	1.79731032	2.260756699
n			10	10			numerator^	denominator^	
mean			1.189489	1.438639					
b	-0.961104392								
a	2.581861883								

Q5 0.75

5) R Adjusted R-squared and describe what it means. (**1 point**)

- T d R-squared value is 0.7012. This means that 70.12% of the total variation in sulfate reduction is accounted for by the regression model. ✓

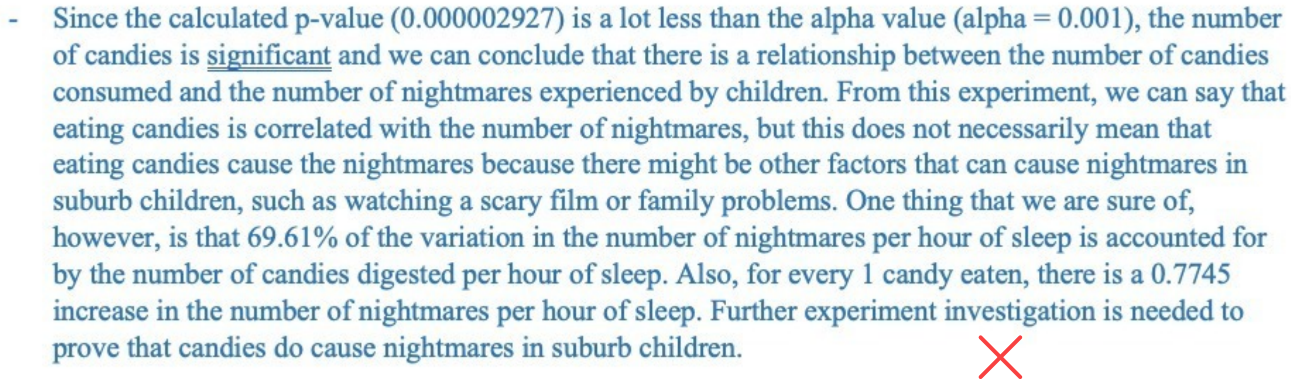


Q6 0

```
Call:
lm(formula = Nightmares.per.hour.asleep ~ Candies.digested.per.hour.asleep,
         = monsters)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.263535 -0.113386  0.007252  0.118401  0.295400 

Coefficients:
                                 Estimate Std. Error t
(Intercept)                        0.2240     0.1305
Candies.digested.per.hour.asleep   0.7745     0.1161    6.673 0.00000293 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1628 on 18 degrees of freedom
Multiple R-squared:  0.7121,	Adjusted R-squared:  0.6961 
F-statistic: 44.53 on 1 and 18 DF,  p-value: 0.000002927
```

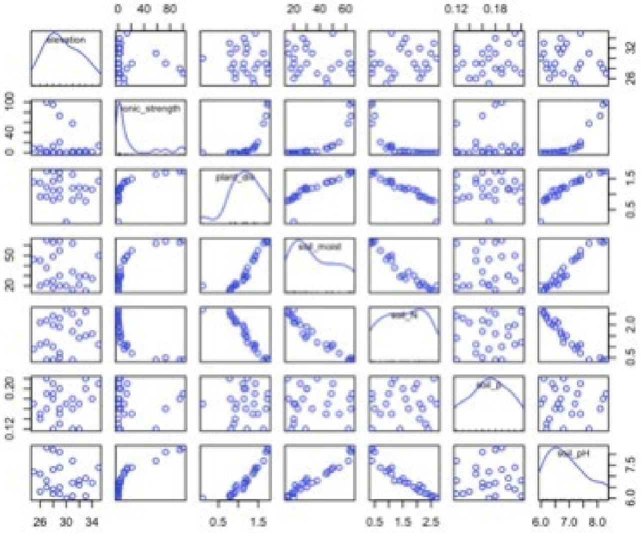
This is a case of spurious correlation (both variables were divided by hours asleep)

When # of candies and # of nightmares are directly compared, there is no correlation (test it in excel)



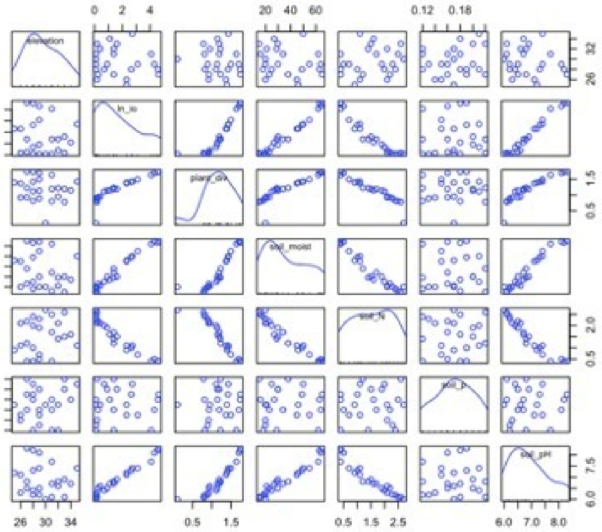
- Since the calculated p-value (0.000002927) is a lot less than the alpha value (alpha = 0.001), the number of candies is significant and we can conclude that there is a relationship between the number of candies consumed and the number of nightmares experienced by children. From this experiment, we can say that eating candies is correlated with the number of nightmares, but this does not necessarily mean that eating candies cause the nightmares because there might be other factors that can cause nightmares in suburb children, such as watching a scary film or family problems. One thing that we are sure of, however, is that 69.61% of the variation in the number of nightmares per hour of sleep is accounted for by the number of candies digested per hour of sleep. Also, for every 1 candy eaten, there is a 0.7745 increase in the number of nightmares per hour of sleep. Further experiment investigation is needed to prove that candies do cause nightmares in suburb children.

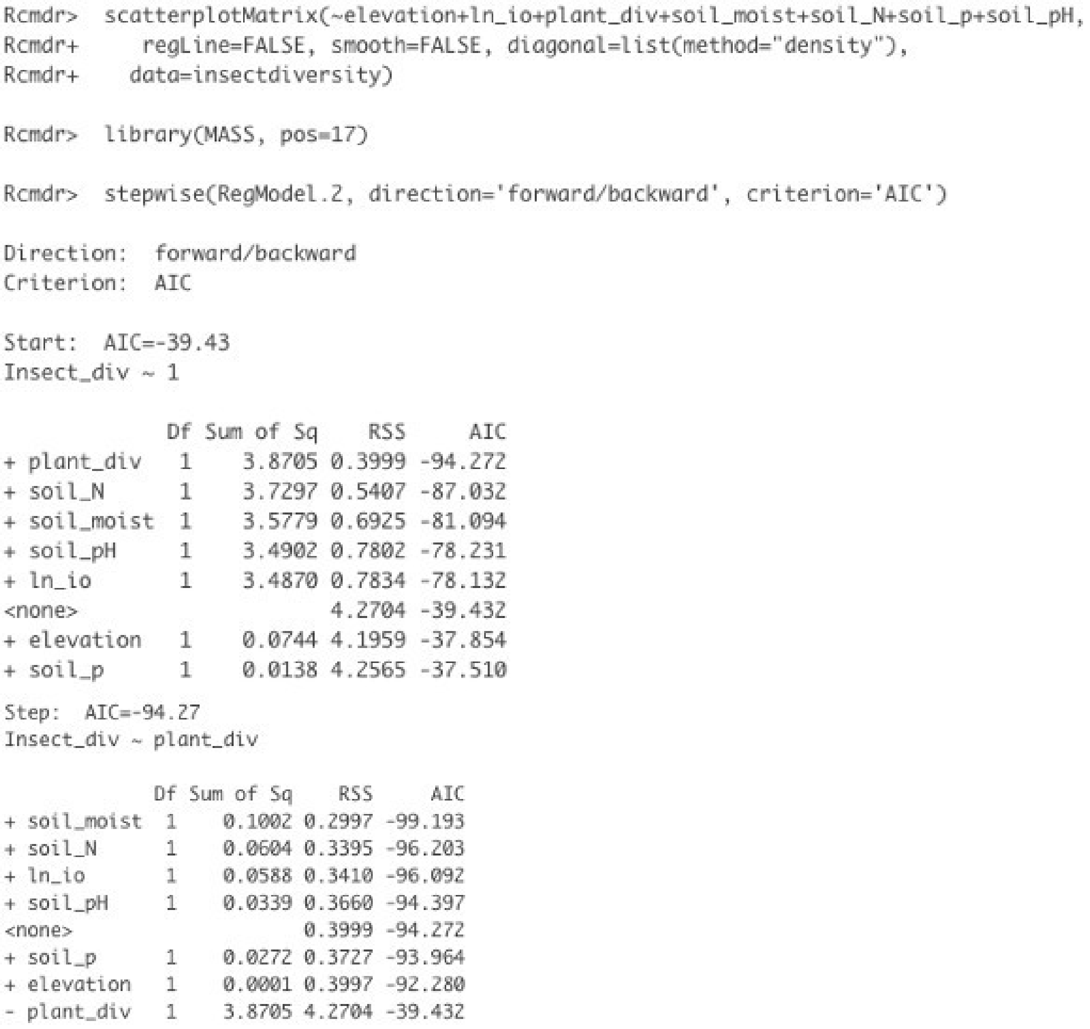
Q7 3



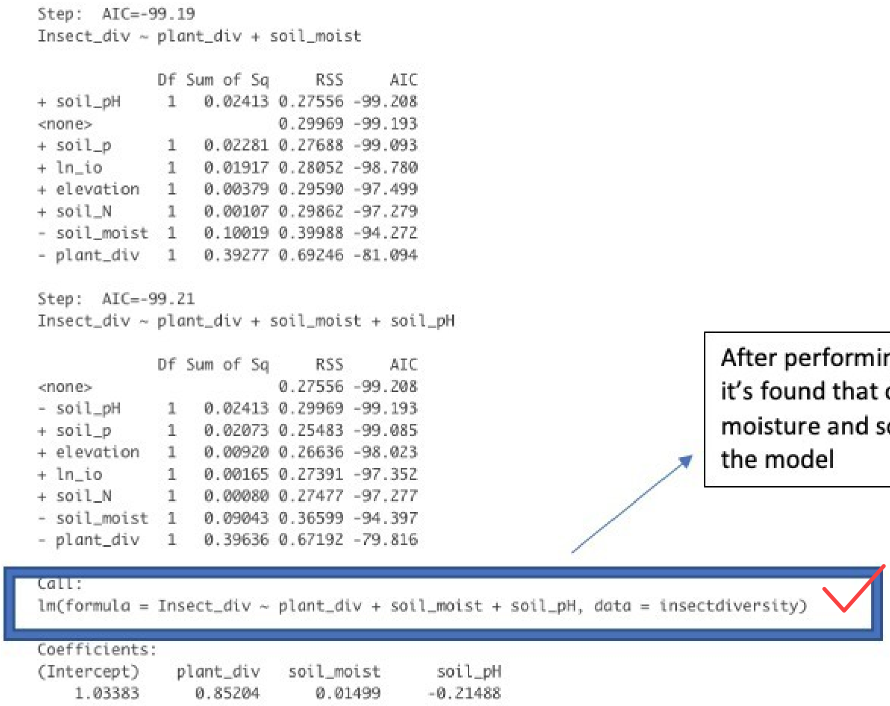
Picture above: Ionic strength distribution is the only one that's skewed (right-skewed), so it's the only variable that will need logarithmic transformation.

Picture below: The scatterplot matrix of the model after the logarithmic transformation of ionic strength. The ionic strength distribution looks more normal now.





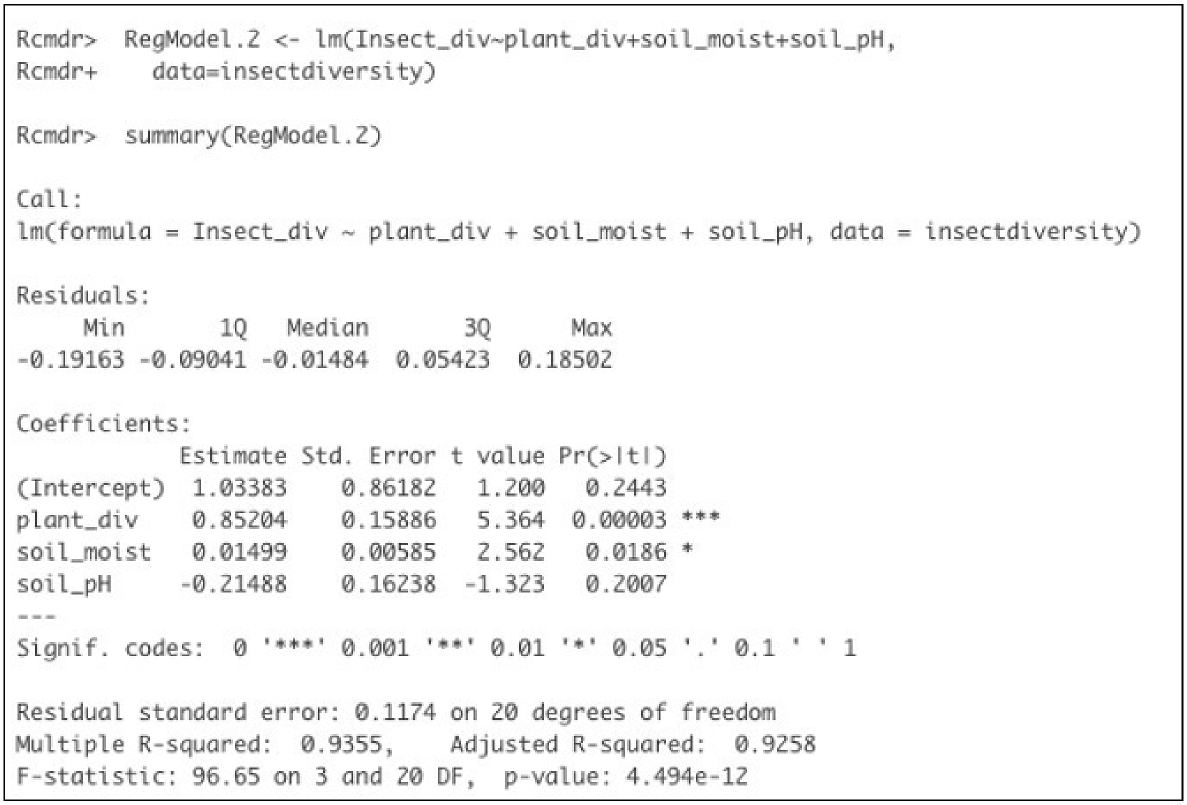
```
Rcmdr>  scatterplotMatrix(~elevation+ln_io+plant_div+soil_moist+soil_N+soil_p+soil_pH,
Rcmdr+      regLine=FALSE, smooth=FALSE, diagonal=list(method="density"),
Rcmdr+     data=insectdiversity)

Rcmdr>  library(MASS, pos=17)

Rcmdr>  stepwise(RegModel.2, direction='forward/backward', criterion='AIC')

Direction:  forward/backward
Criterion:  AIC

Start:  AIC=-39.43
Insect_div ~ 1

             Df Sum of Sq    RSS     AIC
+ plant_div   1    3.8705 0.3999 -94.272
+ soil_N      1    3.7297 0.5407 -87.032
+ soil_moist  1    3.5779 0.6925 -81.094
+ soil_pH     1    3.4902 0.7802 -78.231
+ ln_io       1    3.4870 0.7834 -78.132
<none>                    4.2704 -39.432
+ elevation   1    0.0744 4.1959 -37.854
+ soil_p      1    0.0138 4.2565 -37.510

Step:  AIC=-94.27
Insect_div ~ plant_div

             Df Sum of Sq    RSS     AIC
+ soil_moist  1    0.1002 0.2997 -99.193
+ soil_N      1    0.0604 0.3395 -96.203
+ ln_io       1    0.0588 0.3410 -96.092
+ soil_pH     1    0.0339 0.3660 -94.397
<none>                    0.3999 -94.272
+ soil_p      1    0.0272 0.3727 -93.964
+ elevation   1    0.0001 0.3997 -92.280
- plant_div   1    3.8705 4.2704 -39.432
```



```
Step:  AIC=-99.19
Insect_div ~ plant_div + soil_moist

             Df Sum of Sq     RSS     AIC
+ soil_pH     1   0.02413 0.27556 -99.208
<none>                    0.29969 -99.193
+ soil_p      1   0.02281 0.27688 -99.093
+ ln_io       1   0.01917 0.28052 -98.780
+ elevation   1   0.00379 0.29590 -97.499
+ soil_N      1   0.00107 0.29862 -97.279
- soil_moist  1   0.10019 0.39988 -94.272
- plant_div   1   0.39277 0.69246 -81.094

Step:  AIC=-99.21
Insect_div ~ plant_div + soil_moist + soil_pH

             Df Sum of Sq     RSS     AIC
<none>                    0.27556 -99.208
- soil_pH     1   0.02413 0.29969 -99.193
+ soil_p      1   0.02073 0.25483 -99.085
+ elevation   1   0.00920 0.26636 -98.023
+ ln_io       1   0.00165 0.27391 -97.352
+ soil_N      1   0.00080 0.27477 -97.277
- soil_moist  1   0.09043 0.36599 -94.397
- plant_div   1   0.39636 0.67192 -79.816

Call:
lm(formula = Insect_div ~ plant_div + soil_moist + soil_pH, data = insectdiversity)

Coefficients:
(Intercept)    plant_div   soil_moist      soil_pH
    1.03383      0.85204      0.01499     -0.21488
```

After performing AIC criterion selection in R, it's found that only plant diversity, soil moisture and soil pH have the most impact on the model

After identifying the variables that have the most impact and are therefore the most relevant for the model, another linear regression is ran to identify the intercepts and the slopes of the relevant variables: ✓



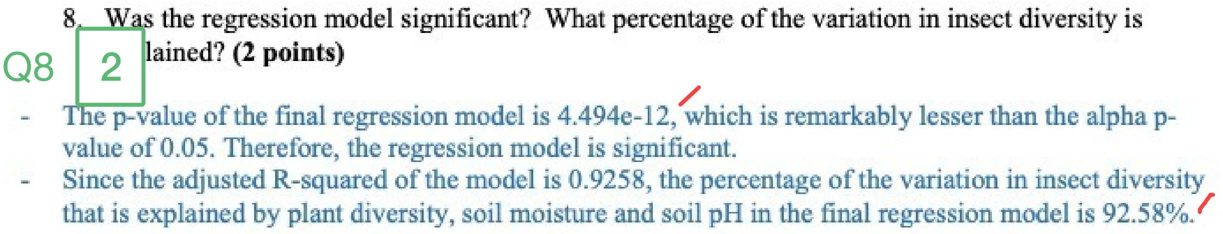
```
Rcmdr>  RegModel.2 <- lm(Insect_div~plant_div+soil_moist+soil_pH,
Rcmdr+    data=insectdiversity)

Rcmdr>  summary(RegModel.2)

Call:
lm(formula = Insect_div ~ plant_div + soil_moist + soil_pH, data = insectdiversity)

Residuals:
     Min       1Q   Median       3Q      Max
-0.19163 -0.09041 -0.01484  0.05423  0.18502

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.03383    0.86182   1.200   0.2443
plant_div    0.85204    0.15886   5.364  0.00003 ***
soil_moist   0.01499    0.00585   2.562   0.0186 *
soil_pH     -0.21488    0.16238  -1.323   0.2007
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1174 on 20 degrees of freedom
Multiple R-squared:  0.9355,    Adjusted R-squared:  0.9258
F-statistic: 96.65 on 3 and 20 DF,  p-value: 4.494e-12
```

$Y_{ij} = \alpha + B_1X_1 + B_2X_2 + B_nX_n + \varepsilon_{ij}$

Y = 1.03383 + 0.85204(plant_div) + 0.01499(soil_moist) – 0.21488(soil_pH) + 0.1174

Y = 1.15123 + 0.85204(plant_div) + 0.01499(soil_moist) – 0.21488(soil_pH) ✓

Q8 2



8. Was the regression model significant? What percentage of the variation in insect diversity is lained? **(2 points)**

- The p-value of the final regression model is 4.494e-12, which is remarkably lesser than the alpha p-value of 0.05. Therefore, the regression model is significant.
- Since the adjusted R-squared of the model is 0.9258, the percentage of the variation in insect diversity that is explained by plant diversity, soil moisture and soil pH in the final regression model is 92.58%.