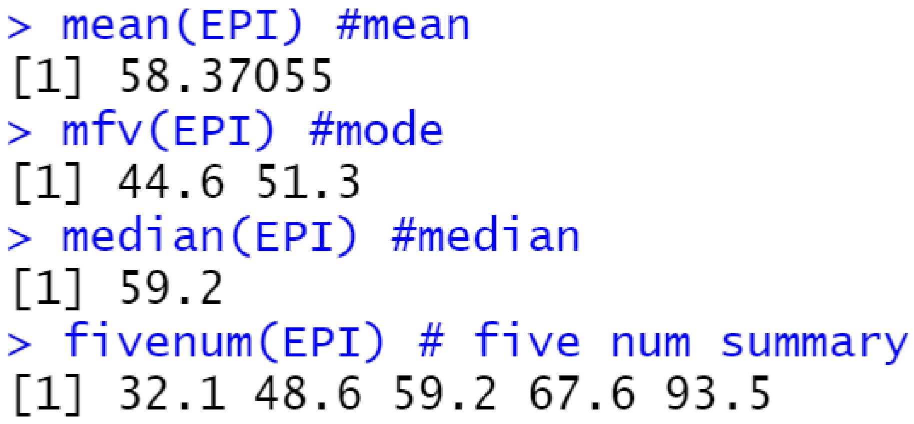
Data Analytics Assignment 2 Report

Part 1.

a). 1. Central Tendency of EPI:



```
> mean(EPI) #mean
[1] 58.37055
> mfv(EPI) #mode
[1] 44.6 51.3
> median(EPI) #median
[1] 59.2
> fivenum(EPI) # five num summary
[1] 32.1 48.6 59.2 67.6 93.5
```

Figure 1. Central Tendencies of EPI

2. Histogram for EPI variable shows the frequency distribution of Environmental Performance Index (EPI) value across all locations.

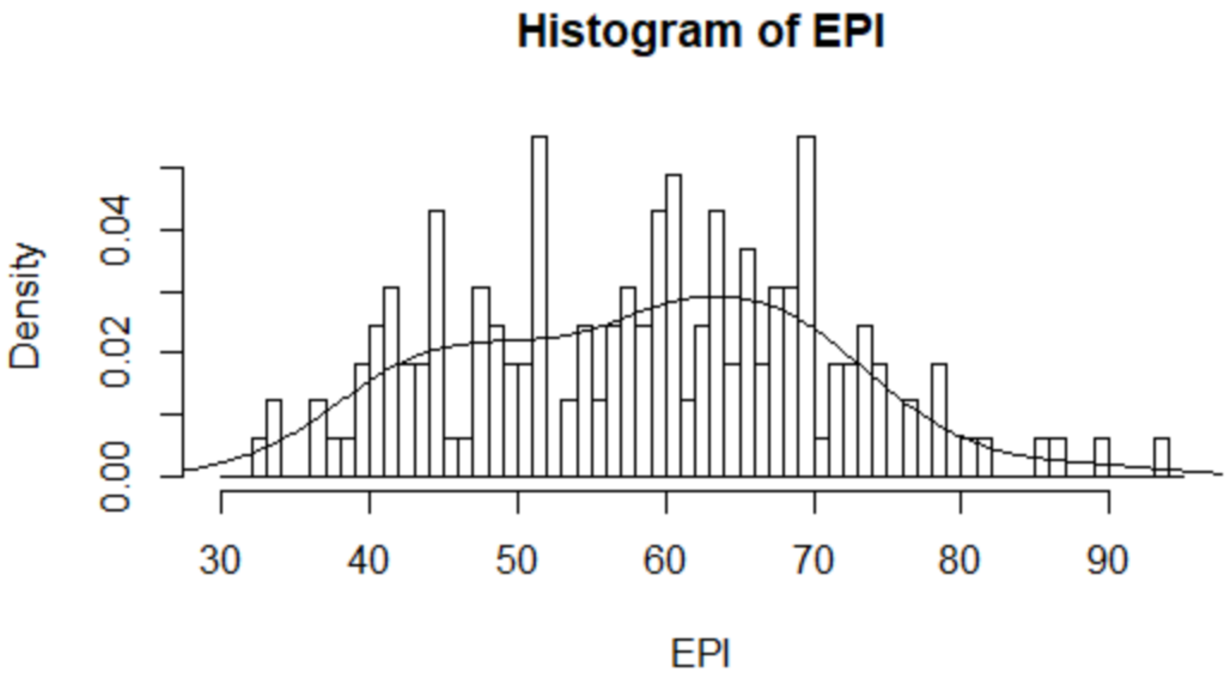


Figure 2. Distribution Histogram of EPI

3. Histogram for DALY variable shows the frequency distribution of disability-adjusted life year (DALY) across all locations.

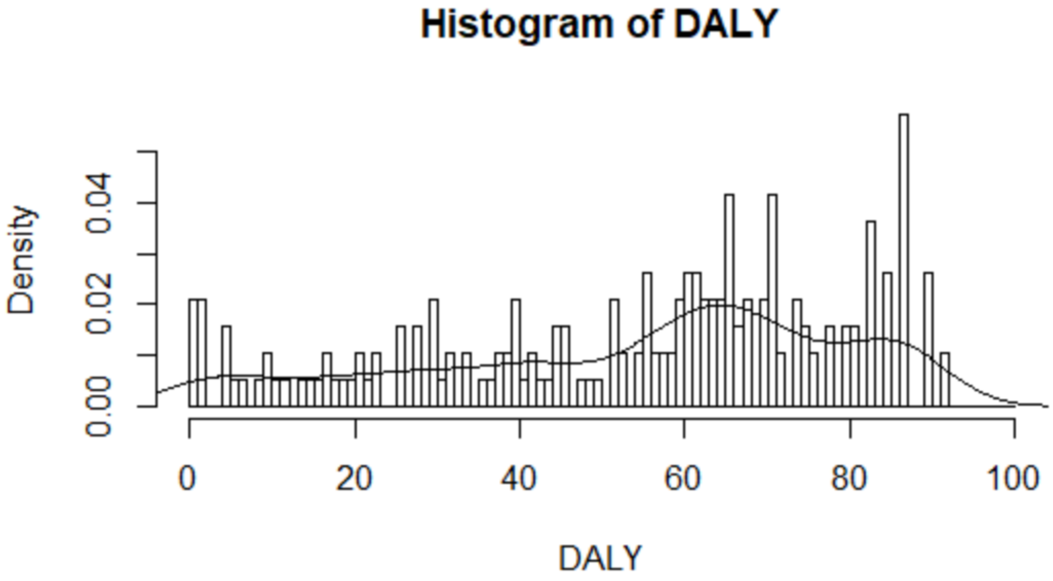


Figure 3. Distribution Histogram of DALY

4. The two box plots below provide visual summary of central tendencies and spreads of ENVHEALTH score, ECOSYSTEM score and compare them side by side, clearly showing the interquartile range and the outliers of each.

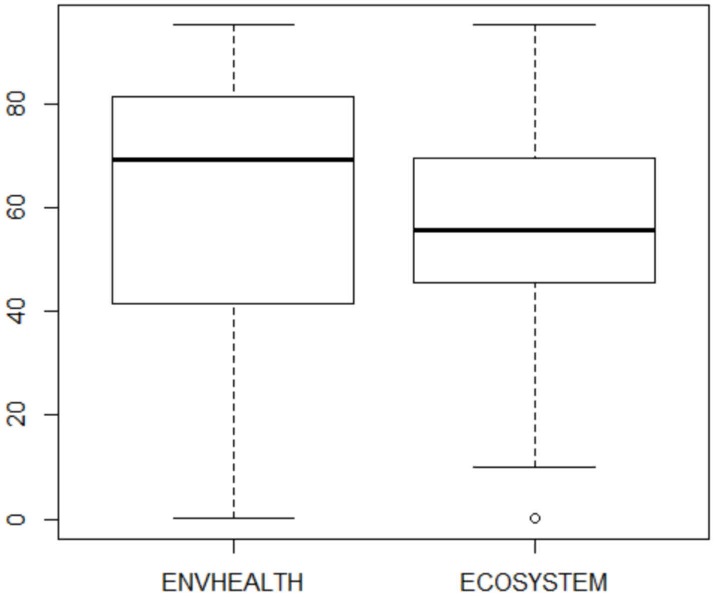


Figure 4. Boxplots of ENVHEALTH and ECOSYSTEM

5. A Q-Q plot basically compares two distributions. If two distributions are exactly the same, all the points should lie on the y = x line. However, the plot below indicates that the distributions of ENVHEALTH and ECOSYSTEM are not the same, meaning that the density. We can further verify this conclusion by plotting and comparing both distribution lines of ECOSYSTEM and ENVHEALTH on a single graph as shown in the Figure 6 below.

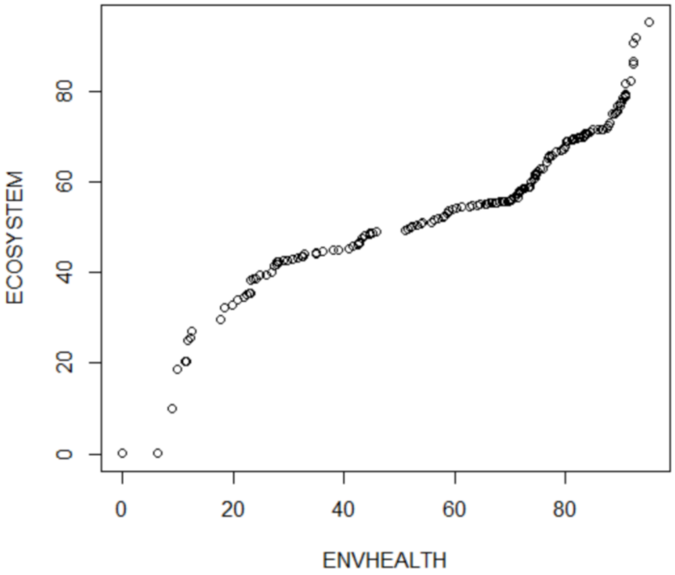


Figure 5. Q-Q Plot for ECOSYSTEM and ENVHEALTH

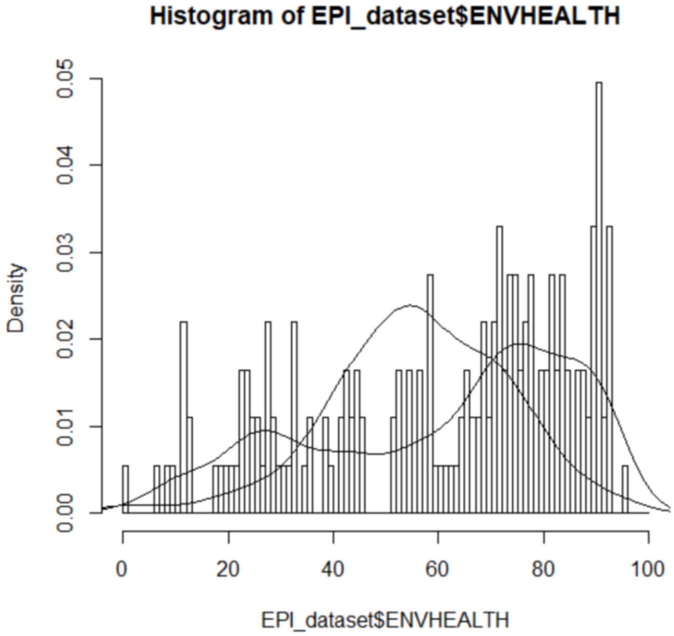
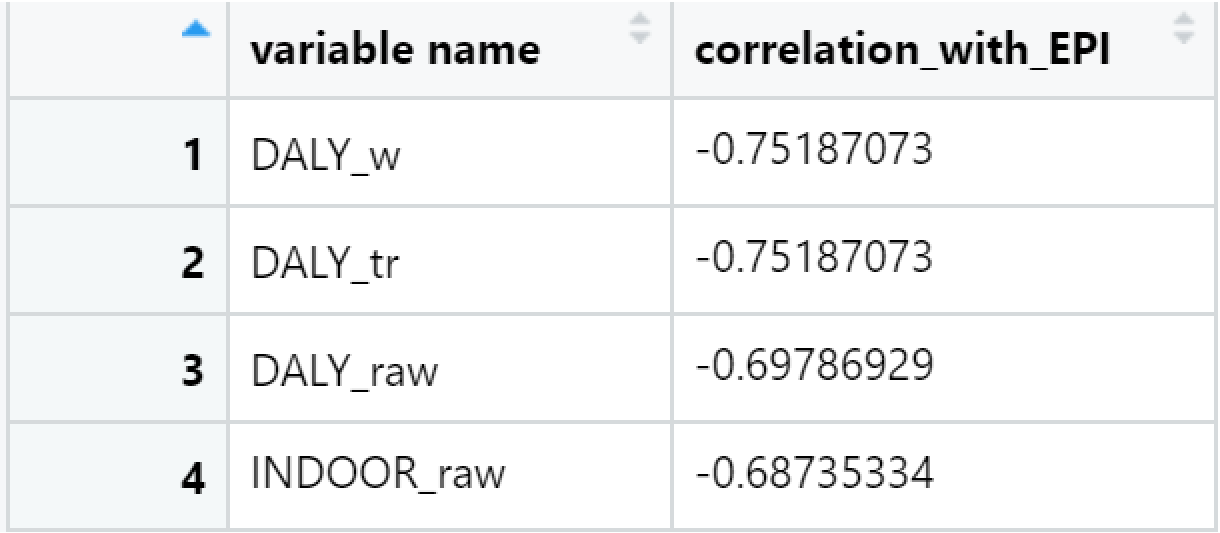


Figure 6. Distribution Lines of ECOSYSTEM and ENVHEALTH

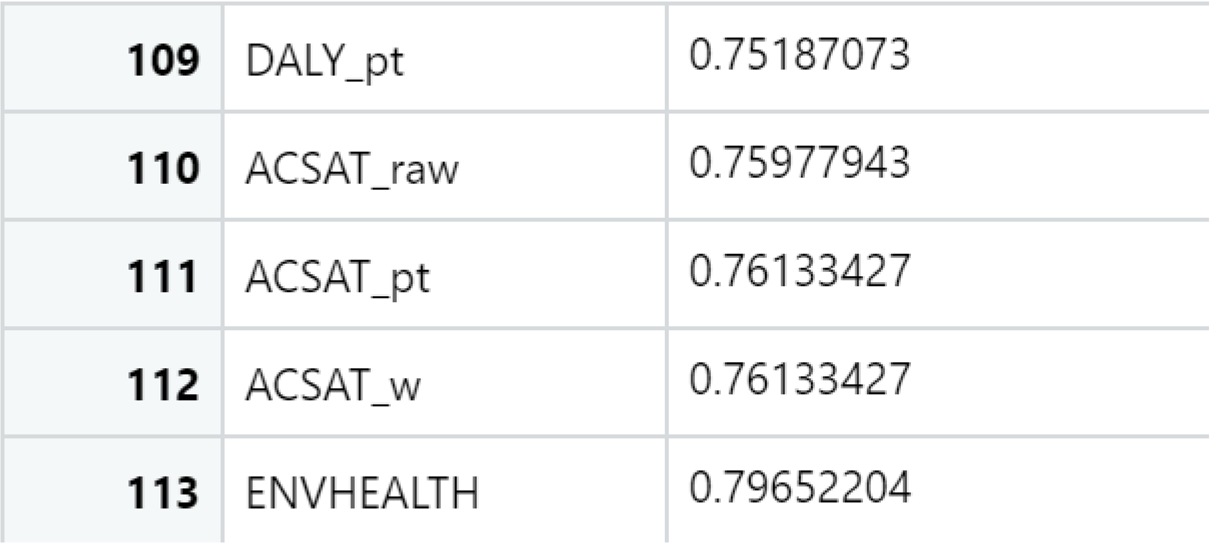
b).

1. The features that are mostly correlated (important) to EPI can be figured out by using the built in *cor()* function to generate a correlation list. Then, the features with high correlation coefficients can be easily identified by sorting the list, as shown in Figure 7 & 8 below.



	variable name	correlation_with_EPI
1	DALY_w	-0.75187073
2	DALY_tr	-0.75187073
3	DALY_raw	-0.69786929
4	INDOOR_raw	-0.68735334

Figure 7. The Most Positively Correlated Features



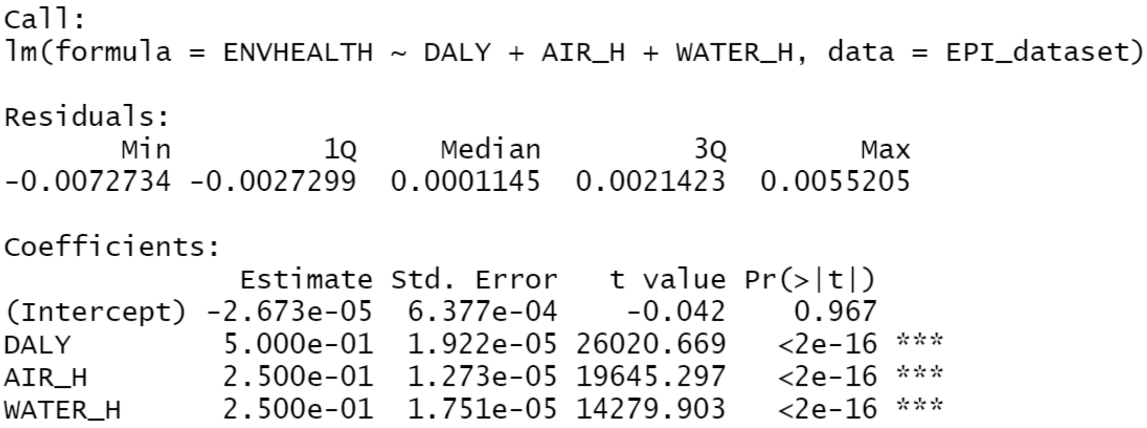
109	DALY_pt	0.75187073
110	ACSAT_raw	0.75977943
111	ACSAT_pt	0.76133427
112	ACSAT_w	0.76133427
113	ENVHEALTH	0.79652204

Figure 8. The Most Negatively Correlated Features

2. A linear regression model is built and used to generate predicted values to show linear relationship between ENVHEALTH and some selected features. According to the summary shown in Figure 9, the equation can be written as below. And a snippet of the predicted ENVHEALTH results for our artificial data is shown in Figure 10. Row one means that with all independent variables being set to 5, the predicted $\boldsymbol{ENVHEALTH}$ value will be 5.000048, with a 95% of confidence interval between 4.993824 and 5.006272.

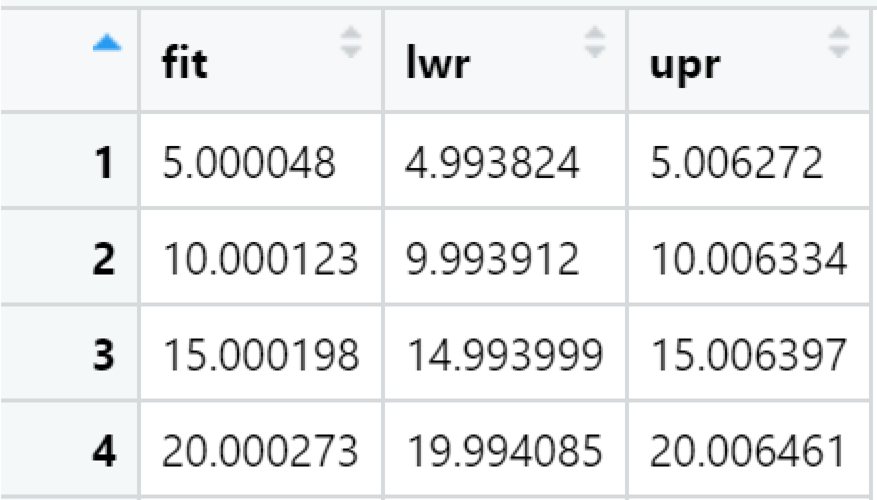
Equation 1.

$$\boldsymbol{ENVHEALTH} = 0.5 * \boldsymbol{DALY} + 0.25 * \boldsymbol{AIR\_H} + 0.25 * \boldsymbol{WATER\_H} - 2.67 * 10^{-5}$$



```
Call:
lm(formula = ENVHEALTH ~ DALY + AIR_H + WATER_H, data = EPI_dataset)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0072734 -0.0027299  0.0001145  0.0021423  0.0055205 

Coefficients:
              Estimate Std. Error   t value Pr(>|t|)    
(Intercept) -2.673e-05  6.377e-04    -0.042    0.967    
DALY         5.000e-01  1.922e-05 26020.669   <2e-16 ***
AIR_H        2.500e-01  1.273e-05 19645.297   <2e-16 ***
WATER_H      2.500e-01  1.751e-05 14279.903   <2e-16 ***
```

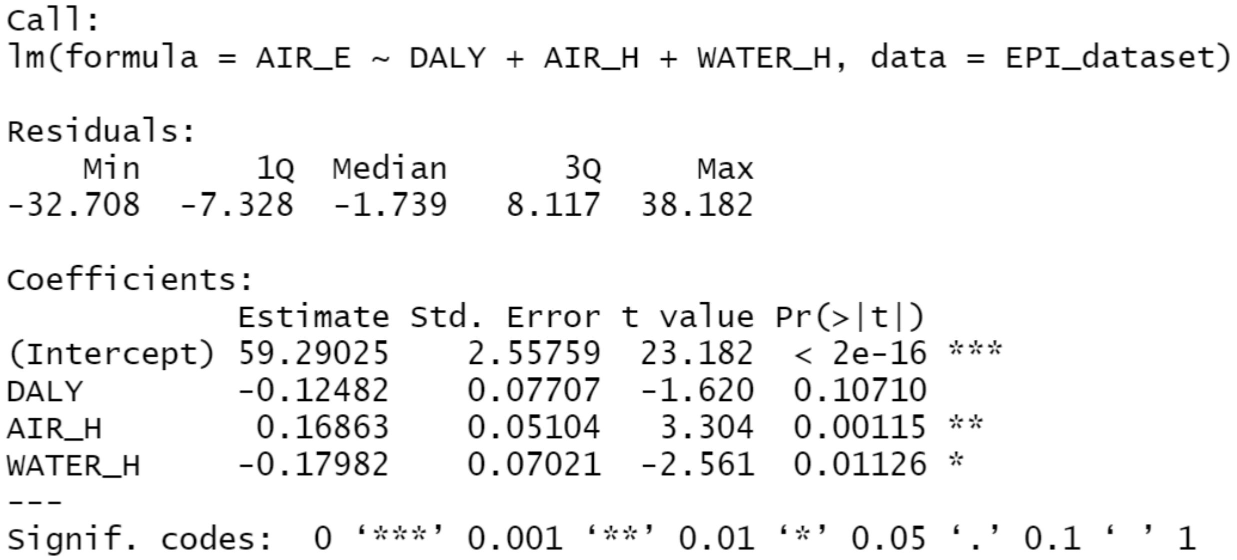
Figure 9. Linear Regression Summary



	fit	lwr	upr
1	5.000048	4.993824	5.006272
2	10.000123	9.993912	10.006334
3	15.000198	14.993999	15.006397
4	20.000273	19.994085	20.006461

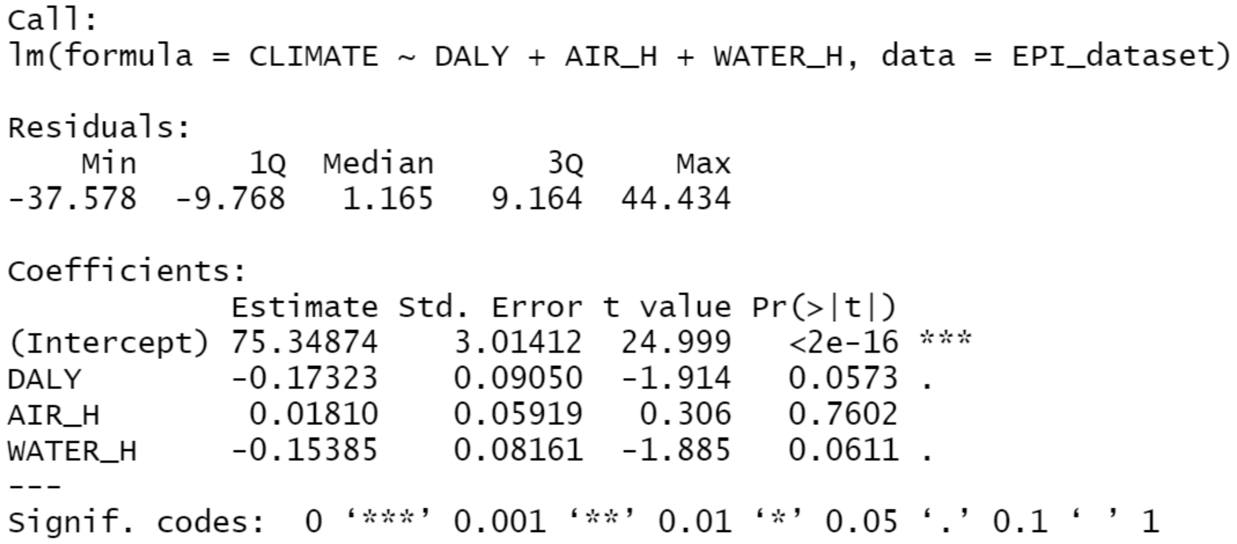
Figure 10. Snippet of the Predicted Results

3. Then, similar processes are applied to variable **AIR_E** and **CLIMATE**. Summaries are shown in Figure 11 & 12 below. Noted, variable **DALY** is not significant enough for explaining **AIR_E**, and **AIR_H** is not significant enough for explaining **CLIMATE** at 95% confidence level.



```
Call:
lm(formula = AIR_E ~ DALY + AIR_H + WATER_H, data = EPI_dataset)

Residuals:
    Min      1Q  Median      3Q     Max 
-32.708  -7.328  -1.739   8.117  38.182 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 59.29025    2.55759  23.182  < 2e-16 ***
DALY        -0.12482    0.07707  -1.620  0.10710    
AIR_H        0.16863    0.05104   3.304  0.00115 ** 
WATER_H     -0.17982    0.07021  -2.561  0.01126 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 11. Summary for AIR_E Regression



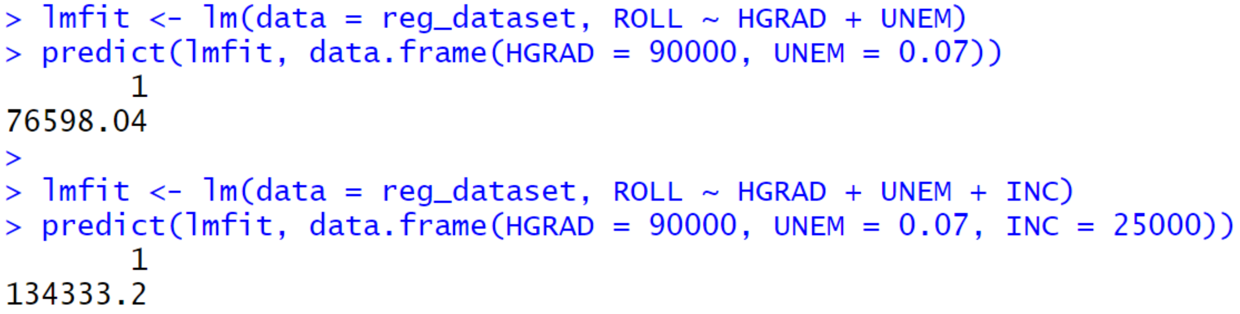
```
Call:
lm(formula = CLIMATE ~ DALY + AIR_H + WATER_H, data = EPI_dataset)

Residuals:
    Min      1Q  Median      3Q     Max 
-37.578  -9.768   1.165   9.164  44.434 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 75.34874    3.01412  24.999   <2e-16 ***
DALY        -0.17323    0.09050  -1.914   0.0573 .  
AIR_H        0.01810    0.05919   0.306   0.7602    
WATER_H     -0.15385    0.08161  -1.885   0.0611 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 12.Summary for CLIMATE Regression

Part 2.

Exercise 1).

In this exercise, we build two linear regressions to find the linear relationship between enrollment (ROLL) and some selected features. Equations and predicted results are shown as below.

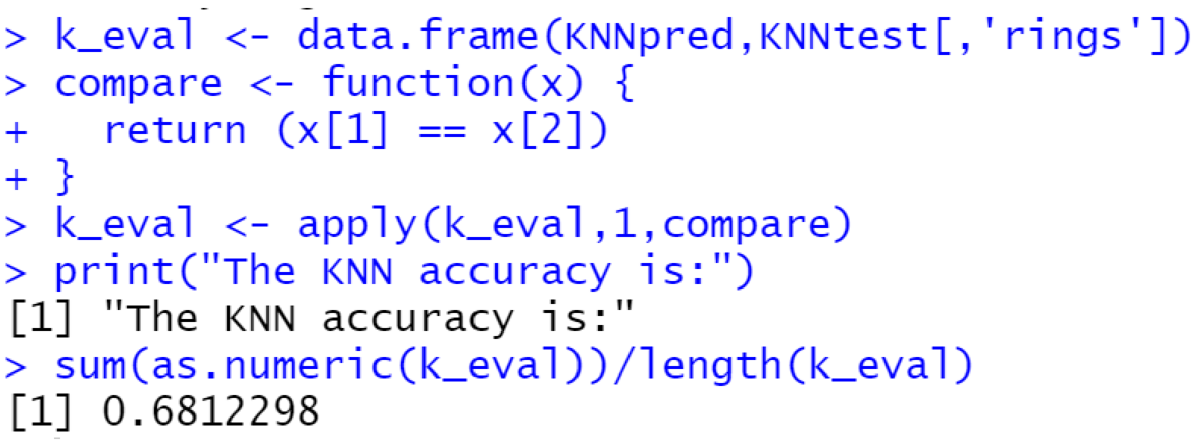


```
> lmfit <- lm(data = reg_dataset, ROLL ~ HGRAD + UNEM)
> predict(lmfit, data.frame(HGRAD = 90000, UNEM = 0.07))
       1 
76598.04 
> 
> lmfit <- lm(data = reg_dataset, ROLL ~ HGRAD + UNEM + INC)
> predict(lmfit, data.frame(HGRAD = 90000, UNEM = 0.07, INC = 25000))
       1 
134333.2 
```

Figure 13. Multiple Regression Exercise

Exercise 2).

In this exercise, we build a KNN model to predict abalones' age label (i.e. young, adult, old) using several selected features. Since the random seed is not fixed, the predicted result will have some variation but the testing (out of sample) accuracy is always similar at around 66% - 68% with a train test split of 7:3 and k = 55.



```
> k_eval <- data.frame(KNNpred,KNNtest[,'rings'])
> compare <- function(x) {
+   return (x[1] == x[2])
+ }
> k_eval <- apply(k_eval,1,compare)
> print("The KNN accuracy is:")
[1] "The KNN accuracy is:"
> sum(as.numeric(k_eval))/length(k_eval)
[1] 0.6812298
```

Figure 14. KNN Accuracy

Exercise 3).

In this exercise, we are using K-means to cluster iris type based on their sepal length, sepal width, petal length, and petal width. If we were to use only two of the features, visualizations of the attribute will be easy to comprehend as shown as Figure 15. But, in this problem, we are using all four of the features, so that actual algorithm will run on a 4-dimensional graph which cannot be graphed with any existing tool. However, we can still see the clustering result in a table format. Configurations and results are shown in Figure 16.

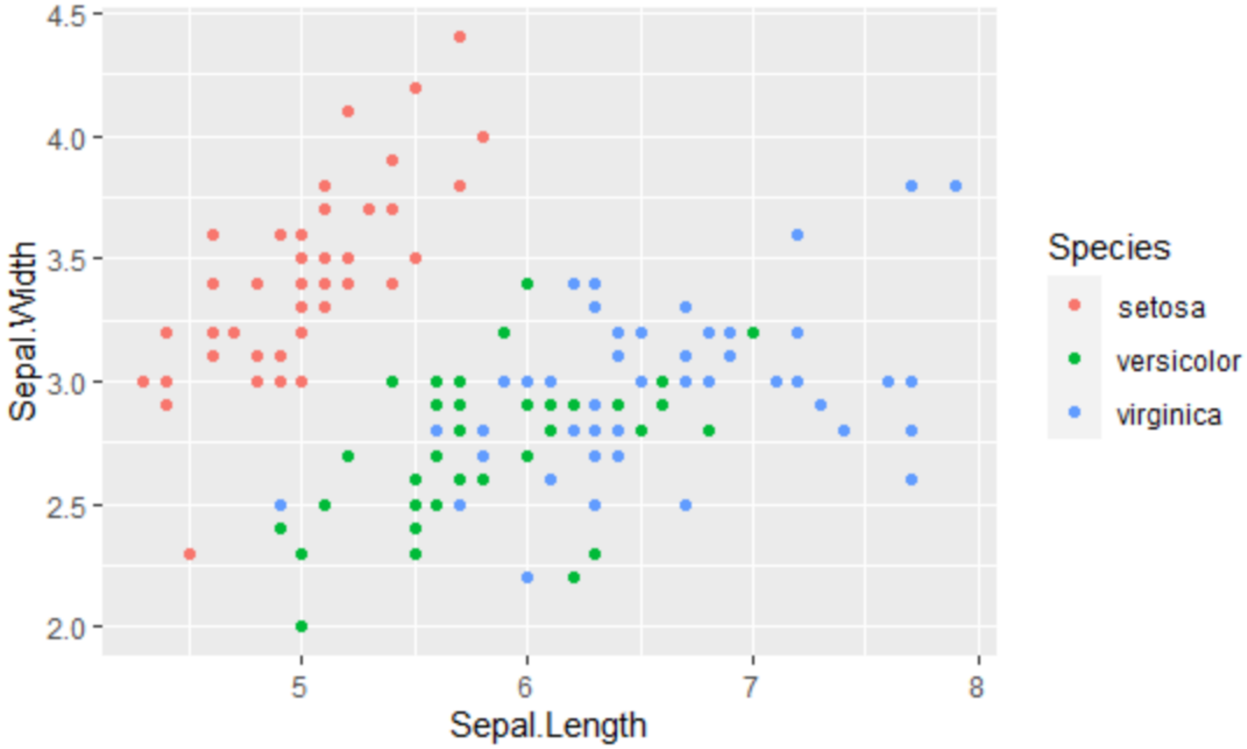
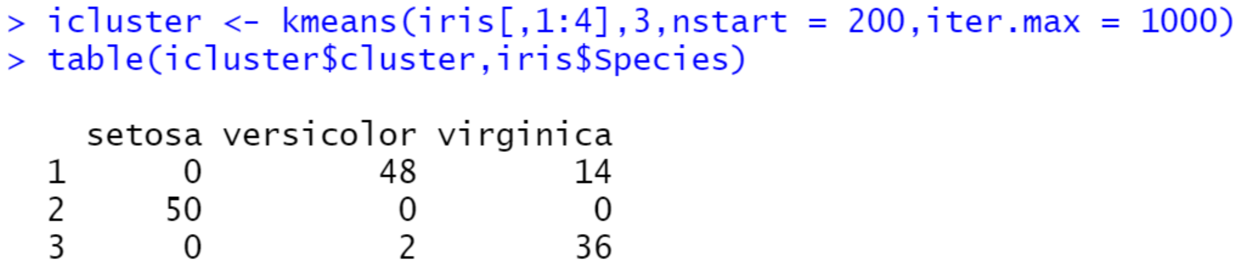


Figure 15. 2-D Visualization



```
> icluster <- kmeans(iris[,1:4],3,nstart = 200,iter.max = 1000)
> table(icluster$cluster,iris$Species)

   setosa versicolor virginica
  1      0         48        14
  2     50          0         0
  3      0          2        36
```

Figure 16. Clustering Result