Tony Min
Data Analytics
October 13, 2023

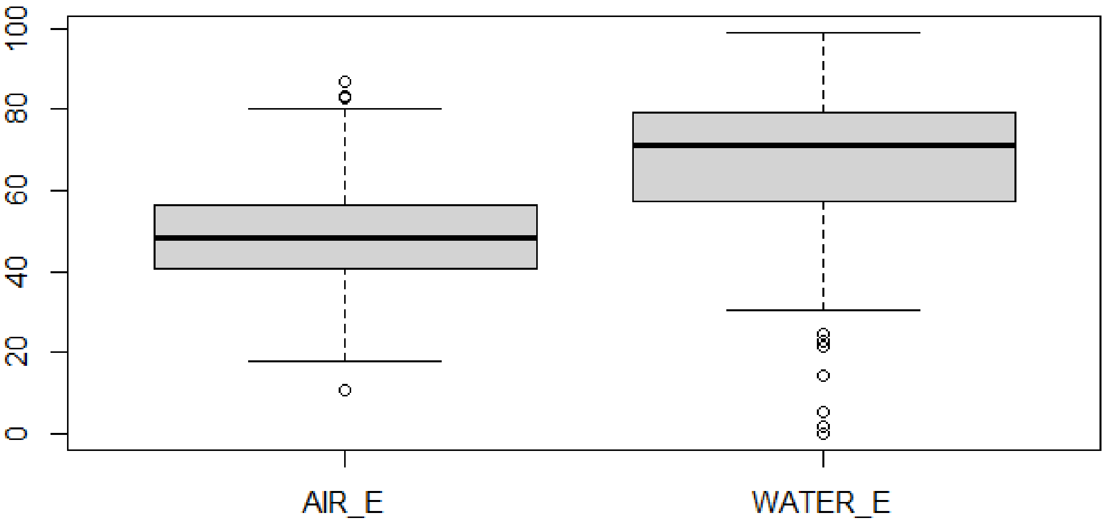
Assignment 2

Part 1
Section (a)

Central Tendency Values for AIR_E and WATER_E

Column	Mean	Median	Mode
AIR_E	49.46	48.24	44.69
WATER_E	67.48	71.17	71.4

Boxplot of AIR_E and WATER_E

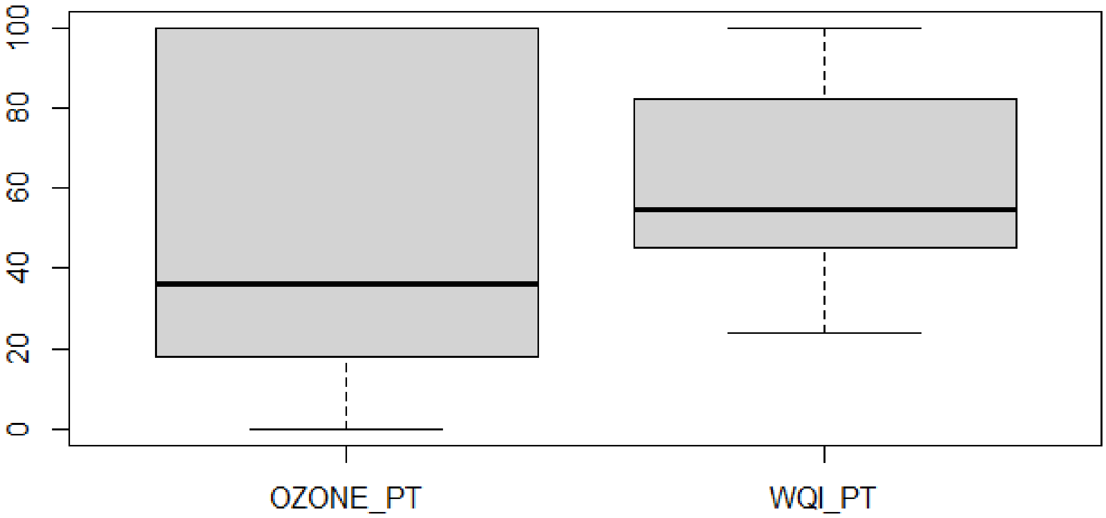


Based on the Boxplot and the central tendency values we found from these data sets we see that their are some outliers in both sets, both in AIR_E and WATER_E but we see that WATER_E has more outliers that are below the the minimum but with AIR_E there are is one outlier below the minimum and 2 outliers above the maximum. We also see that the AIR_E median value is slower than WATER_E which we are able to visualize in the box plot.

Rest of Central Tendency Values

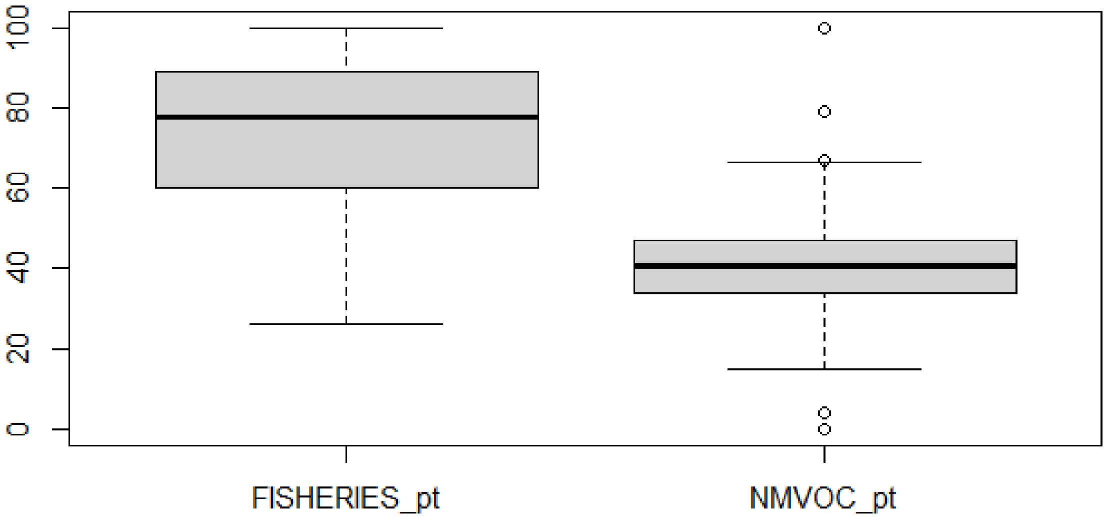
Column	Mean	Median	Mode
NOX_PT	47.51	48.04	28.36
SO2_PT	53.05	53.73	20.63
CLIMATE	55.33	55.43	60.74
AGRICULTURE	70.86	75.29	54.55

Boxplot OZONE_PT and WQI_PT



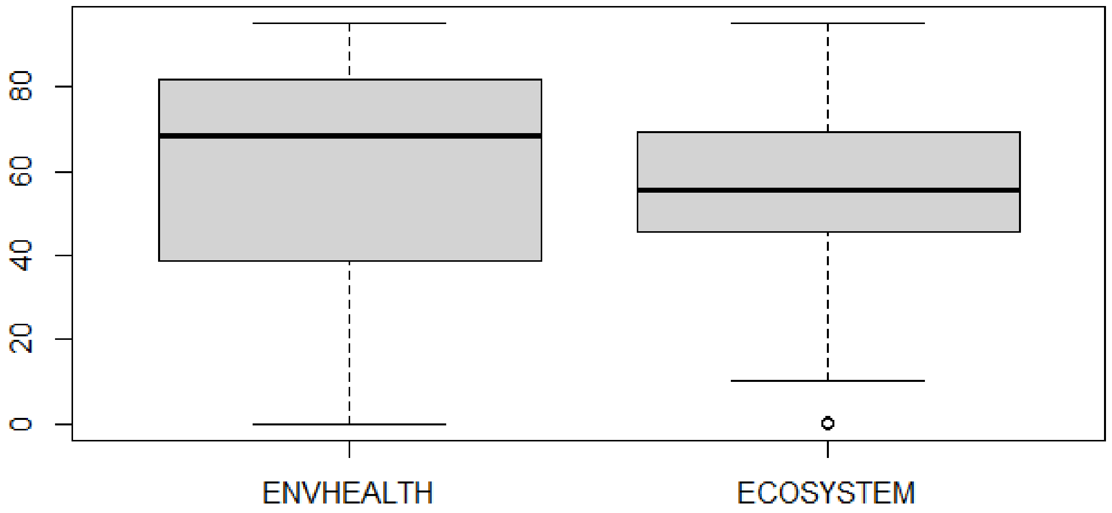
For OZONE and WQI we see that OZONE has a much larger range compared to WQI. From the box plots we see that the median for the OZONE is lower than WQI but Quarter 1 for WQI is actually higher than the median of OZONE as well. From the OZONE boxplot we see that there is also no maximum.

Boxplot of FISHERIES and NMVOC



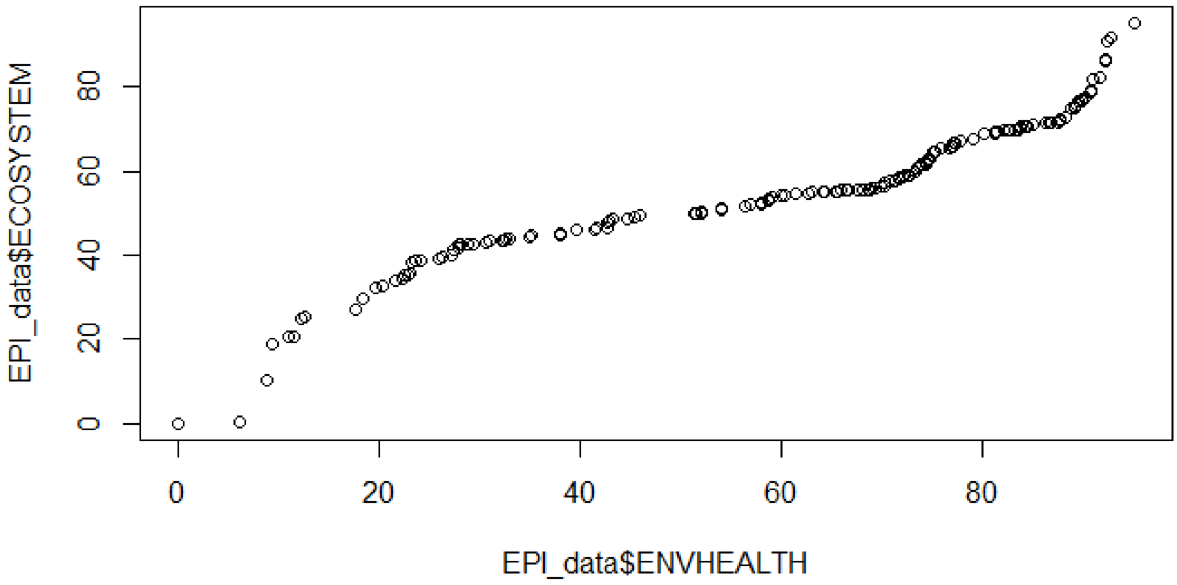
These boxplots show that there is a smaller range between quarters for NMVOC compared to the FISHERIES. There are a couple outliers for NMVOC as well. Also Quarters 2 and 3 for NMVOC are both lower than FISHERIES quarter 2 value.

Boxplot of ENVHEALTH and ECOSYSTEM



Based on the boxplot we see that ECOSYSTEM has one outlier at 0 while ENVHEALTH has a minimum at 0 and maximum of 100. The range between quarter 1 and 3 for ENVHEALTH is larger than the range that of ECOSYSTEM.

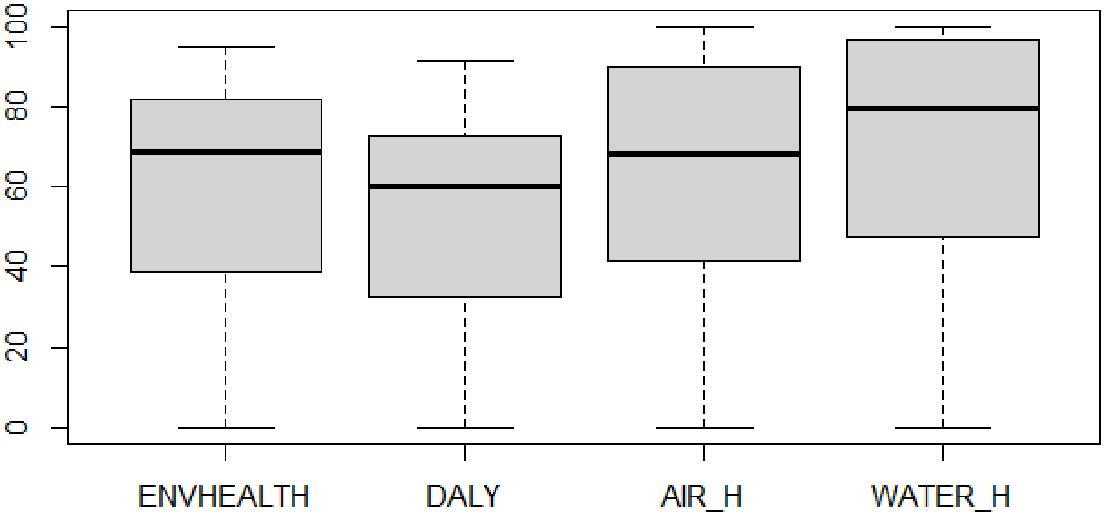
QQ-Plot for ENVHEALTH and ECOSYSTEM



In this QQ-Plot we see how it does not follow the line of equality therefore this data does not follow a normal distribution. It seems like the data has a slight curve up towards the upper middle half, 60-100.

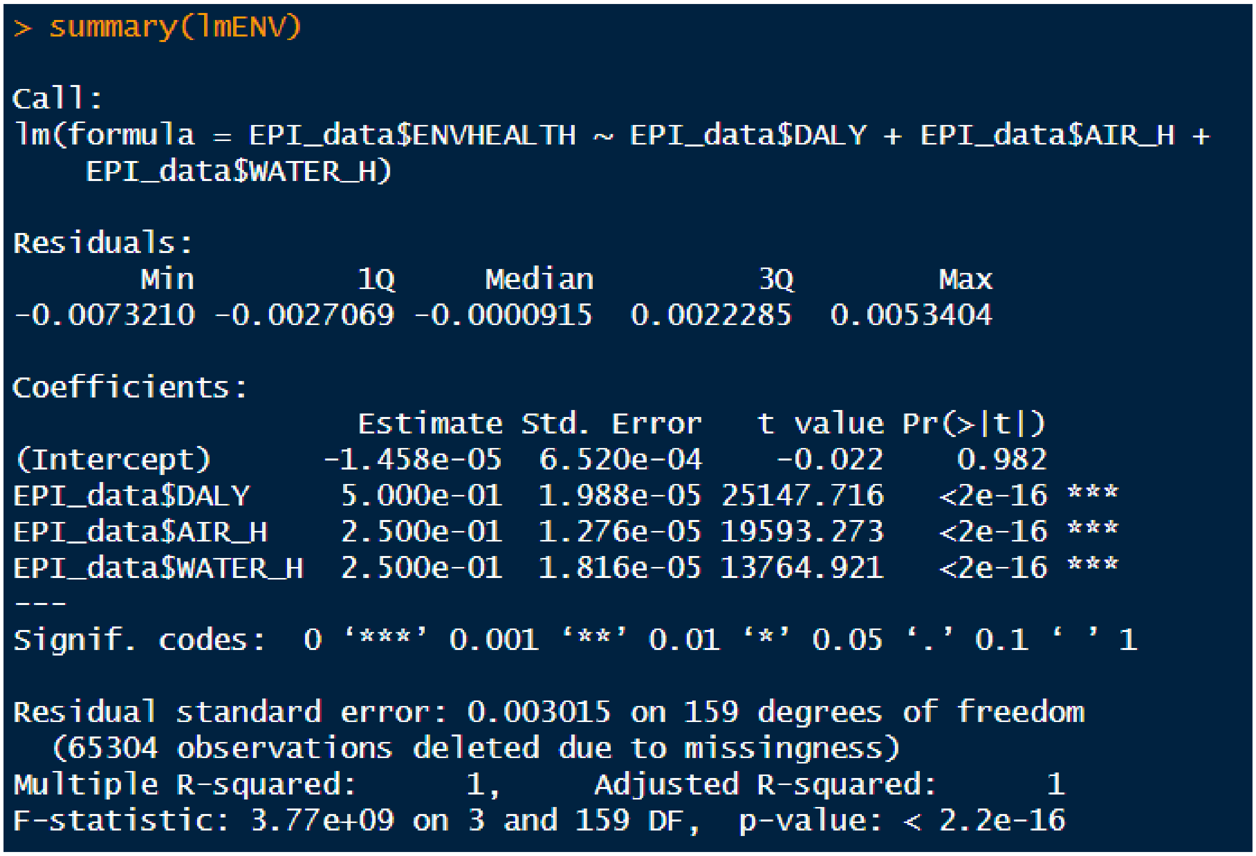
Section (b)

Boxplots for ENVHEALTH, DALY, AIR_H and WATER_H



The box plot shows that AIR_H and WATER_H have the largest range of distributions while the smallest one seems to be DALY. Ordering the medians from smallest to largest values we could say the DALY is the smallest then AIR_H, ENVHEALTH and WATER_H as the largest.

Next we created a linear regression model to predict the dependent variable, ENVHEALTH based on the three independent variables DALY, AIR_H and WATER_H. Using the summary() function on our linear regression model we find out that we get coefficients of .5 for DALY, .25 for AIR_H, and .25 for WATER_H with an intercept of -1.458e-05. Also the summary function returns a multiple R-squared value which is 1 in this case meaning that our model explains all the variance in ENVHEALTH, this may seem good but it could also indicate overfitting. Also with the significant codes we see that all the independent variables are highly significant based on the three stars (***). This result shows that the model is significant but also could mean that there is some overfitting happening.



```
> summary(lmENV)

Call:
lm(formula = EPI_data$ENVHEALTH ~ EPI_data$DALY + EPI_data$AIR_H +
    EPI_data$WATER_H)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0073210 -0.0027069 -0.0000915  0.0022285  0.0053404

Coefficients:
                   Estimate Std. Error   t value Pr(>|t|)
(Intercept)      -1.458e-05  6.520e-04    -0.022    0.982
EPI_data$DALY     5.000e-01  1.988e-05 25147.716   <2e-16 ***
EPI_data$AIR_H    2.500e-01  1.276e-05 19593.273   <2e-16 ***
EPI_data$WATER_H  2.500e-01  1.816e-05 13764.921   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.003015 on 159 degrees of freedom
  (65304 observations deleted due to missingness)
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 3.77e+09 on 3 and 159 DF,  p-value: < 2.2e-16
```

Next we created new data points to represent new values we will generate for the independent variables. We then created prediction and confidence intervals for the first linear regression model we created.

Next we created a linear regression model to predict AIR_E using the variables DALY, AIR_H and WATER_H. With this model we see that the multiple R-squared value of .1834 which indicates that the model explains 18.34% of the variance in AIR_E. The F-statistic is 11.9 with a p-value of 4.515e-07 indicates that the model is statistically significant. We see that Model 1 is statistically significant and it does attempt to predict AIR_E with DALY, AIR_H and WATER_H. We did create prediction and confidence intervals for this model as well.

Our final linear regression model was to predict CLIMATE with DALY, AIR_H and WATER_H. With this model we see that the multiple R-squared value is .2919 which means that our model explains 29.19% of the variance in CLIMATE. From the F-statistic 21.85 and p-value of 2.709e-12 we see the model as a whole is also statistically significant. We know that Model 2 is also statistically significant and does attempt to predict CLIMATE with DALY, AIR_H and WATER_H as the independent variables. Afterwards we also created prediction and confidence intervals for the model.

Section (c)
We used the Shapiro-Wilk test to assess is a sample of data comes from a normal distribution, if the p-value is below 0.05 we say that the null hypothesis is rejected therefore the data is not normally distributed.

2010 EPI Data Results

Variable	Sample Size	P-value	Reject/Accept?
ENVHEALTH	163	8.178e-10	Reject
DALY	163	1.523e-06	Reject
AIR_H	163	3.206e-07	Reject
WATER_H	163	1.348e-10	Reject

EPI Data Results

Variable	Sample Size	P-value	Reject/Accept
ENVHEALTH	182	1.083e-08	Reject
DALY	192	1.891e-07	Reject
AIR_H	197	8.994e-09	Reject
WATER_H	197	1.679e-12	Reject

From these two tables we see that all from both EPI and 2010 EPI data sets the selected variables all had between 3 and 5,000 in order to use the shapiro-wilk test. The results from both data sets also made sense if the 2010 dataset rejected the null hypothesis meaning that they were not normally distributed then the whole EPI dataset should also show that these variables are not normally distributed, which it did.

Part 2
Section (a)

For this part we are creating a linear regression model to predict ROLL using UNEM and HGRAD. We then predicted fall enrollment (ROLL) given that the unemployment rate (UNEM) is 7% and 90,000 spring high school graduates (HGRAD), resulting in 81,437 students.

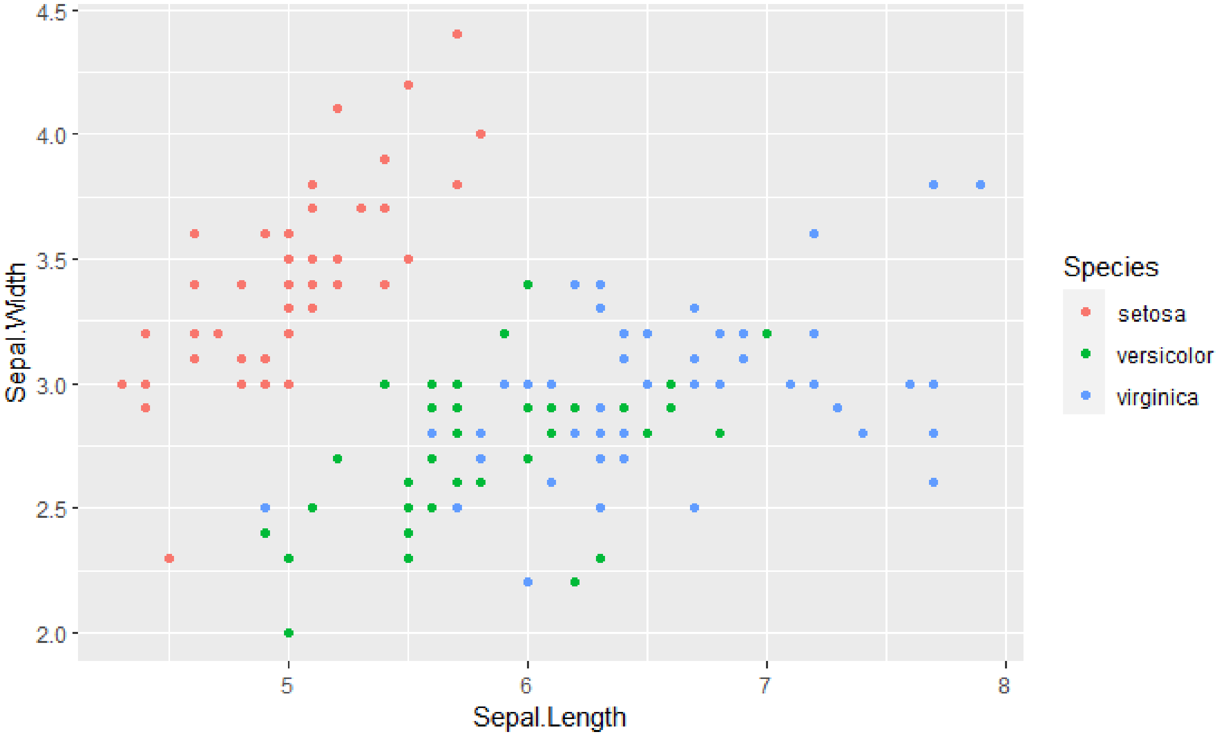
Next adding per capita income to the model (INC) as an independent, we set INC = 25000 with the same unemployment rate and spring high school graduates we predict about 137,453 students enrolled in the fall.

Section (b)

We first normalized the data first so each value is between 0 and 1. We then split the whole data set into training and testing sets. We then use the knn function to predict the number of rings (age) of abalone.

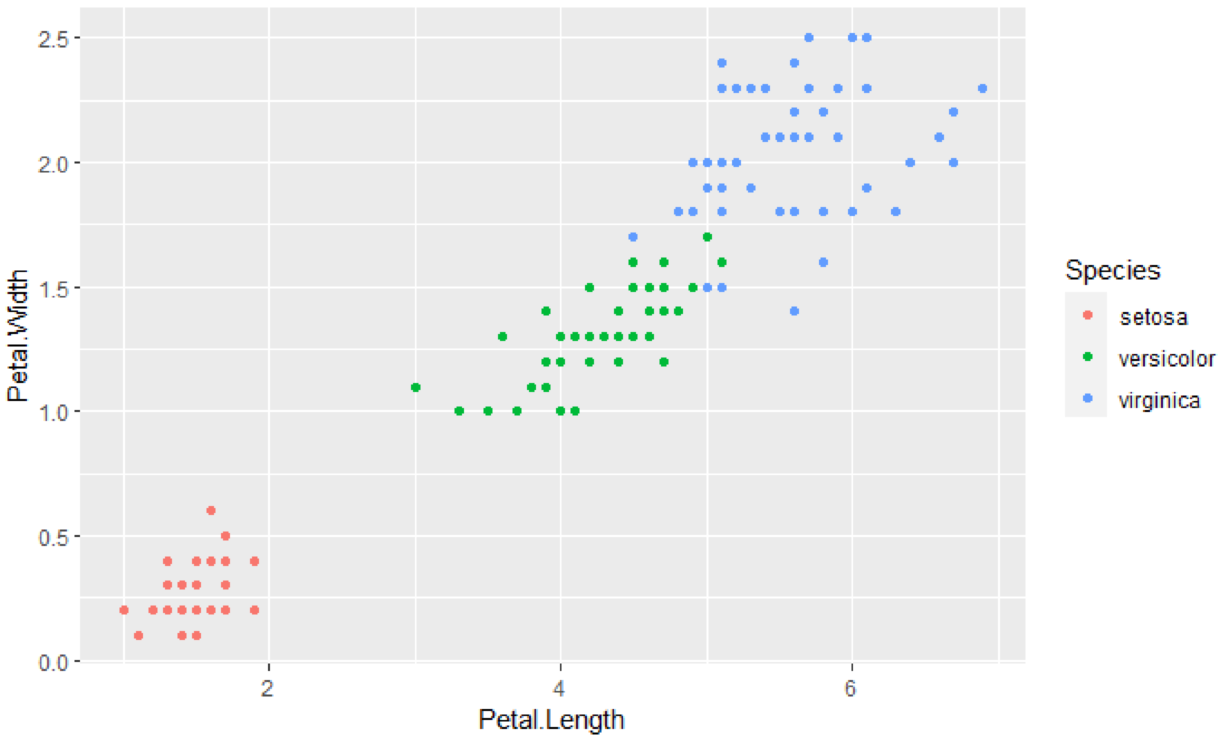
Section (c)

Using the iris data set we create a plot of sepal length vs sepal width.



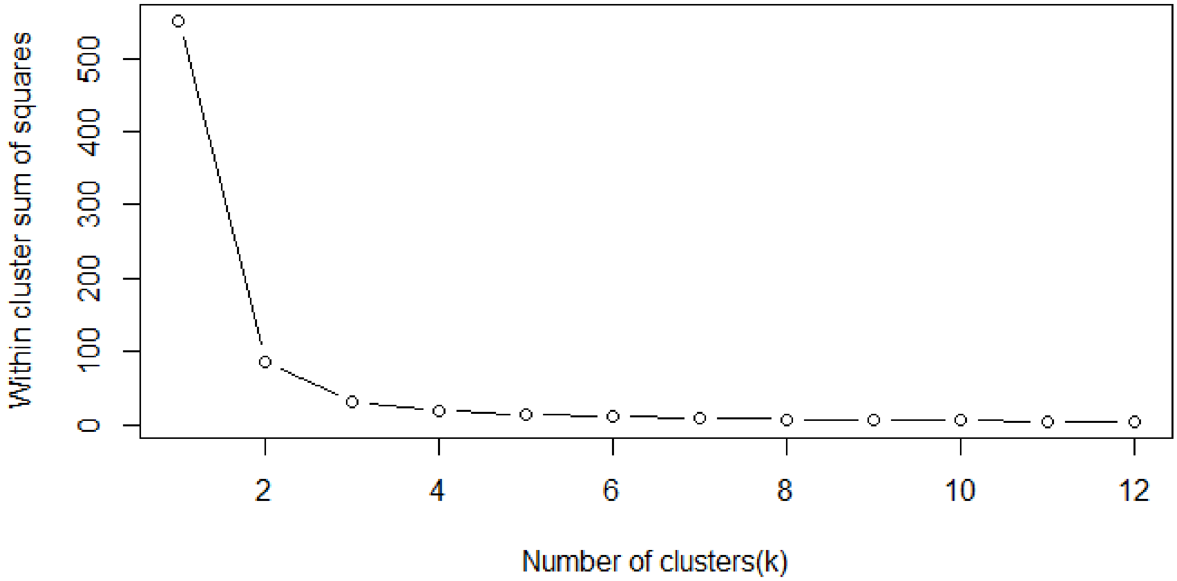
With this plot we see the setosa species are separated from the other two species. While versicolor and virginica are a little overlapped in some areas but we can see a shorted sepal length and width is classified as versicolor and larger lengths and widths are classified to virginica.

Next we created a plot for petal length vs petal width.



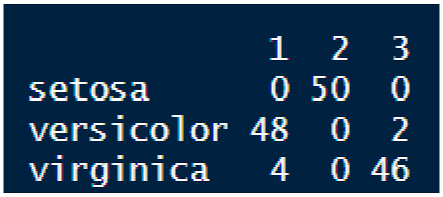
Now in this plot we see quite the separation between species. First we have setosa who have small petal widths and lengths while virginica have larger widths and lengths. Finally, versicolor is in between the lengths and widths of setosa and virginica.

Next we have a plot of squared based on the number of clusters.



In this graph we see that in the ‘elbow’ it sits at around 3 and is the optimal number of clusters.

Finally we create a table (shown below)



```
             1  2  3
setosa       0 50  0
versicolor  48  0  2
virginica    4  0 46
```

It seems like most of the data was classified correctly but we do see that 2 versicolor species are classified into cluster 3 instead of 1. Then 4 virginica have been put in cluster 1 instead of 3.