Izaak King

Professor Eleish

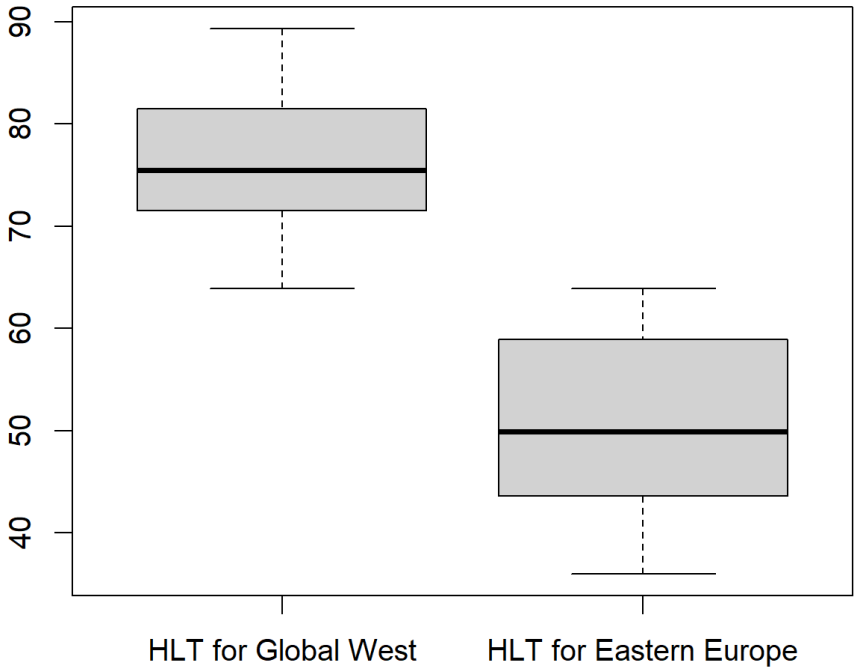
Data Science

October 14, 2025

"Assignment 2"

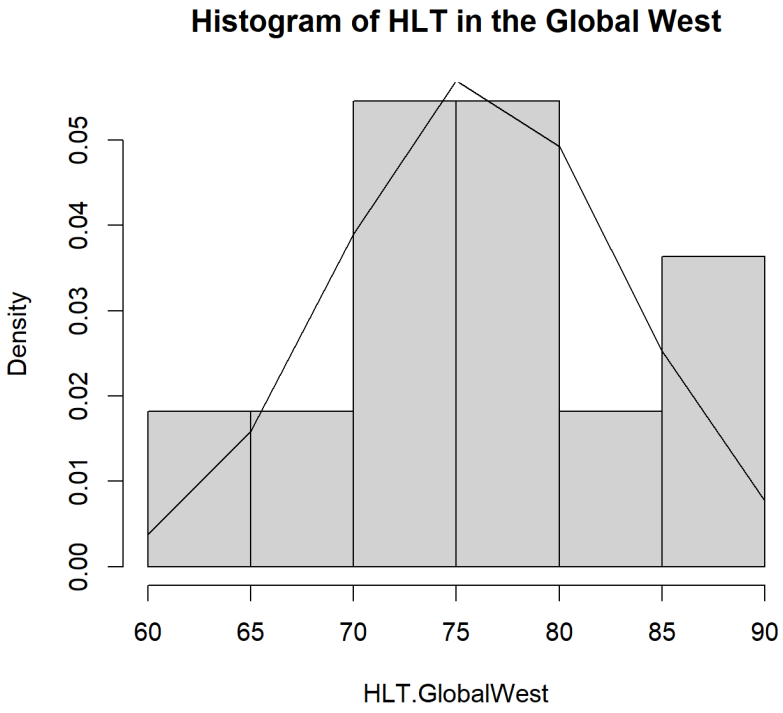
For this assignment I chose subset regions of the Global West and Eastern Europe during my analysis. The variable I chose to examine is HLT, which measures the environment health in our dataset.

1. Variable Distribution
    a. Boxplots for each region's environment health

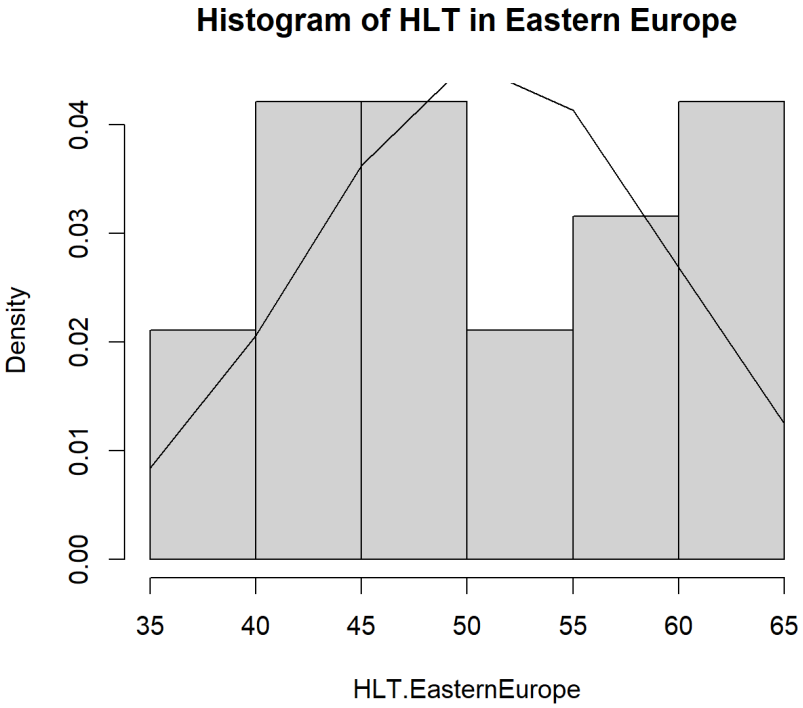


b. Histograms of each region

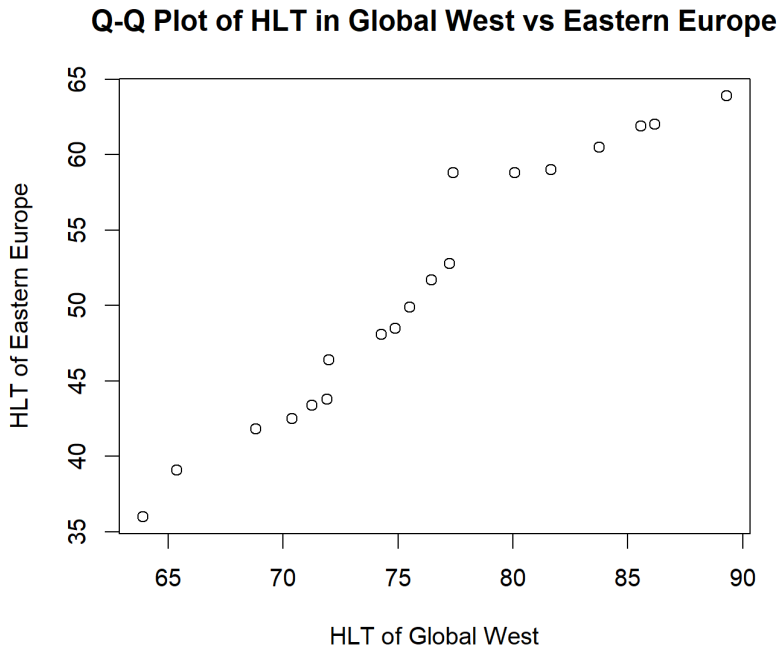
i. Histogram of environmental health in Global West



ii. Histogram of environmental health in Eastern Europe



c. QQ Plot of environmental health between Global West and Eastern Europe



2. Linear Models

I examined two linear models, one with Population as a predictor variable and environmental health (HLT) as a response variable, and one with GDP as a predictor variable and HLT as a response variable. After experimenting with cleaning the data by removing outliers and using logs, I found the two best predictors to be as follows: Log Cleaning on Population and Outlier Removal on GDP

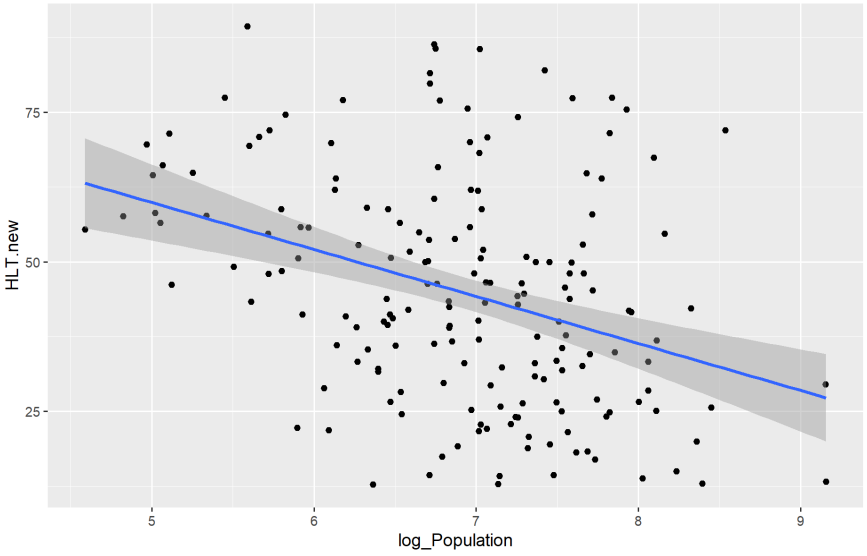
2.1 Linear Models for the full data set:

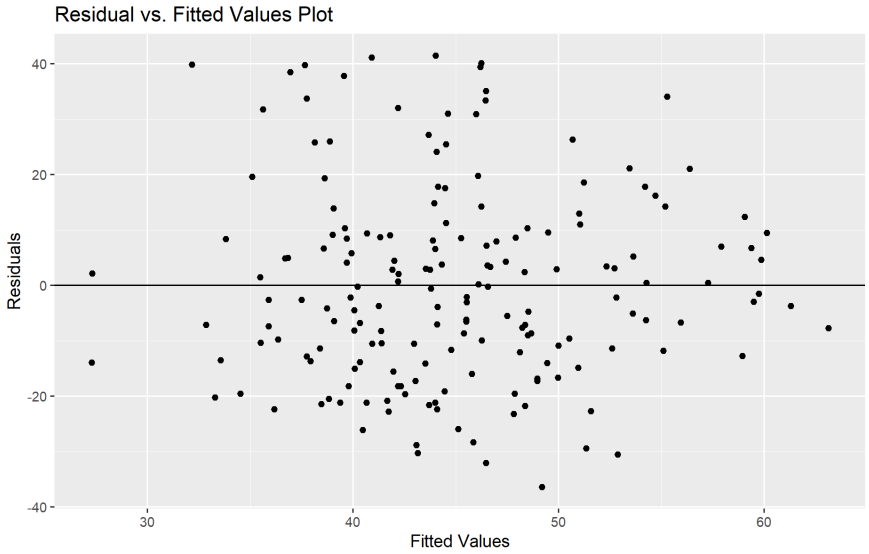
I. Linear Model 1:

- Predictor: Population, with log cleaning applied
- Response: Environmental Health (HLT)

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      99.130     10.709   9.257  < 2e-16 ***
log_Population   -7.846      1.538  -5.100 8.68e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.67 on 177 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.1281,    Adjusted R-squared:  0.1232
F-statistic: 26.01 on 1 and 177 DF,  p-value: 8.678e-07
```



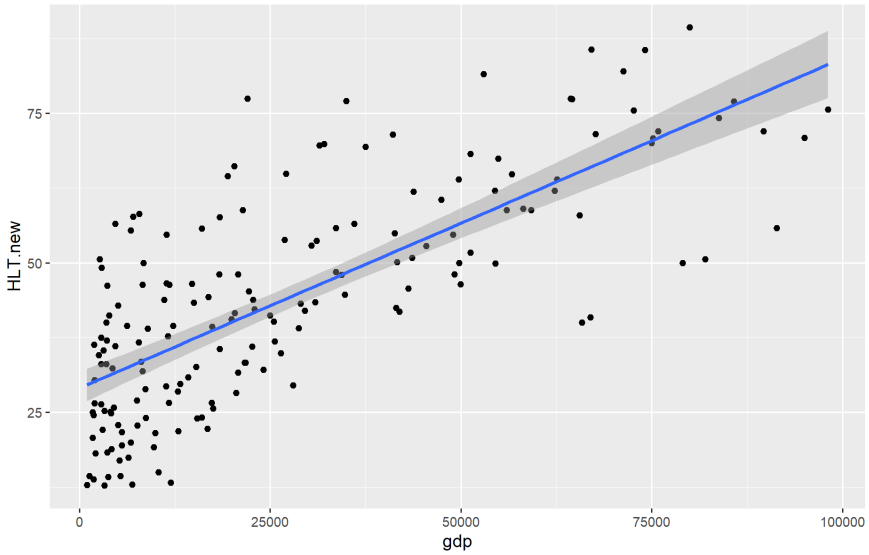


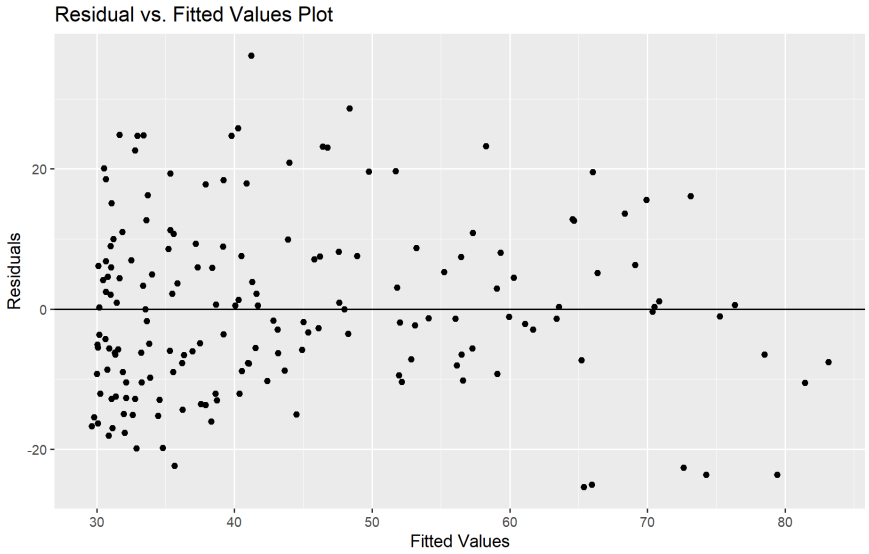
II. Linear Model 2:

- Predictor: GDP, with outliers removed
- Response: Environmental Health (HLT)

```
Coefficients:
             Estimate Std. Error t value
(Intercept) 2.908e+01  1.392e+00   20.89
gdp         5.508e-04  3.772e-05   14.60
            Pr(>|t|)
(Intercept)   <2e-16 ***
gdp           <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.37 on 172 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.5535,    Adjusted R-squared:  0.5509
F-statistic: 213.2 on 1 and 172 DF,  p-value: < 2.2e-16
```





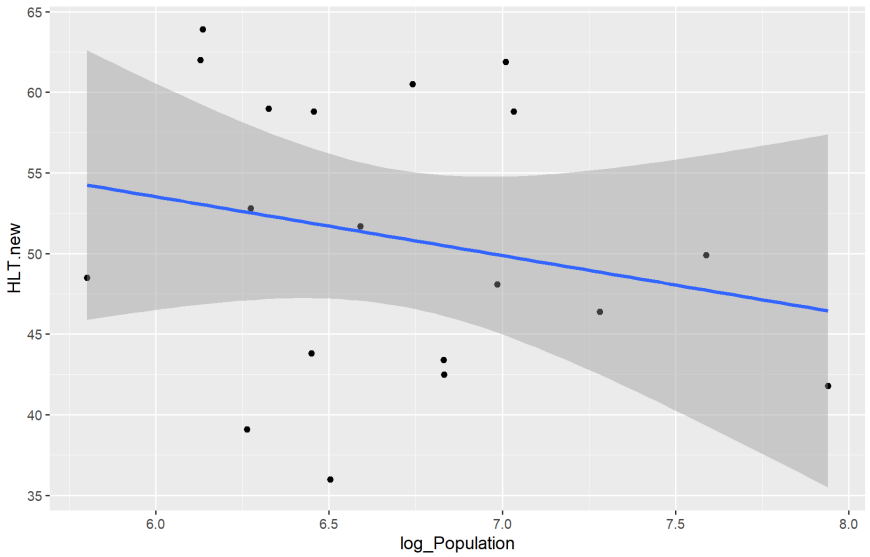
2.2 Models with subset of Eastern Europe:

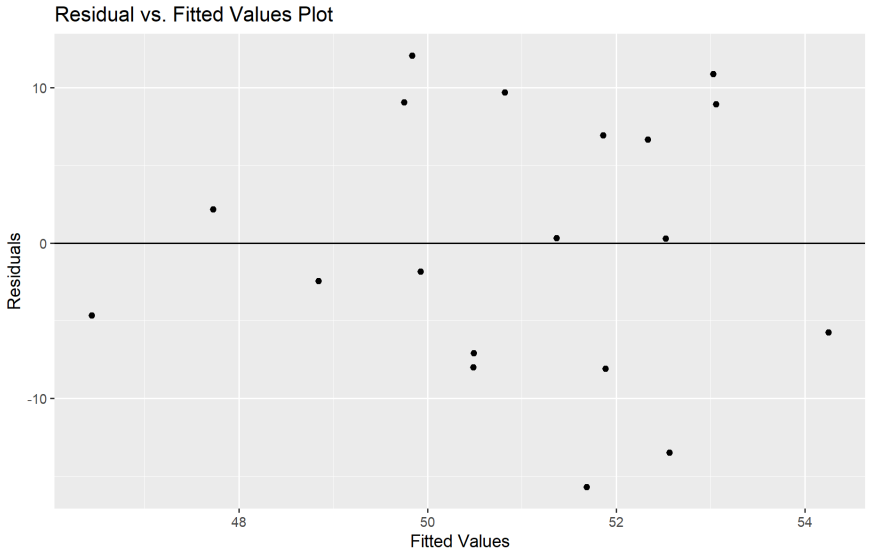
I. Linear Model 1:

- Predictor: Population, with log cleaning applied
- Response: Environmental Health (HLT)

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      75.434     25.799   2.924  0.00947 **
log_Population   -3.651      3.843  -0.950  0.35535
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.703 on 17 degrees of freedom
Multiple R-squared:  0.05043,   Adjusted R-squared:  -0.005431
F-statistic: 0.9028 on 1 and 17 DF,  p-value: 0.3554
```



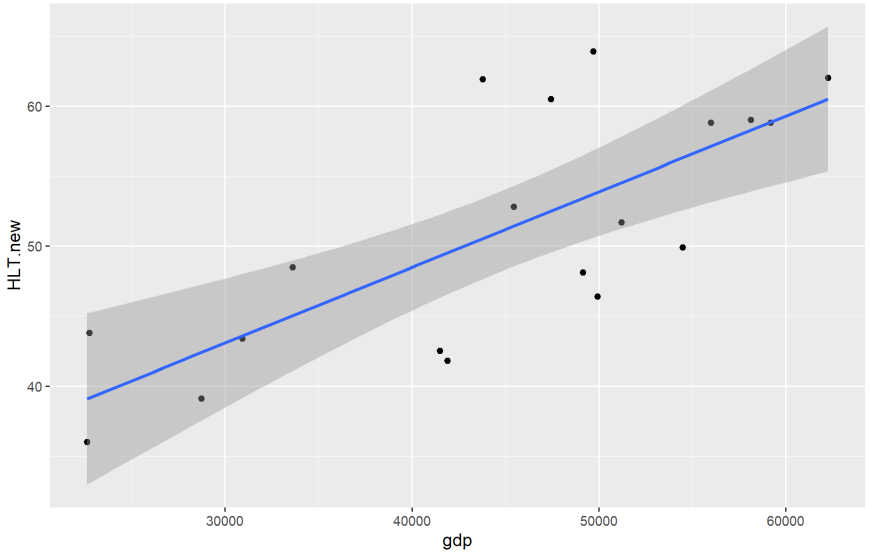


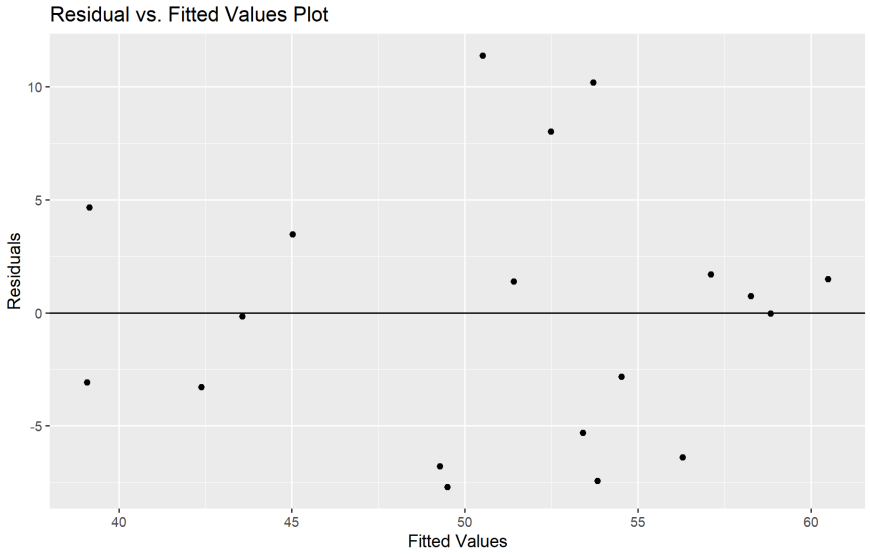
II. Linear Model 2:

- Predictor: GDP, with outliers removed
- Response: Environmental Health (HLT)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.688e+01  5.362e+00   5.013 0.000107 ***
gdp         5.398e-04  1.161e-04   4.649 0.000230 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.926 on 17 degrees of freedom
Multiple R-squared:  0.5598,    Adjusted R-squared:  0.5339
F-statistic: 21.62 on 1 and 17 DF,  p-value: 0.0002296
```





Based on our analysis, we conclude that the linear model using the outlier-cleaned GDP as the predictor variable and HLT as the response variable is the best fit. This model gave a much smaller p-value and higher R-squared values for both the full data set and also the subset region, Eastern Europe.

3. Classification (kNN)

For this part of the assignment I trained kNN models using “region” as the class label, and three input variables, BDH which measures Biodiversity & Habitat, AIR which measures Air Quality, and PCC which measures Climate Change. I trained models with k values of 3, 5, and 7. Next, I trained another set of models using “region” as the class label again, now with input variables AGR which measures Agriculture, WRS which measures Water Resources, and H2O which measures Sanitation and Drinking Water. I trained models with k values of 3, 5, and 7 once again. I then examined which models perform better.

3.1 First kNN models

- Class label: Region
- Input Variables: Biodiversity & Habitat (BDH), Air Quality (AIR), Climate Change (PCC)

I. Model with k = 3:

	Actual							
Predicted	Asia-Pacific	Eastern Europe	Former Soviet States	Global West	Greater Middle East	Latin America & Caribbean	Southern Asia	Sub-Saharan Africa
Asia-Pacific	17	1	2	0	0	5	2	1
Eastern Europe	1	11	0	0	0	1	0	1
Former Soviet States	0	3	3	0	1	1	0	2
Global West	0	1	0	19	0	2	0	0
Greater Middle East	2	0	3	0	11	3	0	2
Latin America & Caribbean	2	1	3	3	3	14	0	2
Southern Asia	1	0	0	0	0	0	4	0
Sub-Saharan Africa	2	2	1	0	1	6	2	38

```
> correct <- sum(diag(confusionMatrix0))
> total <- length(epi.data$region)
> accuracy <- correct / total
> print(paste("Amount Correct:", correct))
[1] "Amount Correct: 117"
> print(paste("Accuracy:", round(accuracy, 3)))
[1] "Accuracy: 0.65"
```

II. Model with k = 5:

Predicted \ Actual	Asia-Pacific	Eastern Europe	Former Soviet States	Global West	Greater Middle East	Latin America & Caribbean	Southern Asia	Sub-Saharan Africa
Asia-Pacific	14	0	1	0	3	5	2	5
Eastern Europe	0	11	2	0	0	1	0	2
Former Soviet States	0	3	1	0	0	3	0	1
Global West	1	2	0	20	0	2	0	0
Greater Middle East	3	0	2	0	10	3	0	4
Latin America & Caribbean	2	1	4	2	1	14	1	4
Southern Asia	1	0	0	0	0	0	3	0
Sub-Saharan Africa	4	2	2	0	2	4	2	30

```
> correct <- sum(diag(confusionMatrix1))
> accuracy <- correct / total
> print(paste("Amount Correct:", correct))
[1] "Amount Correct: 103"
> print(paste("Accuracy:", round(accuracy, 3)))
[1] "Accuracy: 0.572"
```

III. Model with k = 7:

Predicted \ Actual	Asia-Pacific	Eastern Europe	Former Soviet States	Global West	Greater Middle East	Latin America & Caribbean	Southern Asia	Sub-Saharan Africa
Asia-Pacific	15	0	1	0	3	3	3	4
Eastern Europe	0	9	1	0	1	1	0	2
Former Soviet States	0	1	0	0	0	2	0	1
Global West	1	2	0	20	0	3	0	0
Greater Middle East	3	0	3	0	6	3	1	5
Latin America & Caribbean	1	3	2	2	1	17	1	3
Southern Asia	1	0	0	0	0	0	1	0
Sub-Saharan Africa	4	4	5	0	5	3	2	31

```
> correct <- sum(diag(confusionMatrix2))
> accuracy <- correct / total
> print(paste("Amount Correct:", correct))
[1] "Amount Correct: 99"
> print(paste("Accuracy:", round(accuracy, 3)))
[1] "Accuracy: 0.55"
```

3.2 Second kNN models

- Class label: Region
- Input Variables: Agriculture (AGR), Water Resources (WRS), Sanitation and Drinking Water (H2O)

I. Model with k = 3:

Predicted \ Actual	Asia-Pacific	Eastern Europe	Former Soviet States	Global West	Greater Middle East	Latin America & Caribbean	Southern Asia	Sub-Saharan Africa
Asia-Pacific	13	0	1	0	1	3	3	1
Eastern Europe	3	13	1	0	0	2	0	0
Former Soviet States	1	1	6	1	1	2	0	0
Global West	1	4	0	20	0	0	0	0
Greater Middle East	0	0	1	1	9	4	0	0
Latin America & Caribbean	4	1	3	0	5	19	1	1
Southern Asia	0	0	0	0	0	0	2	0
Sub-Saharan Africa	3	0	0	0	0	2	2	44

```
> correct <- sum(diag(confusionMatrix3))
> accuracy <- correct / total
> print(paste("Amount Correct:", correct))
[1] "Amount Correct: 126"
> print(paste("Accuracy:", round(accuracy, 3)))
[1] "Accuracy: 0.7"
```

II. Model with k = 5:

Predicted \ Actual	Asia-Pacific	Eastern Europe	Former Soviet States	Global West	Greater Middle East	Latin America & Caribbean	Southern Asia	Sub-Saharan Africa
Asia-Pacific	9	1	2	0	2	7	4	1
Eastern Europe	1	12	2	2	1	1	0	0
Former Soviet States	2	0	4	0	1	0	0	0
Global West	2	5	0	19	0	0	0	0
Greater Middle East	0	0	1	1	8	3	0	0
Latin America & Caribbean	7	1	3	0	4	18	1	3
Southern Asia	2	0	0	0	0	0	2	0
Sub-Saharan Africa	2	0	0	0	0	3	1	42

```
> correct <- sum(diag(confusionMatrix4))
> accuracy <- correct / total
> print(paste("Amount Correct:", correct))
[1] "Amount Correct: 114"
> print(paste("Accuracy:", round(accuracy, 3)))
[1] "Accuracy: 0.633"
```

III. Model with k = 7:

Predicted	Actual: Asia-Pacific	Eastern Europe	Former Soviet States	Global West	Greater Middle East	Latin America & Caribbean	Southern Asia	Sub-Saharan Africa
Asia-Pacific	7	1	2	0	1	4	3	1
Eastern Europe	1	12	1	2	1	0	0	0
Former Soviet States	2	1	4	0	0	2	0	0
Global West	2	4	0	19	1	1	0	0
Greater Middle East	1	0	1	1	7	5	0	0
Latin America & Caribbean	6	1	4	0	6	19	1	3
Southern Asia	1	0	0	0	0	0	0	0
Sub-Saharan Africa	5	0	0	0	0	1	4	42

```
> correct <- sum(diag(confusionMatrix5))
> accuracy <- correct / total
> print(paste("Amount Correct:", correct))
[1] "Amount Correct: 110"
> print(paste("Accuracy:", round(accuracy, 3)))
[1] "Accuracy: 0.611"
```

Thus, based on our models, we can conclude the second kNN model using input variables Agriculture, Water Resources, Sanitation and Drinking Water performs better with our class label being the Region. This model consistently outperformed our other model, no matter if the k-value was 3, 5, or 7. Ultimately, the model performed best with k = 3, as it had the highest accuracy of 0.7 or 70%.