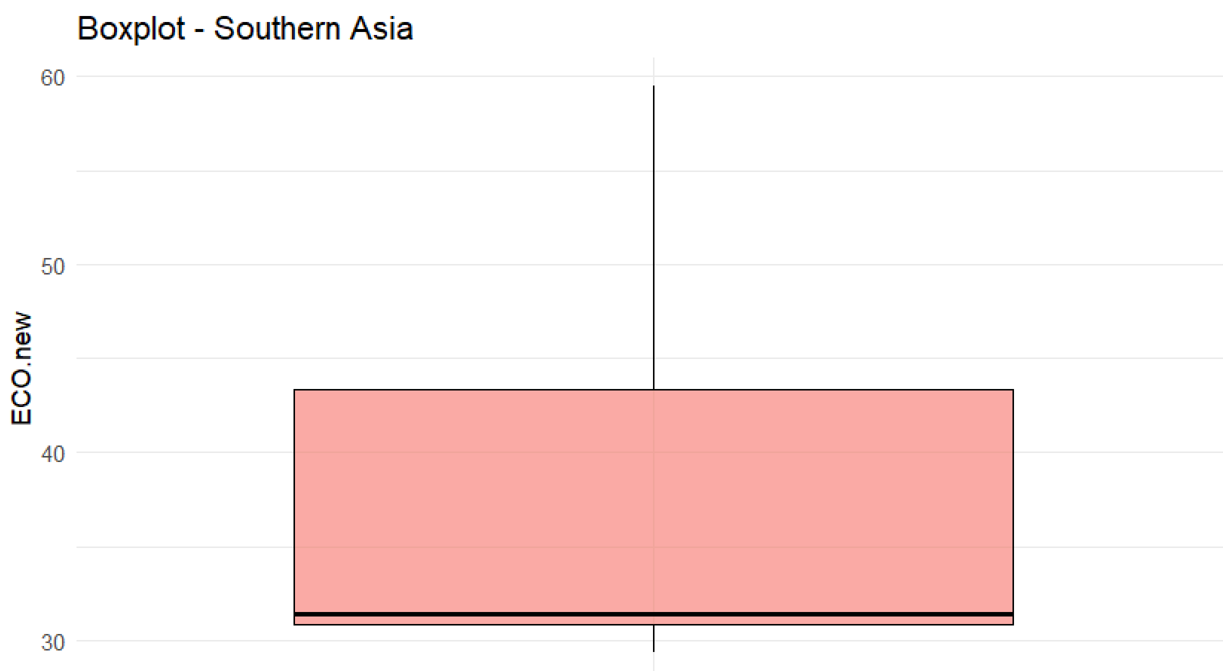
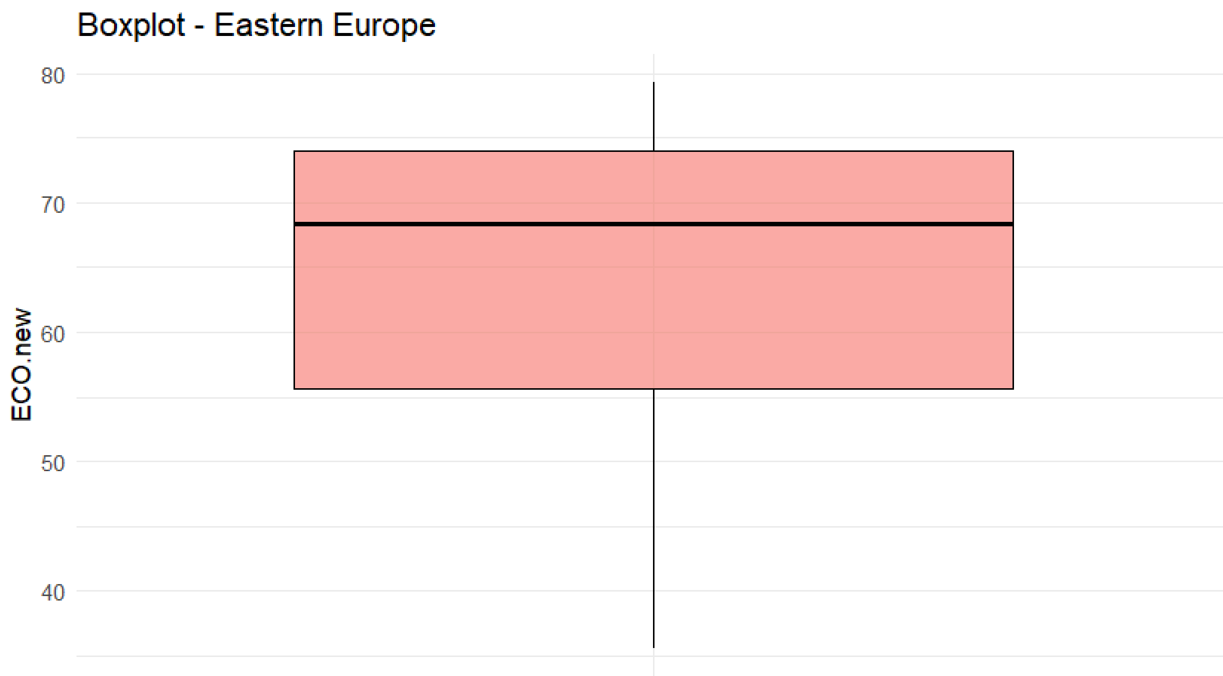
Alex Litchfield
10/10/2025
Assignment 2 Writeup

Line 12 - Display view of the read csv file

	code	iso	country	region	population	gdp	EPI.old	EPI.new	ECO.old	ECO.n
1	4	AFG	Afghanistan	Southern Asia	41454761	2116	18.0	30.7	21.1	
2	8	ALB	Albania	Eastern Europe	2811655	2273	45.9	52.1	50.3	
3	12	DZA	Algeria	Greater Middle East	46164219	1834	38.6	41.9	39.7	
4	24	AGO	Angola	Sub-Saharan Africa	36749906	991	31.6	39.7	35.9	
5	28	ATG	Antigua and Barbuda	Latin America & Caribbean	93316	31474	54.4	55.5	52.4	
6	32	ARG	Argentina	Latin America & Caribbean	45538401	3038	45.9	46.8	41.7	
7	51	ARM	Armenia	Former Soviet States	2943393	2497	42.5	44.7	46.8	
8	36	AUS	Australia	Global West	26451124	71310	59.0	63.0	60.7	
9	40	AUT	Austria	Global West	9130429	74981	68.9	69.0	78.4	
10	31	AZE	Azerbaijan	Former Soviet States	10318207	2548	40.4	40.4	44.7	
11	44	BHS	Bahamas	Latin America & Caribbean	399440	37517	54.6	56.0	54.7	
12	48	BHR	Bahrain	Greater Middle East	1569666	66975	37.1	35.9	45.9	
13	50	BGD	Bangladesh	Southern Asia	171466990	1037	25.5	27.8	27.3	
14	52	BRB	Barbados	Latin America & Caribbean	282336	22035	50.5	53.1	34.1	
15	112	BLR	Belarus	Former Soviet States	9115680	3360	49.3	58.1	60.4	
16	56	BEL	Belgium	Global West	11712893	75199	62.0	66.7	61.6	
17	84	BLZ	Belize	Latin America & Caribbean	411106	14958	46.5	47.4	55.8	

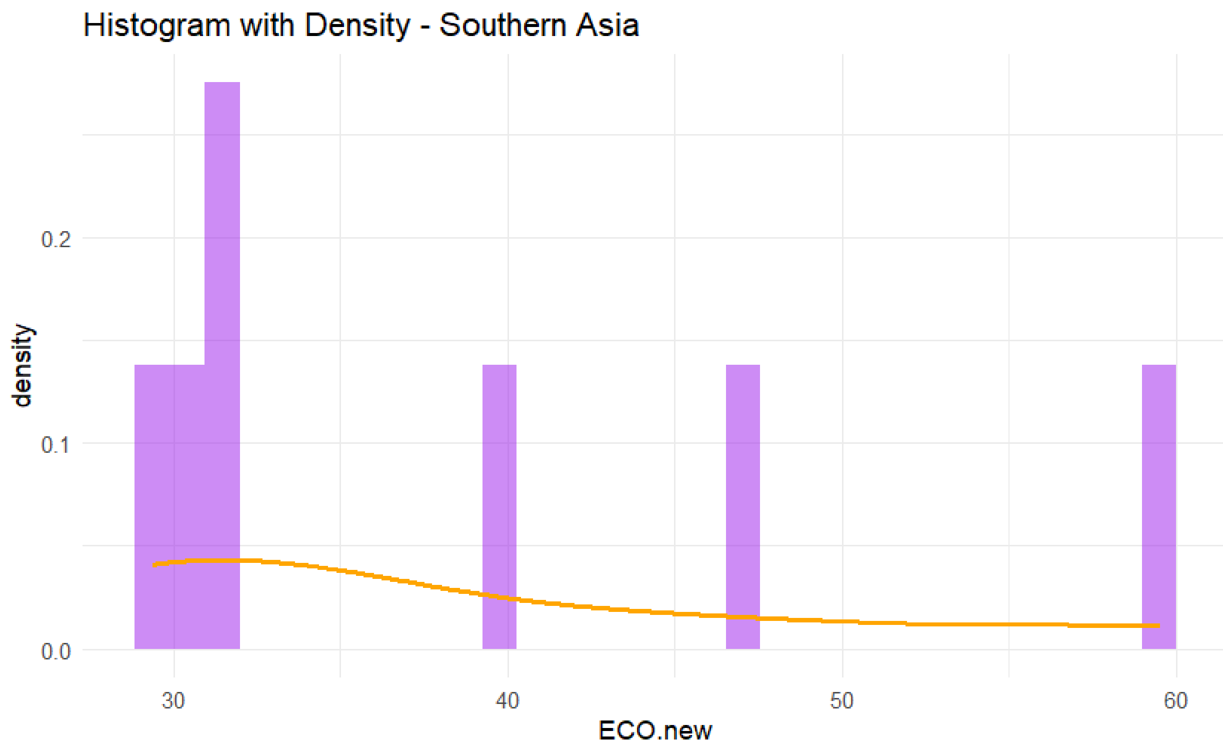
Line 35-50 - Create a boxplot for the variable “ECO.new” for each of the 2 selected regions, Southern Asia and Eastern Europe

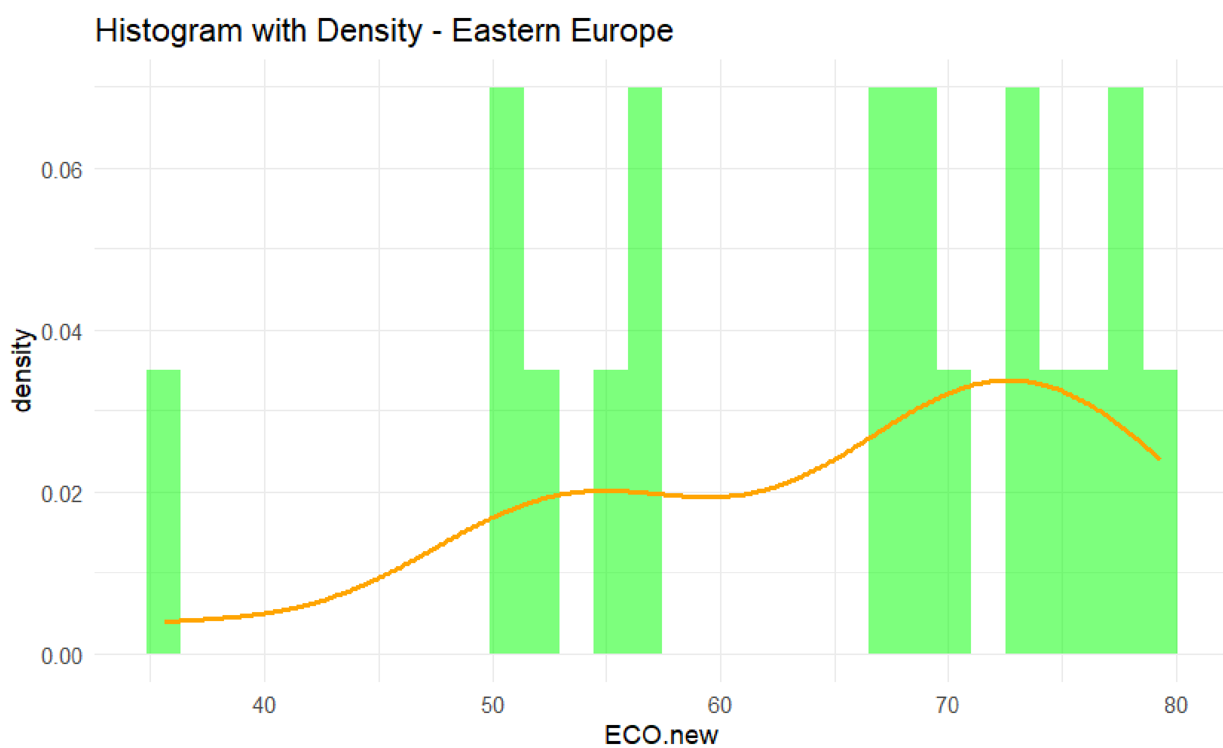




Line 52-66 - Create a histogram for the variable “ECO.new” for each of the 2 selected regions, Southern Asia and Eastern Europe. Each histogram has a density line overlaid.

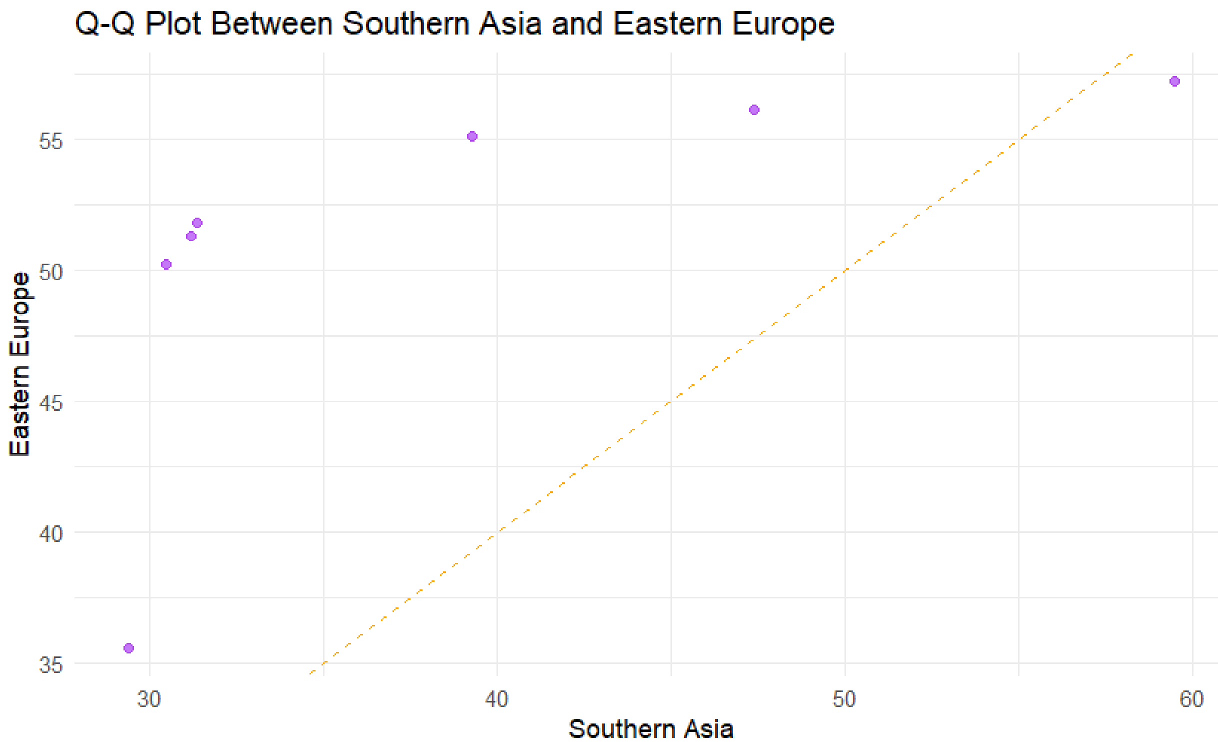
Alex Litchfield
10/10/2025
Assignment 2 Writeup





Alex Litchfield
10/10/2025
Assignment 2 Writeup

Line 71-91 - Plot a QQ for the same variable (“ECO.new”) between the 2 selected regions, Southern Asia and Eastern Europe.



Line 98-137 - Using the variable “ECO.new” as the variable as the response, fit 2 linear models. The predictor in linear model 1 is GDP and the predictor in linear model 2 is population. Log10 is applied as a transformation on each to improve the models. For each model I print the model summary stats, plot the most significant predictor vs the response, and plot the residuals.

Alex Litchfield
10/10/2025
Assignment 2 Writeup

Line 102-103 - Print model summaries

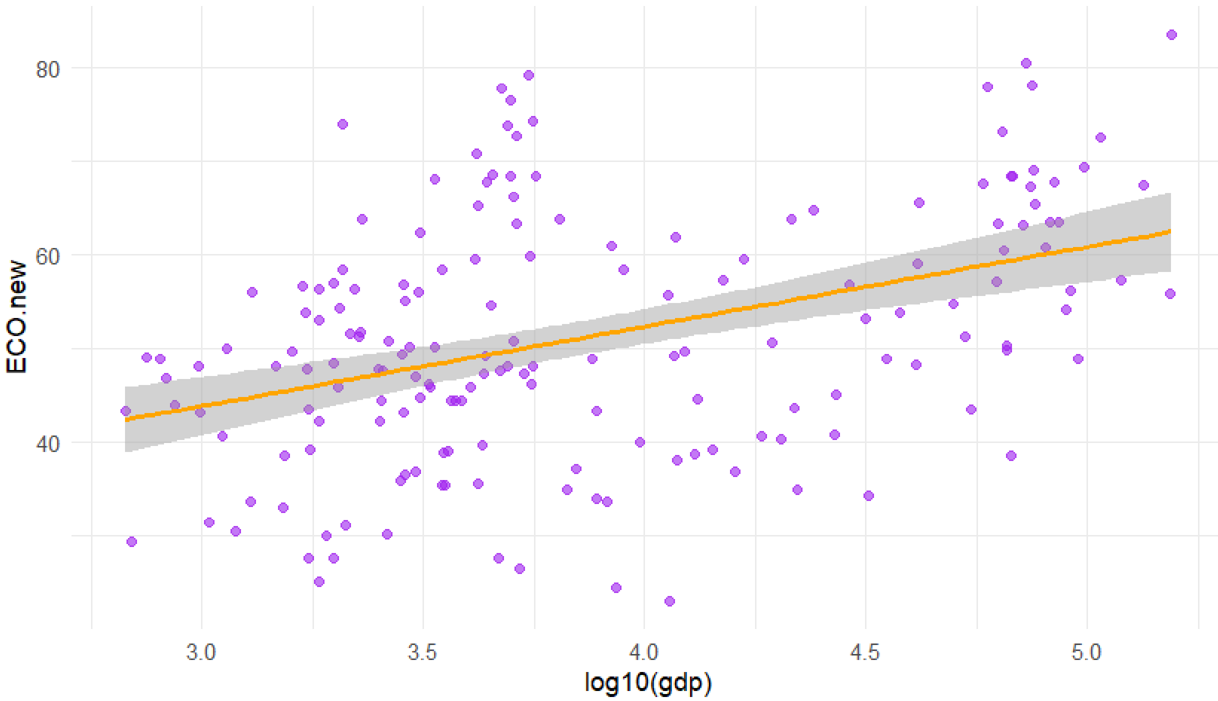
```
> summary(fullDatasetModel1)

Call:
lm(formula = (ECO.new) ~ log10(gdp), data = modifiedData)

Residuals:
     Min       1Q   Median       3Q      Max 
-29.7790  -9.3822   0.3461   8.3034  29.1476 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   18.354      5.762   3.186  0.00171 ** 
log10(gdp)     8.510      1.469   5.794 3.18e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.14 on 173 degrees of freedom
Multiple R-squared:  0.1625,	Adjusted R-squared:  0.1577 
F-statistic: 33.57 on 1 and 173 DF,  p-value: 3.177e-08
```

```
> summary(fullDatasetModel2)

Call:
lm(formula = (ECO.new) ~ log10(population), data = modifiedData)

Residuals:
    Min      1Q  Median      3Q     Max 
-27.895  -8.807  -1.623   9.327  32.577 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        49.4813     8.2033   6.032 9.56e-09 ***
log10(population)   0.2647     1.1763   0.225    0.822    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.26 on 173 degrees of freedom
Multiple R-squared:  0.0002927,	Adjusted R-squared:  -0.005486 
F-statistic: 0.05065 on 1 and 173 DF,  p-value: 0.8222
```

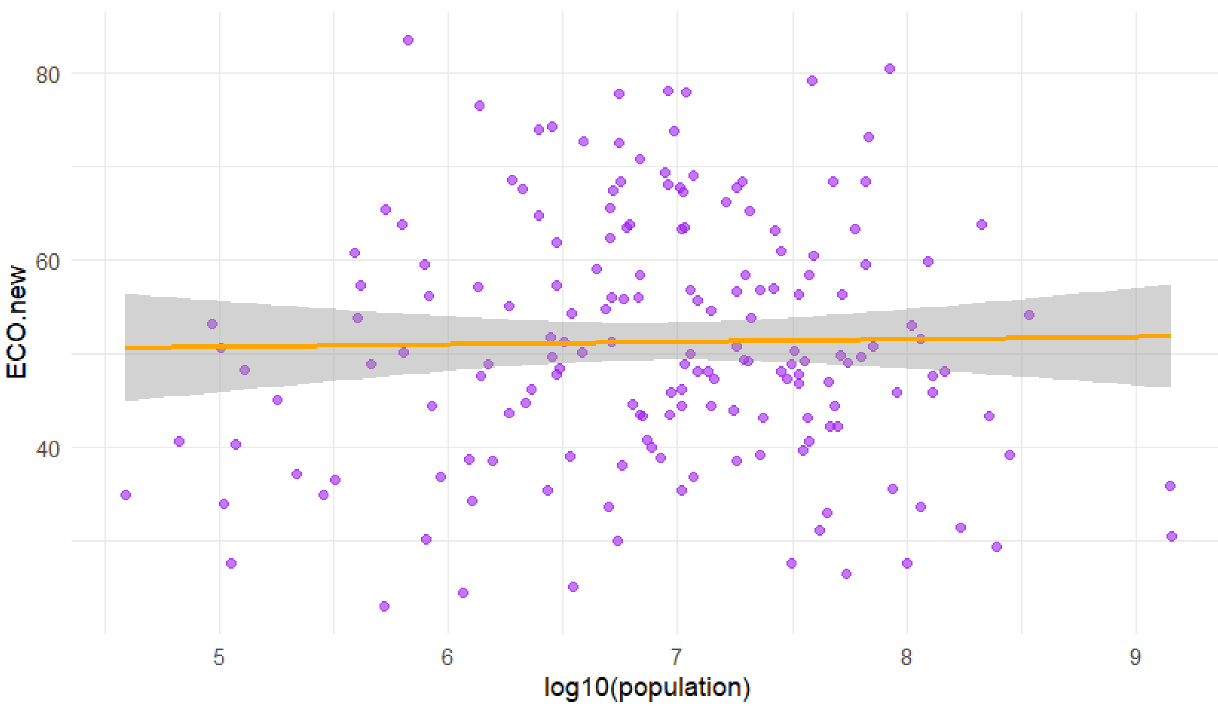
Line 110-121 - Plot the most significant predictor vs response

Alex Litchfield
10/10/2025
Assignment 2 Writeup

ECO.new vs log10(GDP)

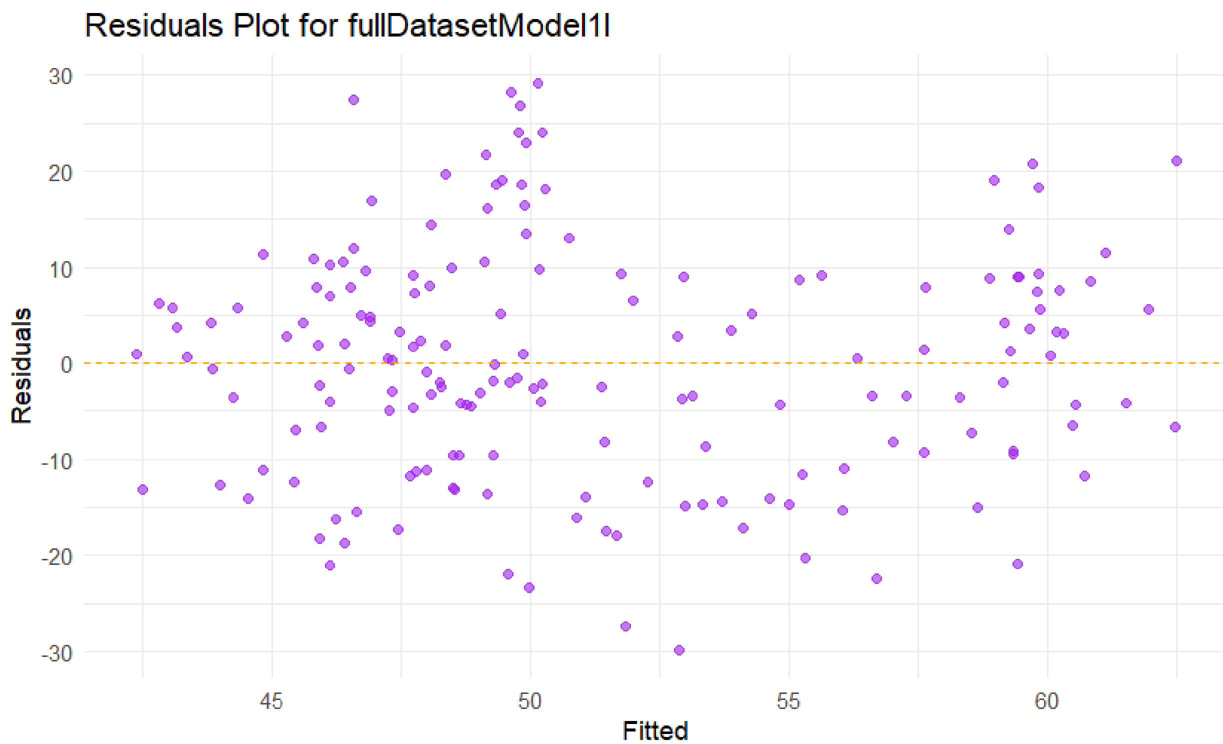


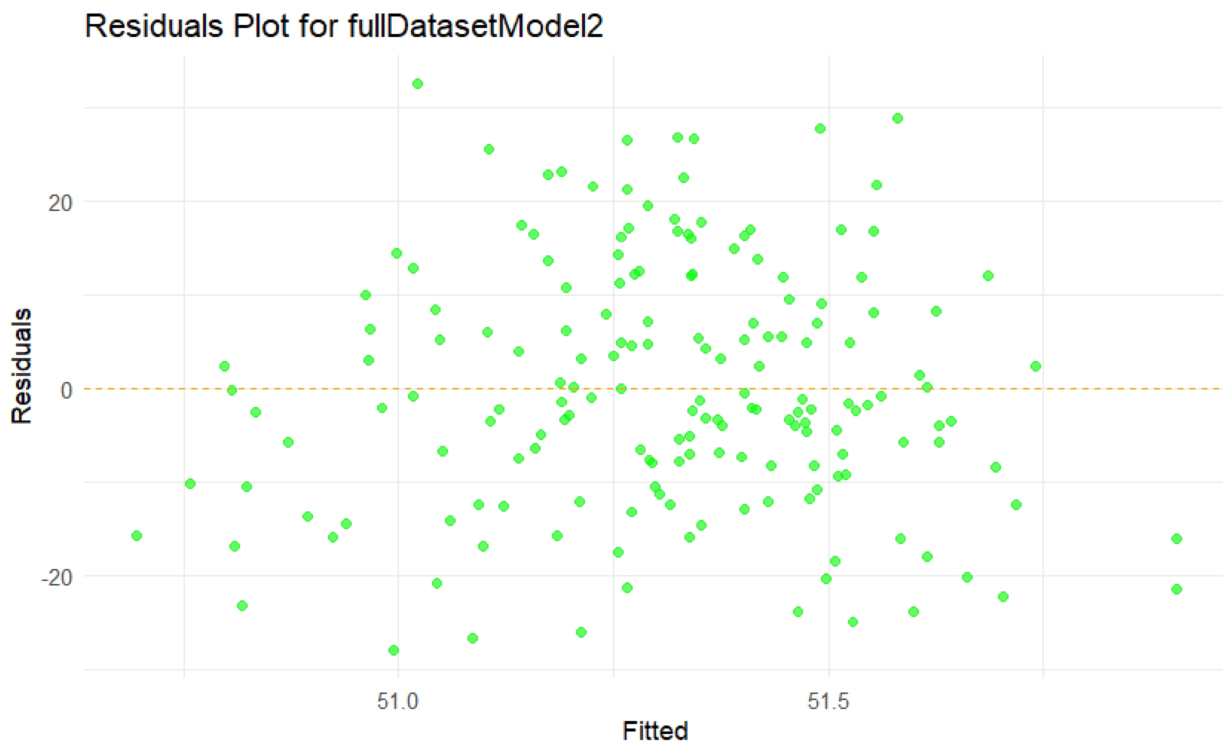
ECO.new vs log10(Population)



Alex Litchfield
10/10/2025
Assignment 2 Writeup

Line 124-137 - Plot the residuals





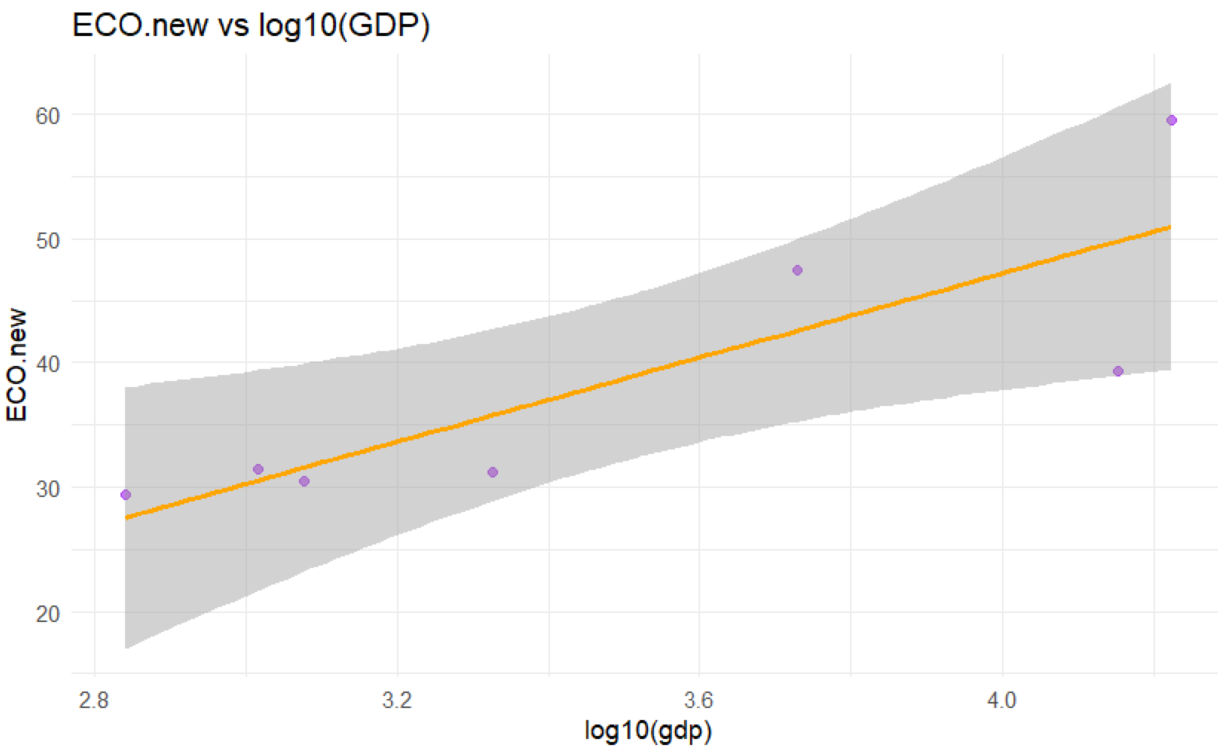
Line 141-176 - Repeat the previous models with a subset of 1 region (Southern Asia)

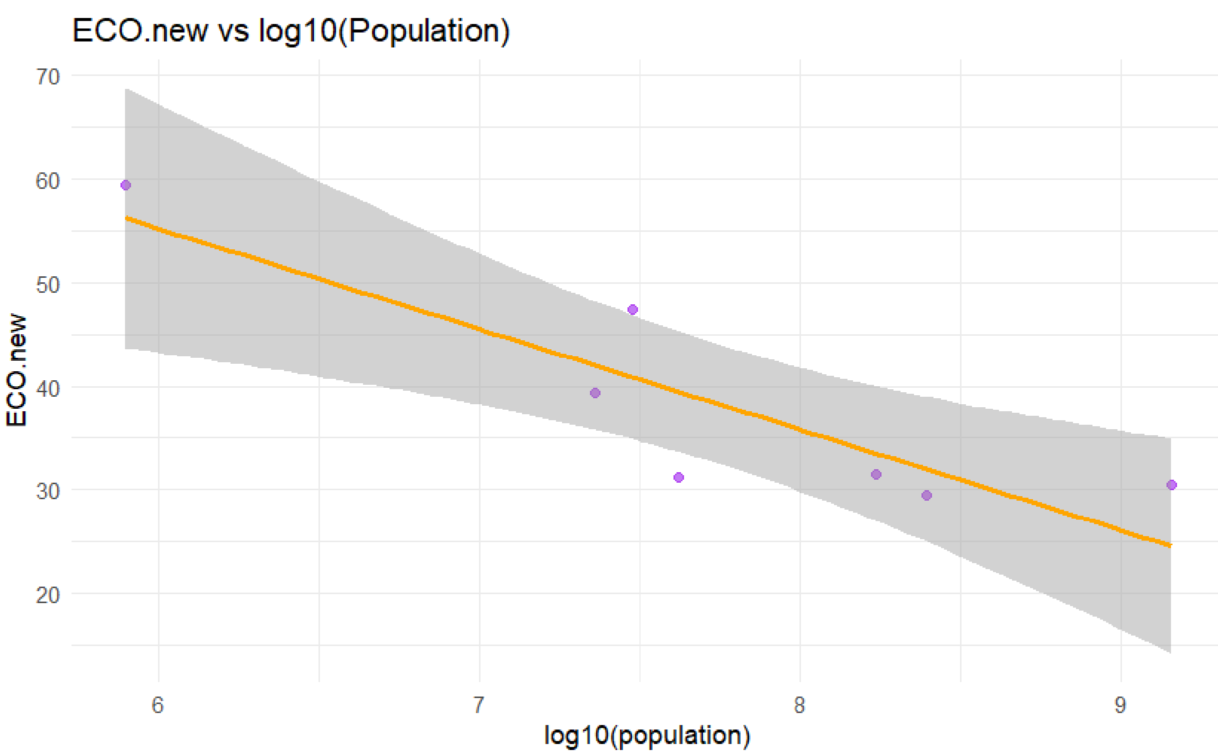
Alex Litchfield
10/10/2025
Assignment 2 Writeup

Line 145-146 - Print model summaries

```
> summary(subRegionData1)

Call:
lm(formula = (ECO.new) ~ log10(gdp), data = subsetRegion1)

Residuals:
      1        2        3        4        5        6        7 
-4.5586   0.8859   8.5260  -1.0508   4.8232   1.8605 -10.4862 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  -20.550     17.481  -1.176   0.2927  
log10(gdp)    16.932      4.967   3.409   0.0191 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.812 on 5 degrees of freedom
Multiple R-squared:  0.6991,	Adjusted R-squared:  0.639 
F-statistic: 11.62 on 1 and 5 DF,  p-value: 0.01907

> summary(subRegionData2)

Call:
lm(formula = (ECO.new) ~ log10(population), data = subsetRegion1)

Residuals:
     1      2      3      4      5      6      7 
-8.314 -2.126  3.262  5.944  6.517 -2.578 -2.704 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)   
(Intercept)        113.495     18.405   6.167  0.00163 **
log10(population)   -9.712      2.362  -4.112  0.00925 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.934 on 5 degrees of freedom
Multiple R-squared:  0.7717,	Adjusted R-squared:  0.7261 
F-statistic: 16.91 on 1 and 5 DF,  p-value: 0.009249
```

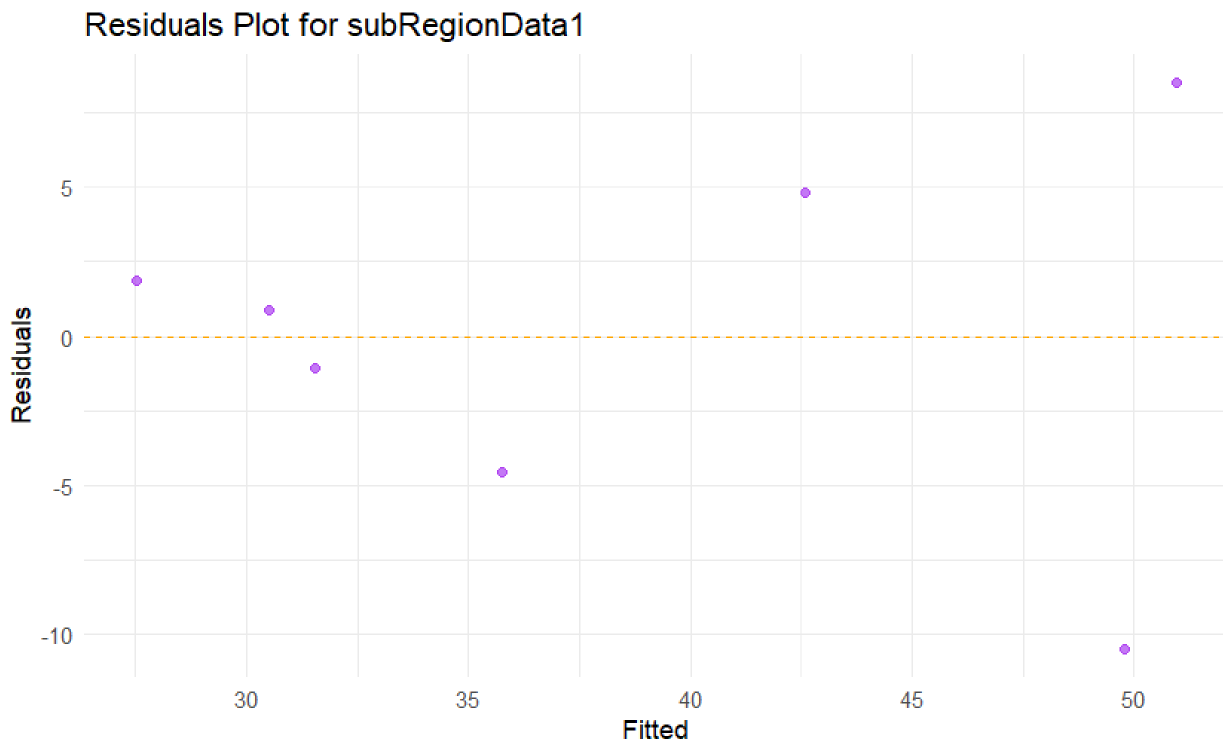
Line 149-160 - Plot the most significant predictor vs the response

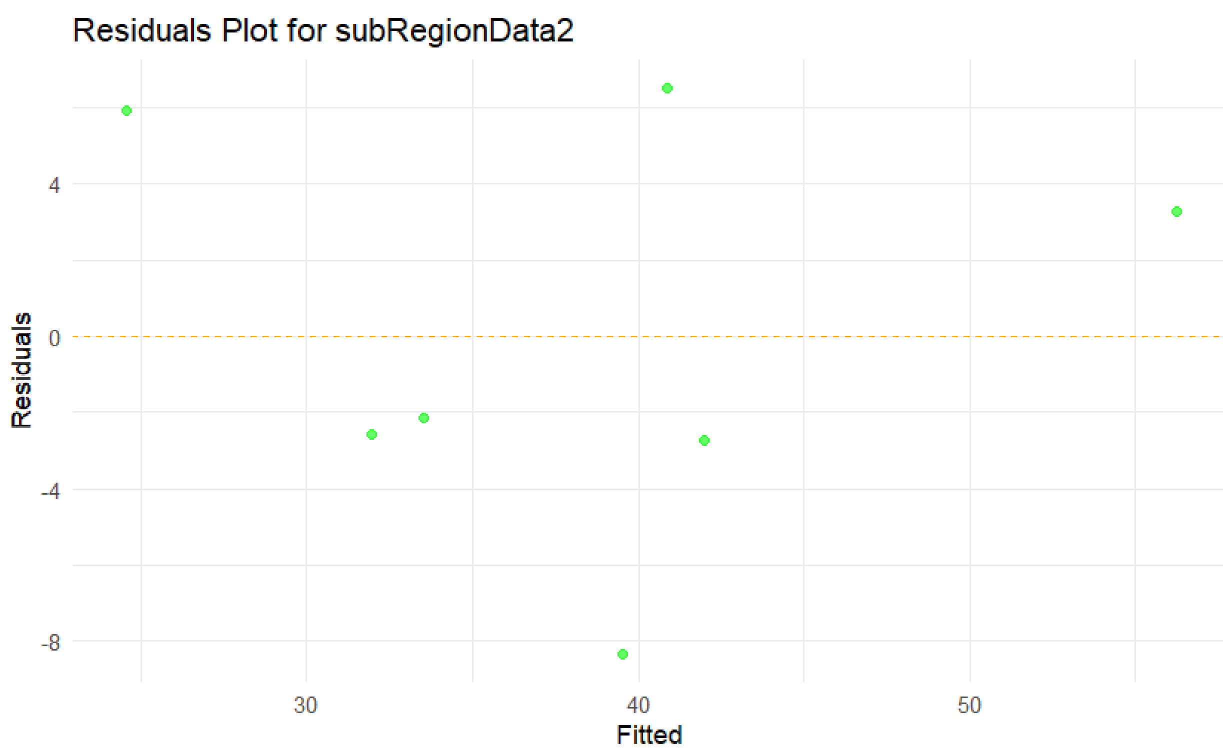




Line 163-176 - Plot the residuals

Alex Litchfield
10/10/2025
Assignment 2 Writeup





The first linear model using the entire modified dataset is the best fit because it has the highest accuracy with many different datapoints, as seen by the density line and lower p value (3.177e-08 vs 0.8222). However, the second linear model is better if we are using the subset region for the data because it has a lower p value (0.009249 vs 0.01907) and the density lines are similar.

Line 181-205 - Train a kNN model using “region” as a class label and choose 3 variables (not population or gdp) as inputs to the model. I used "EPI.new", "ECO.new", and "BDH.new". Evaluate the model using a confusion matrix and calculate the accuracy of correct classifications. Accuracy = correctly classified/total data points. The model runs with different versions of k until it finds the one with the best accuracy, and the confusion matrix is only printed for the iteration of k that fits this standard.

```
[1] "k value= 1 Accuracy = 0.833333333333333"
[1] "k value= 2 Accuracy = 1"
Confusion Matrix and Statistics

                Reference
Prediction        Eastern Europe Southern Asia
  Eastern Europe               4             0
  Southern Asia                0             2

               Accuracy : 1
                 95% CI : (0.5407, 1)
    No Information Rate : 0.6667
    P-Value [Acc > NIR] : 0.08779

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.6667
         Detection Rate : 0.6667
   Detection Prevalence : 0.6667
      Balanced Accuracy : 1.0000

       'Positive' Class : Eastern Europe
```

Line 210-227 - Repeat the previous model with 3 other variables and the same k value. I used "SPI.new", "BER.new", and "RLI.new" for the variables and the k value of 2.

Alex Litchfield
10/10/2025
Assignment 2 Writeup

```
> print(paste("k value=", 3, "Accuracy =", accuracy))
[1] "k value= 3 Accuracy = 1"
> print(confusionMatrixVar)
Confusion Matrix and Statistics

                Reference
Prediction       Eastern Europe Southern Asia
  Eastern Europe              6             0
  Southern Asia               0             0

               Accuracy : 1
                 95% CI : (0.5407, 1)
    No Information Rate : 1
    P-Value [Acc > NIR] : 1

                  Kappa : NaN

 Mcnemar's Test P-Value : NA

            Sensitivity :  1
            Specificity : NA
         Pos Pred Value : NA
         Neg Pred Value : NA
             Prevalence :  1
         Detection Rate :  1
   Detection Prevalence :  1
      Balanced Accuracy : NA

       'Positive' Class : Eastern Europe
```

The first version of the model (using k=2) is better because the confusion matrix in model 1 has more diversity in the confusion matrix than model 2. The accuracy of both models is equivalent, so I would only go off of the diversity of the accurate predictions provided by each model.