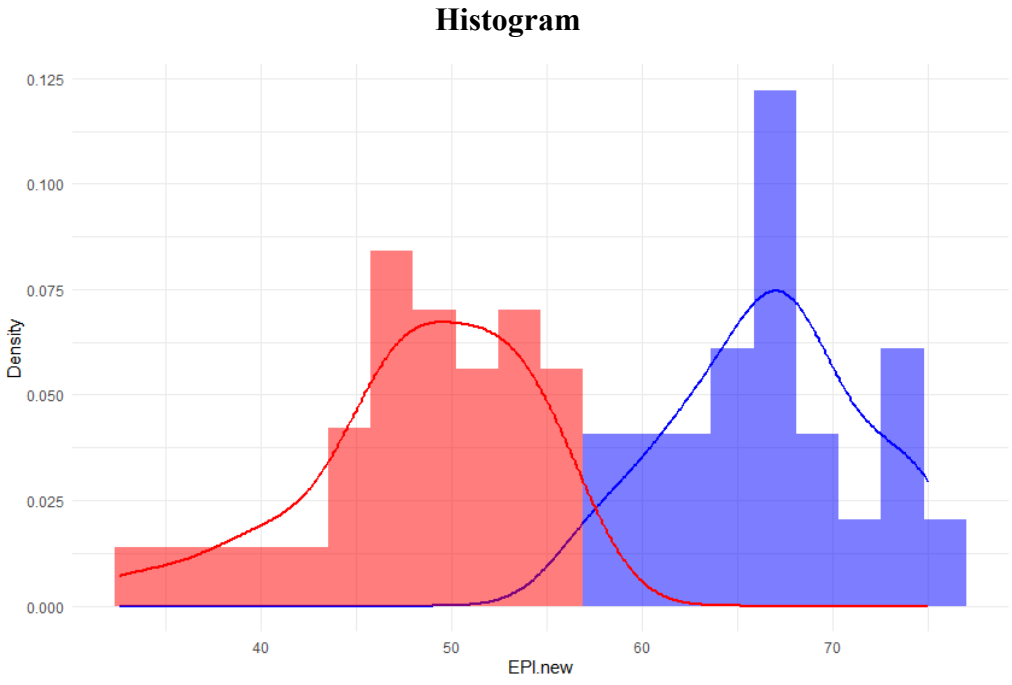
Inah Lee
Data Analytics
2/17/2025

Environmental Indicator Modeling and Regional Classification

Variable Distributions

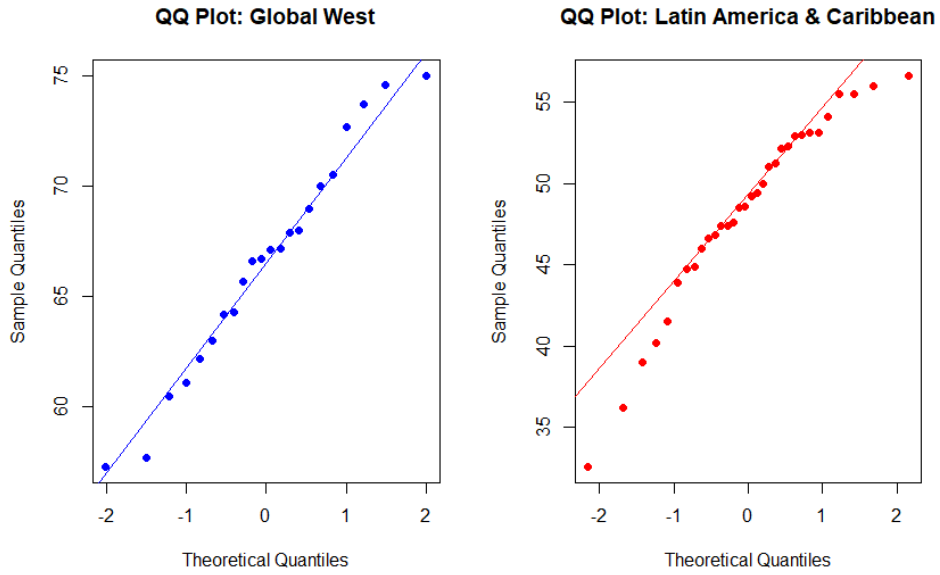
From the epi_data dataset, two regions were selected for comparison: Global West and Latin America & Caribbean. The Global West primarily consists of developed countries, while Latin America & Caribbean includes several developing countries.

west_data <- subset(epi_data, region == "Global West") color = "blue"
latin_data <- subset(epi_data, region == "Latin America & Caribbean") color = "red"



The Global West region generally exhibits higher EPI.new scores, indicating stronger overall environmental performance. In contrast, the Latin America & Caribbean region shows a wider distribution of scores, with a concentration in the middle to lower ranges. This suggests greater variability in environmental performance among countries in Latin America & Caribbean compared to the more consistently high scores observed in the Global West.

QQ Plot



For the Global West region, the points in the Q-Q plot closely follow the normal reference line, indicating that the distribution of scores is approximately normal with relatively small deviations. In contrast, for the Latin America & Caribbean region, there are more pronounced deviations from the reference line, particularly at the lower and upper ends of the distribution. This pattern suggests that countries in the Global West have relatively similar environmental performance scores, whereas the Latin America & Caribbean region exhibits greater variability between countries.

Linear Models

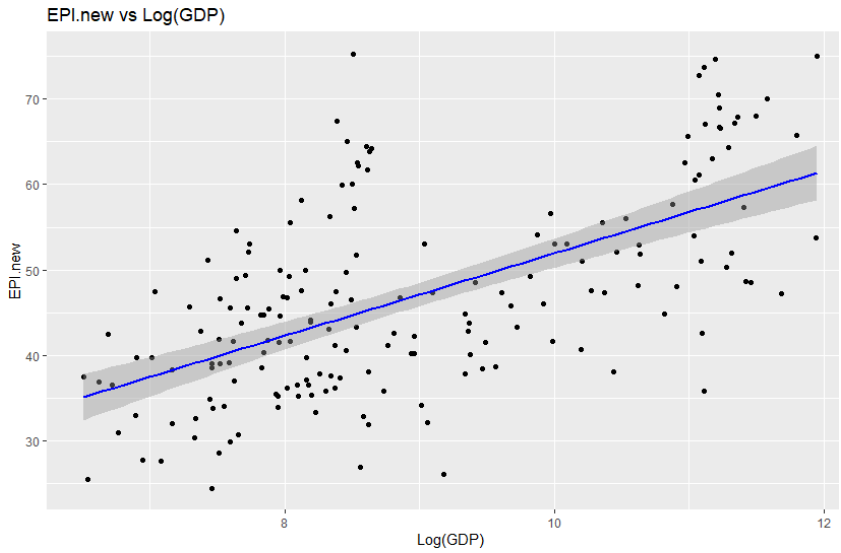
Two response variables were selected: EPI.new and ECO.new. Linear regression models were constructed to evaluate the relationships between these response variables and two predictors: GDP and population.

```
epi_model_gdp <- lm(EPI.new ~ gdp, data = epi_data)
eco_model_gdp <- lm(ECO.new ~ gdp, data = epi_data)
epi_model_pop <- lm(EPI.new ~ population, data = epi_data)
eco_model_pop <- lm(ECO.new ~ population, data = epi_data)
```

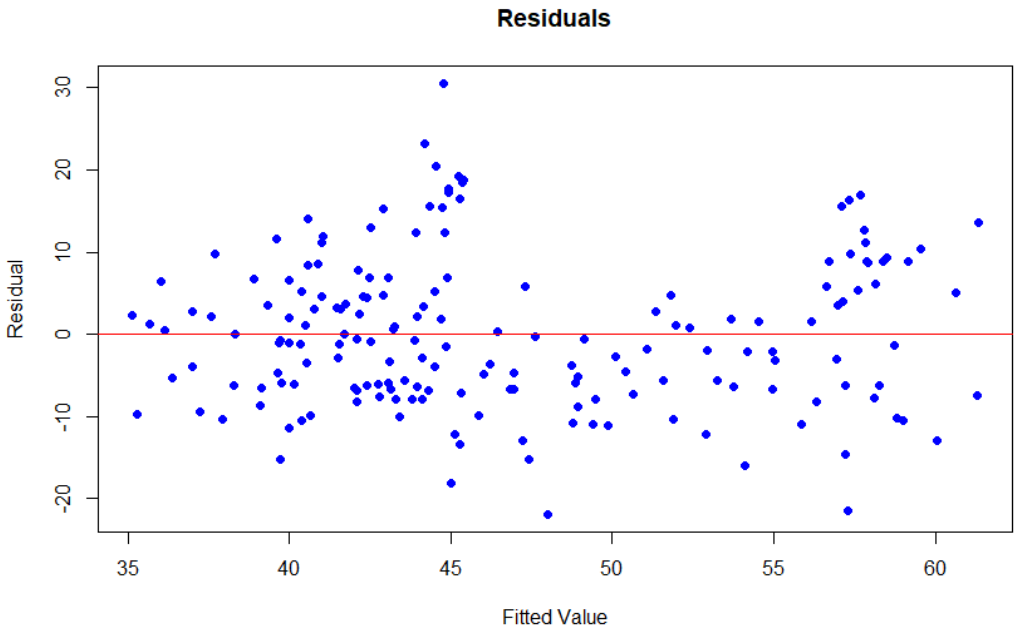
Model results:
EPI.new ~ GDP: Intercept = 42.40, GDP coefficient = 2.06e-04 ($p < 2e-16$), $R^2 = 0.327$
ECO.new ~ GDP: Intercept = 47.46, GDP coefficient = 1.70e-04 ($p < 1e-8$), $R^2 = 0.171$
EPI.new ~ Population: Intercept = 47.50, Population coefficient = -1.35e-08 ($p = 0.0147$), $R^2 = 0.033$
ECO.new ~ Population: Intercept = 51.85, Population coefficient = -1.45e-08 ($p = 0.021$), $R^2 = 0.030$

Interpretation

GDP showed a stronger relationship with environmental performance compared to population size. For EPI.new, GDP explained 32.7% of the variance, while for ECO.new, GDP explained only 17.1%. The population had very low explanatory power, with R^2 values of 3.3% for EPI.new and 3.0% for ECO.new. These results suggest that while GDP has a notable influence on environmental performance, other factors are also important. Among the tested predictors, GDP was the most significant predictor, particularly for EPI.new.



A log transformation was applied to GDP to stabilize variance. The transformed data showed that EPI.new generally increases as GDP increases, although some variability remains, indicating that factors other than GDP also influence environmental performance.



The residuals are randomly distributed around zero, indicating that the regression model fits the data reasonably well. However, a few observations have residuals greater than 20 or less than -20, suggesting the presence of potential outliers or cases where the model underperforms in prediction.

Repeat with one region

For a regional analysis, the Global West subset was selected, and a linear regression model was applied using EPI.new as the response variable and GDP (log-transformed) as the predictor.

```
west_model_gdp <- lm(EPI.new ~ gdp, data = west_data)
> summary(west_model_gdp)
```

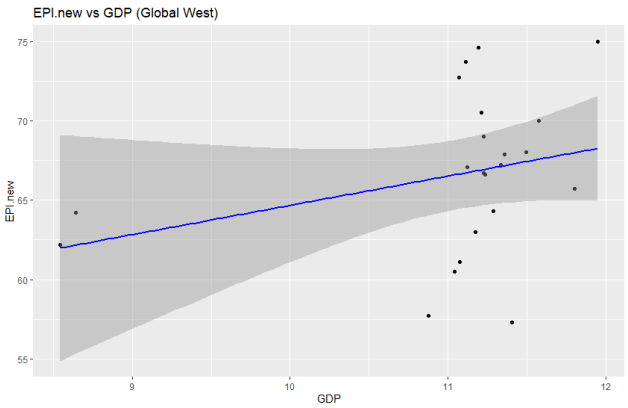
Result:

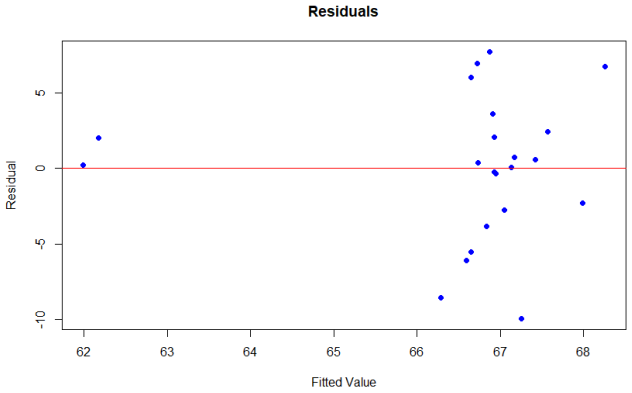
Intercept = 62.21 ($p < 0.001$)

GDP coefficient = 5.79e-05 ($p = 0.0861$)

$R^2 = 0.1401$

F-statistic = 3.259, $p = 0.0861$





When compared with the global model, the Global West model shows a slightly higher R^2 value. However, because countries in the Global West already have relatively high GDP levels, GDP has a weaker influence on EPI.new within this region. This suggests that GDP is not the most suitable predictor for environmental performance in the Global West, and other socio-environmental factors may better explain the variation.

Classification (k-NN)

A subset of the data containing only two regions, Global West and Latin America & Caribbean, was used for classification. Three predictors, EPI.new, ECO.new, and GHN.new, were selected. The dataset was split into a 70 percent training set and a 30 percent testing set using the createDataPartition() function.

A k-nearest neighbors (k-NN) model with k = 20 was trained to classify regions, and performance was evaluated using a confusion matrix and accuracy.

Confusion Matrix

	Reference	
Prediction	Global West	Latin America & Caribbean
Global West	5	0
Latin America & Caribbean	1	9

Performance Metrics

Accuracy: 93.33%

Kappa: 0.8571

Sensitivity: 0.8333

Specificity: 1.0000

Balanced Accuracy: 0.9167

P-value [Acc > NIR]: 0.0052

Interpretation

The model achieved a high classification accuracy of 93.33 percent with k = 20. The dataset contained 22 observations for Global West and 32 for Latin America & Caribbean. Due to the

relatively small number of samples in each region, only a few data points remained in the test set after splitting, which likely contributed to the high observed accuracy.

For several values of k, the first model using EPI.new, ECO.new, and GHN.new as predictors achieved 100 percent accuracy when k was set to 1 or 5. Accuracy decreased slightly to 93.33 percent for k values of 10 and 20, but still remained high. When k was increased to 30, accuracy dropped significantly to 73 percent. This suggests that the high accuracy at lower k values may be due to the small dataset size or to the possibility that the two regions differ markedly in the three selected predictors. The sharp decline at k = 30 indicates that too large a k value reduces the model's ability to distinguish between the two regions effectively.

The process was repeated using three different predictors: GTP.new, LUF.new, and GRI.new. The confusion matrix for this model is shown below.

Confusion Matrix

	Reference	
Prediction	Global West	Latin America & Caribbean
Global West	5	1
Latin America & Caribbean	1	8

Performance Metrics

Accuracy: 86.67%

Kappa: 0.7222

Sensitivity: 0.8333

Specificity: 0.8889

Balanced Accuracy: 0.8611
P-value [Acc > NIR]: 0.0271

For this second model, accuracy was 93.33 percent at k = 1 and 5, dropped to 86.67 percent at k = 10 and 20, and fell further to 66.67 percent at k = 30.

Comparison

The first model, which used EPI.new, ECO.new, and GHN.new as predictors, achieved a higher peak accuracy of 93.33 percent compared to 86.67 percent for the second model. This suggests that EPI.new, ECO.new, and GHN.new are more effective predictors for distinguishing between the Global West and Latin America & Caribbean than GTP.new, LUF.new, and GRI.new.