

exploratory data analysis and models on the epi dataset

date: 2025-10-13

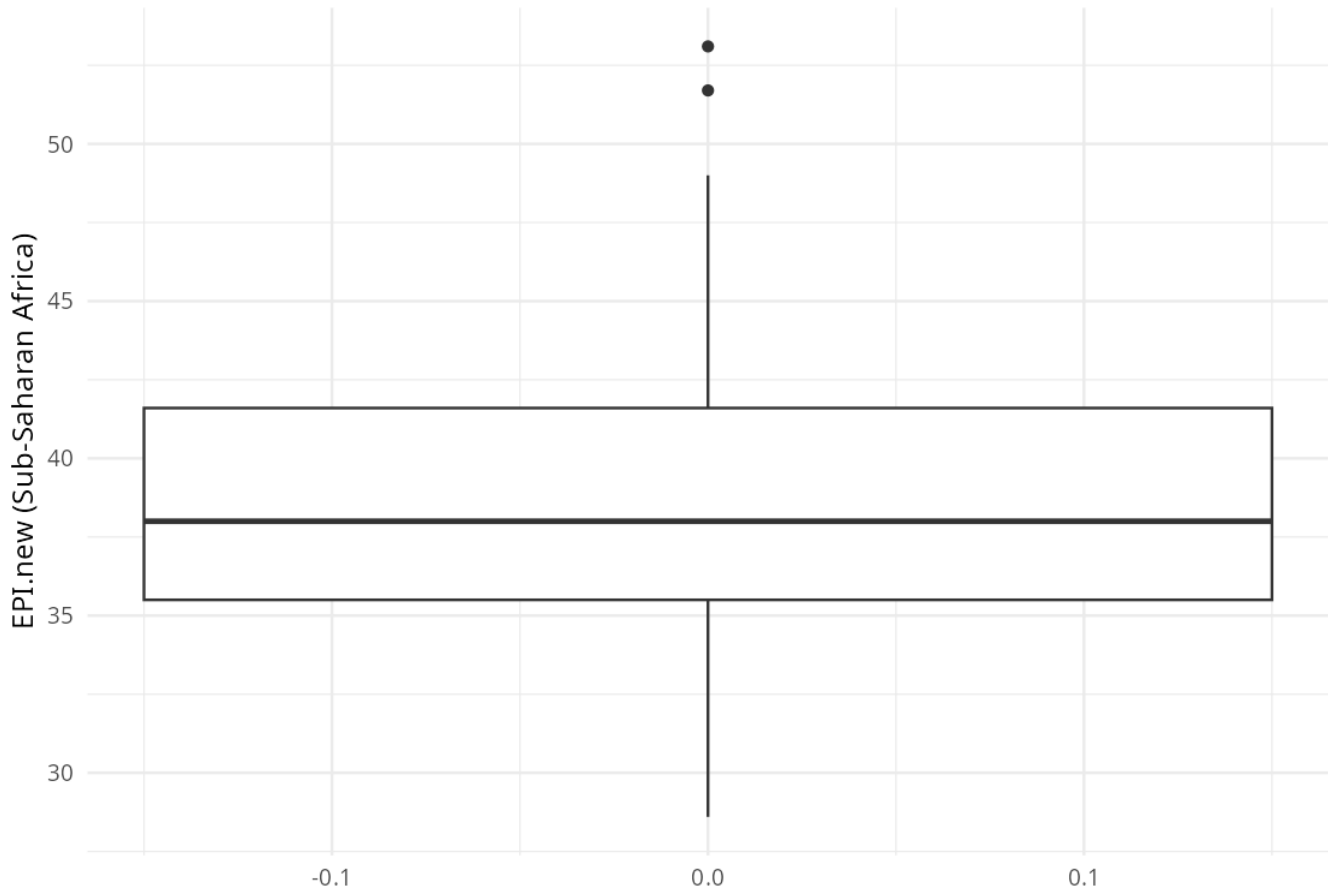
dataset and choices

- file: epi_results_2024_pop_gdp_v2.csv
- region column: region
- response var: EPI.new
- regions: Sub-Saharan Africa vs Latin America & Caribbean

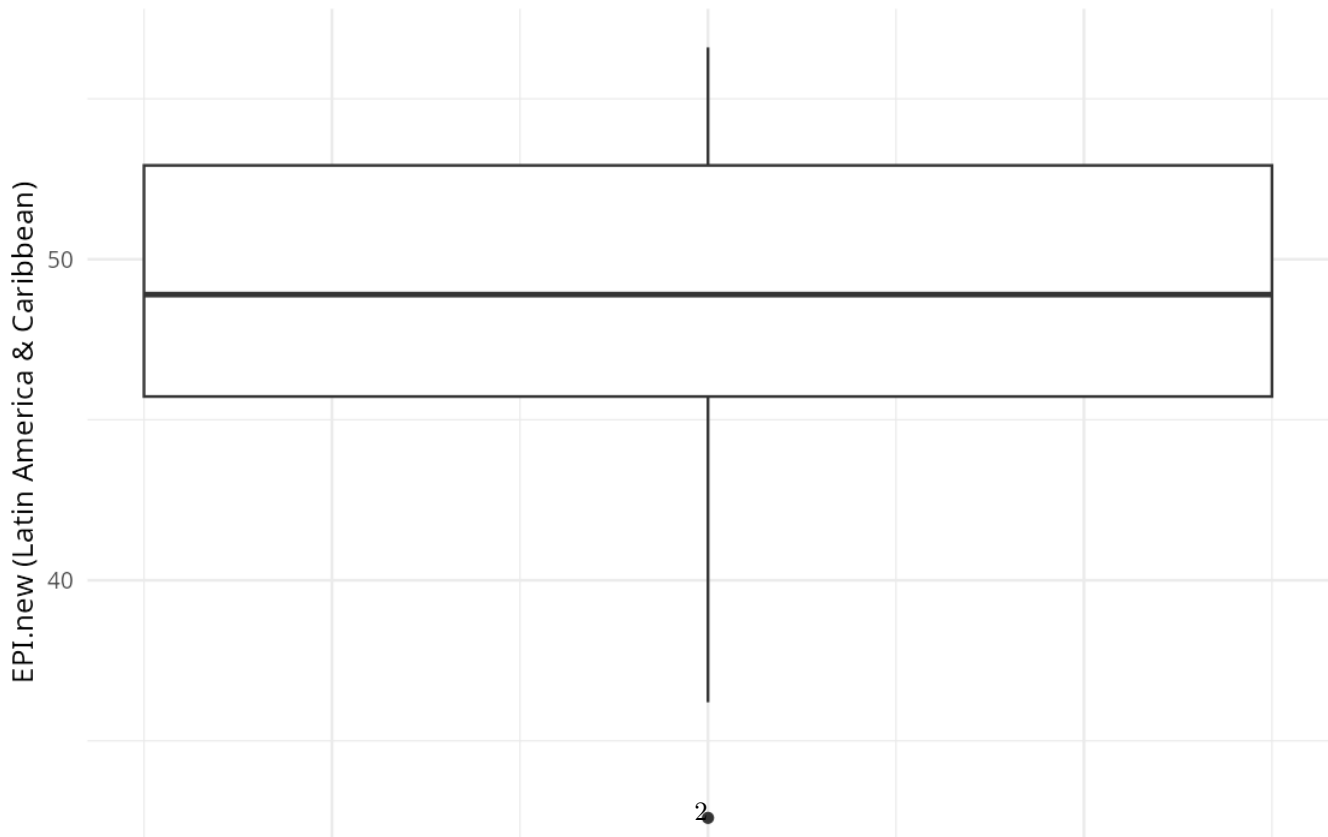
1) variable distributions

1.1 boxplots and histograms (with density!)

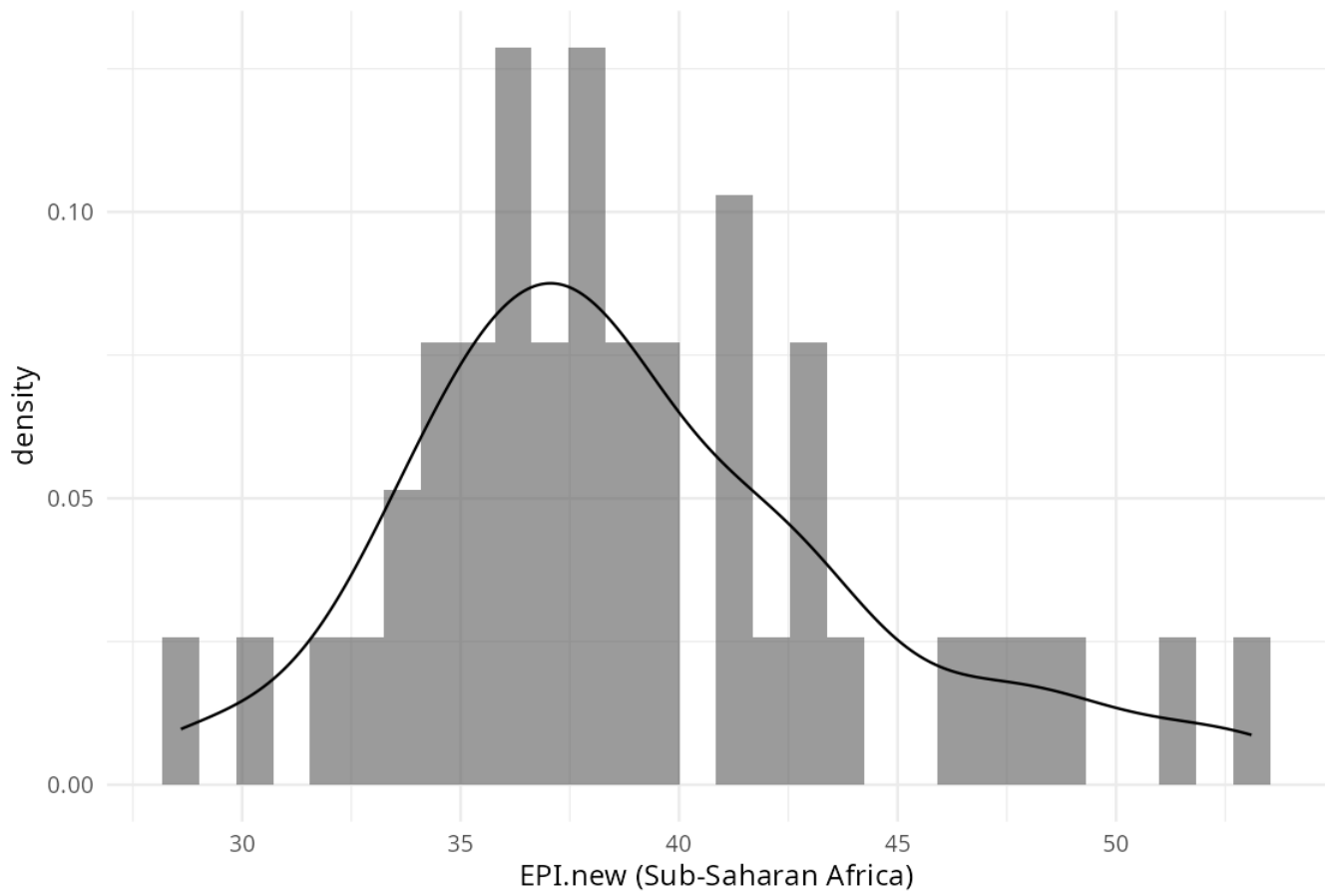
boxplot: EPI.new (Sub-Saharan Africa)



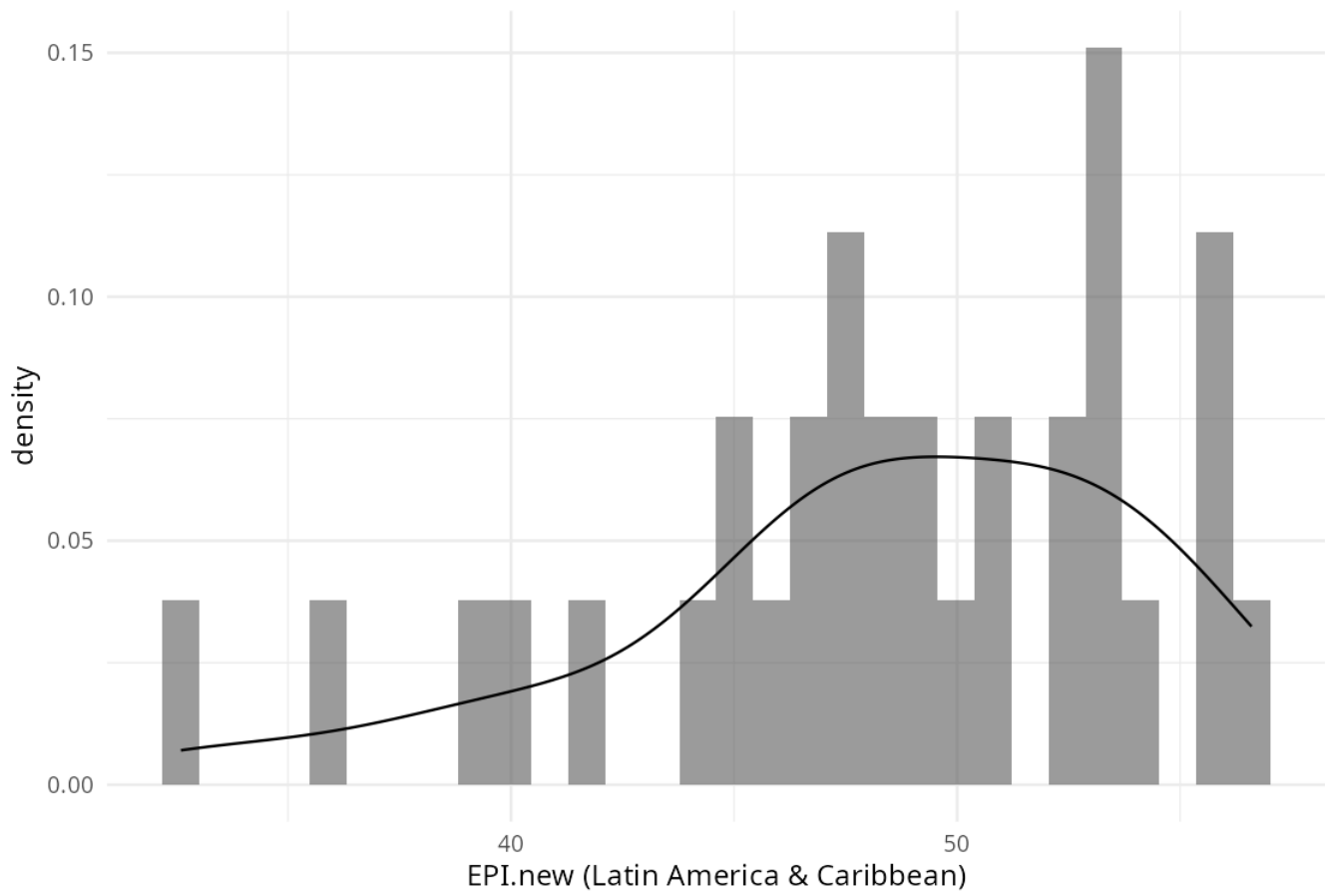
boxplot: EPI.new (Latin America & Caribbean)



histogram + density: EPI.new (Sub-Saharan Africa)

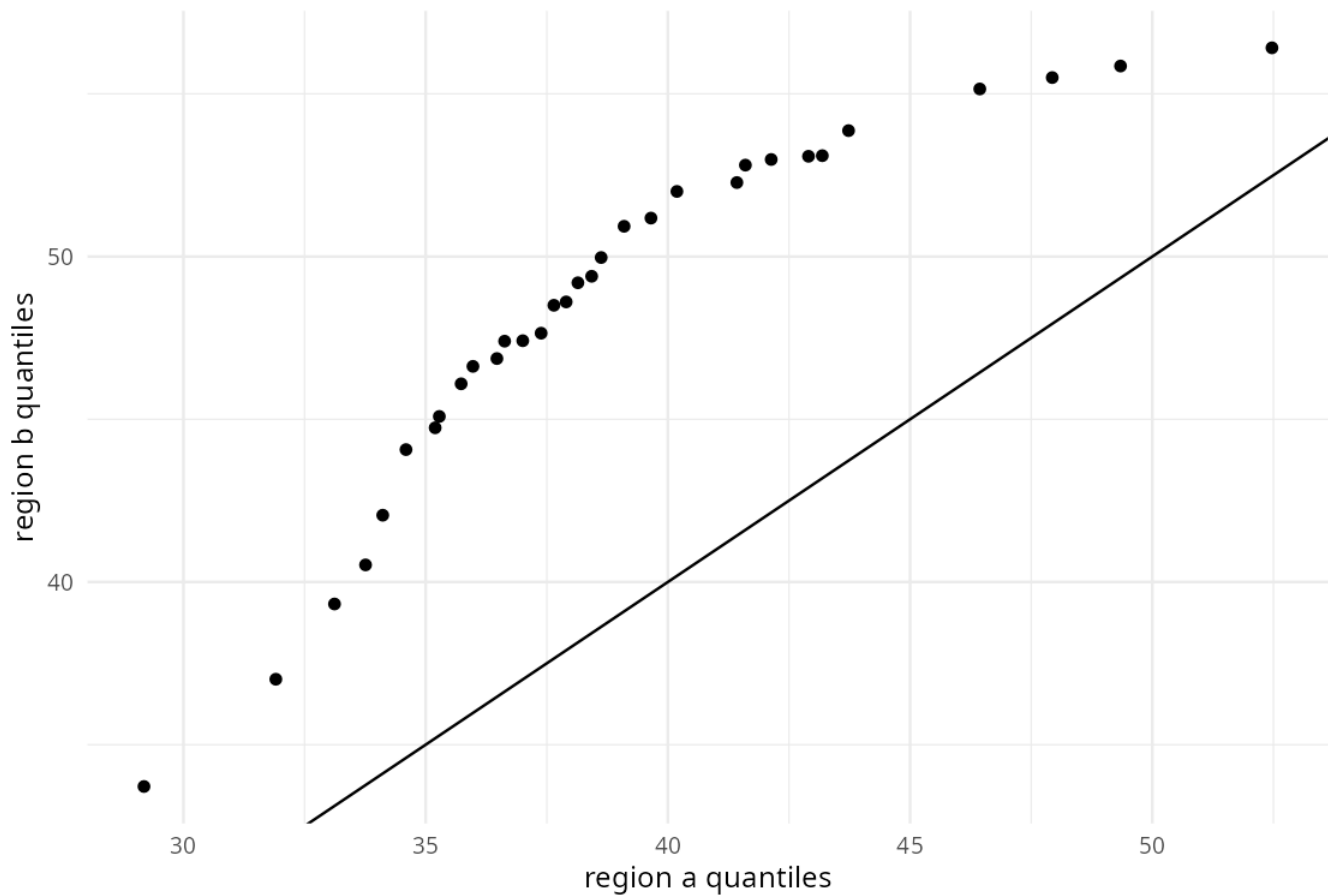


histogram + density: EPI.new (Latin America & Caribbean)



1.2 qq plot (two-sample)

qq plot: Sub-Saharan Africa vs Latin America & Caribbean



2) linear models

full: $\text{EPI.new} \sim \text{gdp}$

full: $\text{EPI.new} \sim \text{gdp} + \text{population}$

2.2 same models on one region (comparison)

on region Sub-Saharan Africa, the better model is **region Sub-Saharan Africa: $\text{EPI.new} \sim \text{gdp} + \text{population}$** ($r^2=0.361$, $\text{aic}=265.4$, $\text{bic}=272.7$).

3) classification (knn, label = region)

model A

		confusion matrix
true	Sub-Saharan Africa	0
	Southern Asia	0
	Latin America & Caribbean	1
	Greater Middle East	0
	Global West	1
	Former Soviet States	1
	Eastern Europe	0
	Asia-Pacific	2
	Asia-Pacific	

- k: 5 | accuracy: 0.5581 | test n: 43 variables: c("AGR.new", "AIR.new", "APO.new")

model B

true	confusion matrix	
	Sub-Saharan Africa	2
	Southern Asia	0
	Latin America & Caribbean	1
	Greater Middle East	0
	Global West	1
	Former Soviet States	0
	Eastern Europe	1
	Asia-Pacific	3
	Asia-Pacific	

- k: 5 | accuracy: 0.5116 | test n: 43 variables: c("BCA.new", "BDH.new", "CBP.new")