

Module 3

Exploratory Analysis

Shmueli et al., Chapters 1–4 (parts)

Frequencies

Correlations

Categories

Extended EDA

Roumen Vesselinov, PhD **Roumen@jhu.edu**

Explore the data · build the model · ask the AI · verify every result

Outcomes and predictors

Var	Description	Var	Description
Y1	GDP per capita, constant 2015 US\$	Y2	Armed conflict (yes/no)
X1	Annual GDP growth rate (%)	X11	Total population
X2	Inflation rate (CPI, annual %)	X12	Population aged 65+ (% of total)
X3	Trade (% of GDP)	X13	Urban population (% of total)
X4	FDI net inflows (% of GDP)	X14	Top 10% income share (inequality)
X5	Fraser Economic Freedom Index	X15	Government effectiveness (WGI)
X6	Oil rents (% of GDP): resource channel	X16	Rule of law (WGI)
X7	Human Development Index	X17	Political stability & absence of violence (WGI)
X8	Life expectancy at birth (years)	X18	V-Dem electoral democracy index
X9	Internet users (% of population)	X19	Historical region (10 categories)
X10	Government expenditure on education (% of GDP)	X20	Year (numeric, for time-trend modeling)

X19: historical region

Value	Region	Value	Region
1	Eastern Europe & post-Soviet Union	6	East Asia
2	Latin America	7	South-East Asia
3	North Africa & Middle East	8	South Asia
4	Sub-Saharan Africa	9	The Pacific
5	Western Europe & North America (incl. Australia, NZ)	10	The Caribbean

X19 is categorical: in models it becomes dummy variables (Slide 8).

Frequencies and summary statistics

Y2	Count	Percent
0	1869	87.95
1	256	12.05

Frequencies of Y2 (armed conflict)

```
> fancy_summary
Variable  Min      Q1    Median      Mean      Q3      Max
1  y1    333.36  1561.68  5410.31  14241.68  18089.59  110873.19
2  y2      0.00      0.00      0.00      0.12      0.00      1.00
3  x1   -36.39    1.74      3.88      3.62      5.88      34.47
4  x2    -5.36    1.47      3.22      4.79      6.15     168.62
5  x3    19.56    53.90     74.54     86.12    103.08     437.33
```

Summary statistics

- Y2: 256 conflict years (12.05%) vs. 1,869 without (87.95%): an imbalanced outcome.
- Summary statistics: check min, max, mean vs. median for every variable.

Correlations

	Y1	Y2	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Y1												
Y2	-0.16											
X1	-0.13	0.02										
X2	-0.23	0.09	0.02									
X3	0.37	-0.21	0.07	-0.10								
X4	0.08	-0.05	0.07	-0.04	0.22							
X5	0.62	-0.22	-0.13	-0.38	0.34	0.11						
X6	0.01	0.07	0.12	0.08	-0.03	-0.04	-0.34					
X7	0.70	-0.16	-0.18	-0.17	0.38	0.10	0.74	0.00				
X8	0.66	-0.12	-0.17	-0.19	0.33	0.10	0.72	-0.04	0.93			
X9	0.67	-0.16	-0.28	-0.24	0.33	0.07	0.65	-0.10	0.81	0.76		
X10	0.34	-0.12	-0.17	-0.09	0.13	0.05	0.29	-0.06	0.41	0.33	0.36	
X11	-0.08	0.36	0.04	0.05	-0.20	-0.05	-0.08	-0.03	-0.07	-0.03	-0.09	-0.08

Correlation matrix

Correlation with Y1

X15	0.82
X16	0.81
X19_5	0.77
X7	0.70
X9	0.67
X8	0.66
X17	0.62
X5	0.62
X13	0.60

Correlation with Y1

- Strongest correlates of Y1: X15 (0.82), X16 (0.81), X19₅ region dummy (0.77), X7 (0.70).
- Predictors correlated with each other signal redundancy.

Reducing categories

- A categorical variable with m categories is typically transformed into m or $m - 1$ dummy variables (handled automatically by most R modeling functions).
- Each dummy variable takes the values 0 or 1: 0 = "no" for the category, 1 = "yes".
- Problem: you can end up with too many variables.
- Solution: reduce by combining categories that are close to each other.
- Use pivot tables to assess the outcome variable's sensitivity to the dummies.

Exception

CART, RF, and NN can handle categorical variables without dummies, but they must be declared as categorical (factors).

A big dataset does not answer questions. It hides them better.

2,782

country-year observations

22

variables, no missing values

149

independent units, once you look

- Everything you want to know about specifying a model (which variables, what functional form, what to control for, your real sample size) is visible in the data before you fit anything. The only cost is looking.
- By the end, the model specification changes four times without a single regression.

Four questions, four sets of pictures

1 Each variable alone

Shape, spread, skew, impossible values, ceilings and floors.

2 Variables together

Correlation structure, nonlinearity, redundancy among predictors.

3 Structure in the rows

Groups, aggregation effects, repeated measures on the same unit.

4 What EDA (Exploratory data Analysis) decided

Transformations, controls, standard errors, effective sample size.

Three things the table already told us

Skew of 23

X2 inflation: mean 5.71, median 3.29, max 557.20. One or two hyperinflation episodes dominate; a linear model gives them enormous leverage.

A minimum of -296%

X4 FDI ranges from -296% to +452% of GDP. Not an error: financial-center economies book huge pass-through capital. But it is not measuring what you think.

SD 3× the mean

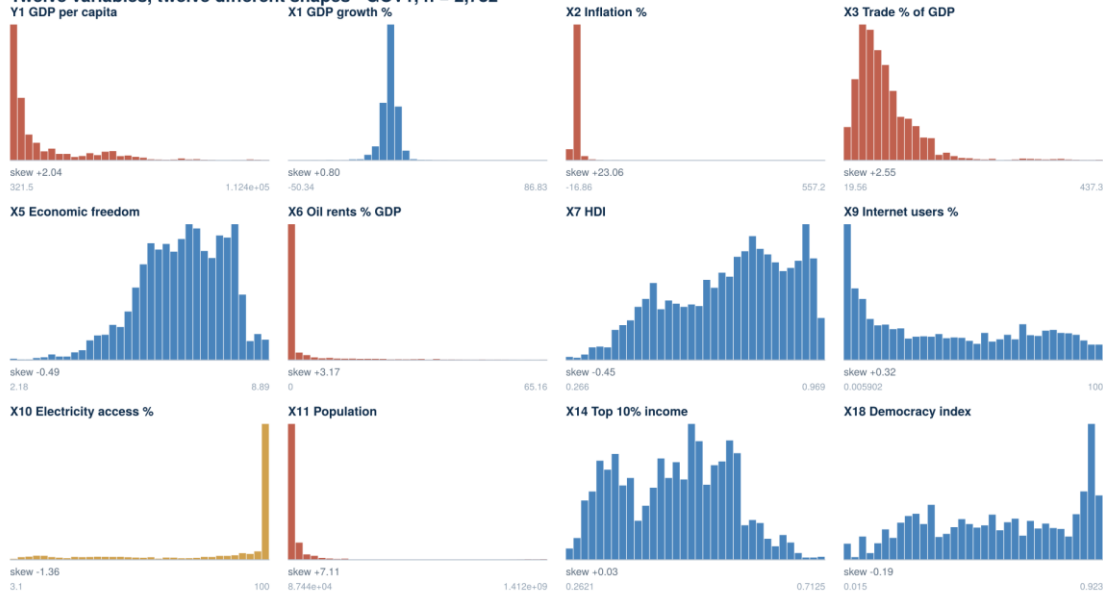
X11 population: mean 47.6m, SD 160.8m, max 1.41bn. Four orders of magnitude; raw, it contributes little except two outliers.

The rules

Mean $\gg$ median $\rightarrow$ right skew. SD $\geq$ mean $\rightarrow$ orders of magnitude. Min/max at 0, 100, 1 $\rightarrow$ floor or ceiling. Median at the max (X10: 99 vs. 100) $\rightarrow$ piled at the ceiling.

Twelve variables, twelve different shapes

Twelve variables, twelve different shapes - GOV1, n = 2,782



- Same data set, same n, same care in construction.
- Four need a transformation before entering a linear model.
- Two are effectively binary in disguise.

What each shape is telling you to do

Shape	Variables	Action
Long right tail	Y1, X6, X11, X2, X3	Consider a log or rank transform; check for zeros first (X6 has minimum 0)
Hard ceiling	X10 electricity, X9 internet	Relationship must flatten near 100%; a linear term misfits both ends
Spike at one value	X2 inflation, X6 oil rents	Really "a few countries differ": a binary indicator may carry more signal
Roughly symmetric	X5, X7, X14, X15, X16	Usable as is in a linear model
Nearly uniform	X9 internet, X18 democracy	Plenty of variation across the range: good for estimation

Extreme is not the same as wrong

What a rule would delete

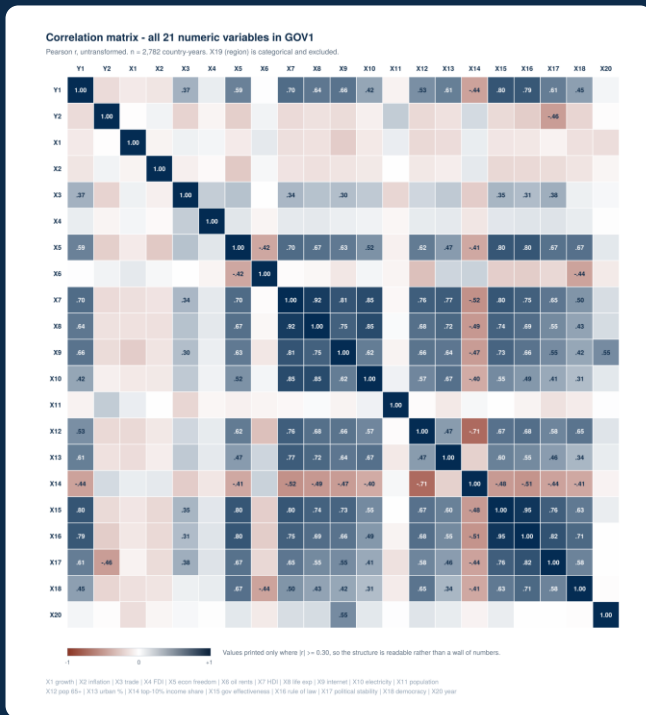
- Tukey's $1.5 \times \text{IQR}$ on Y1 flags 314 country-years (11.3%).
- Every one is a rich country.
- The rule assumes symmetry; on a skewed variable it flags a whole legitimate subpopulation.

What the data say instead

- Cyprus 2020: FDI of -296% of GDP. Real: special-purpose entities hold stocks $\sim 11 \times \text{GDP}$ in a \$25bn economy.
- It does not measure the construct the model is about.
- Statistical reason ("outside $1.5 \times \text{IQR}$ "): weak. Substantive reason: strong.

State the rule, report the count, keep removed rows in a file, and run the analysis both ways.

The whole correlation structure on one slide



- Predictors correlate more with each other than with Y1: X15–X16 0.95, X7–X8 0.92, X7–X10 0.85.
- Two clusters: governance (X15–X17) and development (X7–X10, X13).
- Nearly inert: X4, X6, X11, X20 (irrelevant, or nonlinear).
- Y2 is different: –0.46 with political stability, little else.

What $r = 0.95$ does to a regression

The effect

- Coefficients stay unbiased; standard errors inflate by $1/\sqrt{1 - R^2_j}$.
- At $r = 0.95$, $VIF \approx 10$: SEs about three times larger.
- Symptom: joint F-test significant, no single coefficient is; or signs flip.
- Not an assumption violation: you lose the ability to attribute the effect.

Pairs to watch and responses

- X15 & X16 0.95; X7 & X8 0.92; X7 & X10 0.85; X7 & X9 0.81; X16 & X17 0.82.
- Drop one of each pair on substantive grounds.
- Combine into an index or principal component.
- Keep both, report VIFs, interpret jointly.

Why the gap appears, and what to do

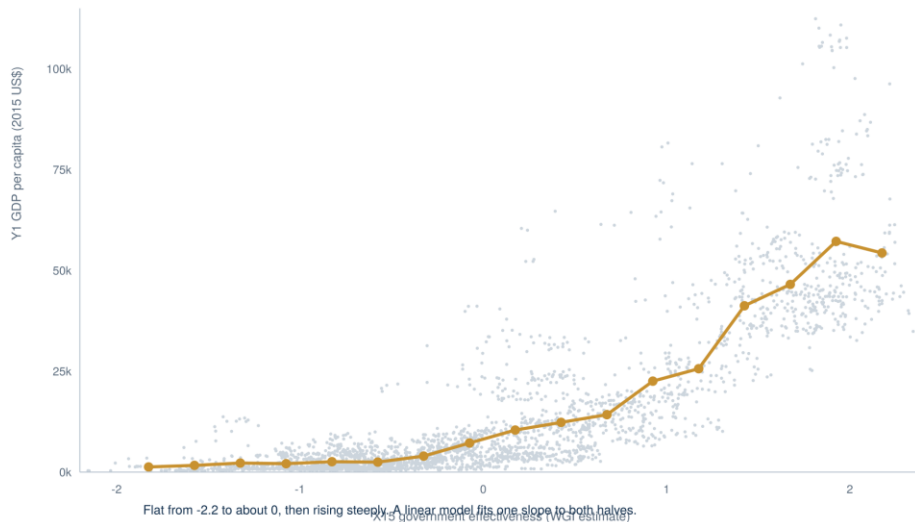
Variable	Pearson	Spearman	Gap	What is going on
X10 electricity	0.42	0.86	+0.44	Ceiling at 100%: ordering nearly perfect, line fits badly
X7 HDI	0.70	0.95	+0.25	HDI bounded in $[0, 1]$; GDP is not: close to exponential
X8 life expectancy	0.64	0.89	+0.25	Large income gains buy small gains at the top
X2 inflation	-0.12	-0.37	-0.25	Hyperinflation episodes; ranks are immune
X15 gov. effectiveness	0.80	0.83	+0.03	Close to linear: the strongest raw correlate

A large gap is a reason to change the functional form (log, quadratic, spline), not to switch to Spearman.

The strongest linear correlate is not linear

Y1 vs X15 government effectiveness - $r = 0.80$ hides a curve

Grey points: 2,782 country-years. Orange: mean of Y1 within each 0.25-wide band of X15.

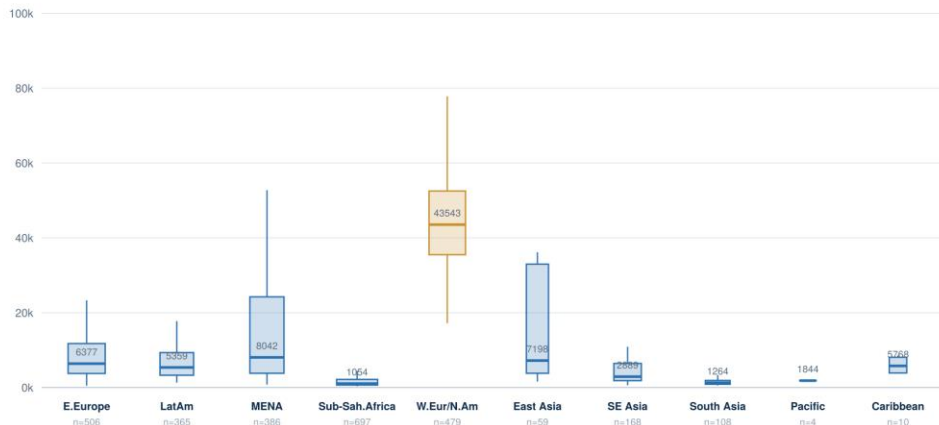


- $r = 0.80$, the highest in the data set.
- Binned means: flat from -2.2 to about 0 , then a steep climb.
- A single slope is fitted to both halves.

One categorical variable explains most of the spread

Y1 by historical region (X19) - one categorical variable explains most of the spread

Box = IQR, line = median, whiskers = 1.5 x IQR. Width of each box is proportional to the square root of n.



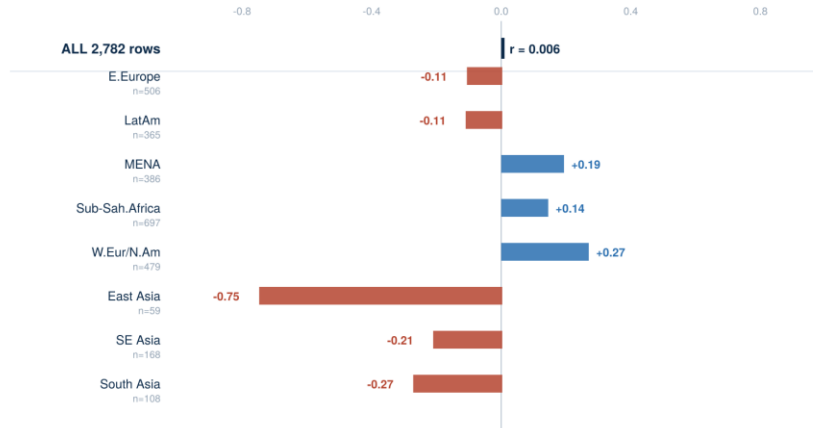
Median GDP per capita ranges from \$1,054 (Sub-Saharan Africa) to \$43,543 (W. Europe & N. America) - a factor of 41.

- Median Y1: \$1,054 (Sub-Saharan Africa) to \$43,543 (W. Europe & N. America), a factor of 41.
- Invisible in a correlation matrix (categorical), yet omitting it biases other predictors.
- Pacific n = 4, Caribbean n = 10: collapse categories with fewer than ~30 rows.

A correlation of 0.01 that is 0.01 nowhere

Oil rents and income: $r = 0.01$ overall, but never 0.01 anywhere

Correlation between X6 (oil rents, % of GDP) and Y1, computed on the pooled sample and then separately within each region.



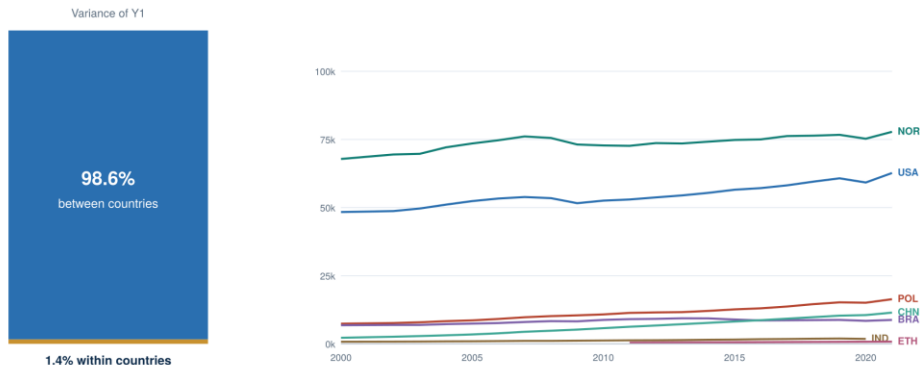
The pooled correlation is an average over groups whose relationships point in opposite directions. East Asia is -0.75; W. Europe & N. America is +0.27.

- Pooled: oil rents and income $r = 0.006$.
- Within regions: -0.75 (East Asia) to $+0.27$ (W. Europe & N. America).
- Gulf states oil-rich and rich; Sub-Saharan exporters oil-rich and poor: pooled, they cancel.
- Always check relationships within groups.

2,782 rows, but only 149 independent units

2,782 rows, but only 149 independent units

Left: where the variance in Y1 actually lives. Right: each line is one country followed from 2000 to 2021.



Countries barely move relative to the gaps between them. Treating 2,782 rows as independent overstates the information in the sample by roughly an order of magnitude.

- 98.6% of the variance in Y1 is between countries; 1.4% within.
- Standard errors computed as if $n = 2,782$: everything looks significant. Cluster by country.
- Random row splits put Norway 2003 in training and 2004 in test: split by country.
- Country fixed effects use only the 1.4%: a different question.

Four things break when the outcome is 0/1

1

Correlation is capped

With 12.5% positives the ceiling is about 0.59; a weak r may be strong. Use AUC or Cohen's d .

2

Mean = prevalence

0.125 is a rate. Histograms, boxplots, violins of Y2 are meaningless.

3

Not a straight line

Probabilities live in $[0, 1]$: an S-curve, the logit link.

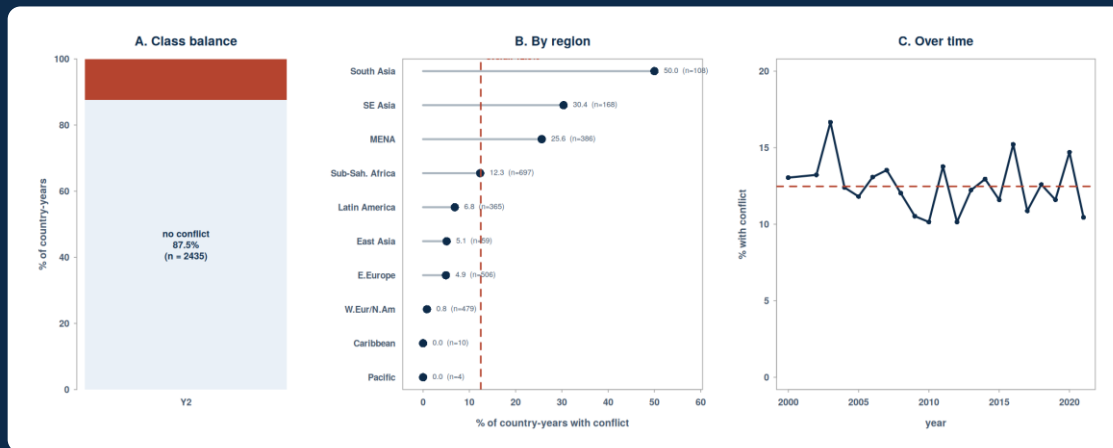
4

Accuracy fails

"No conflict" for every row is 87.5% accurate and useless.

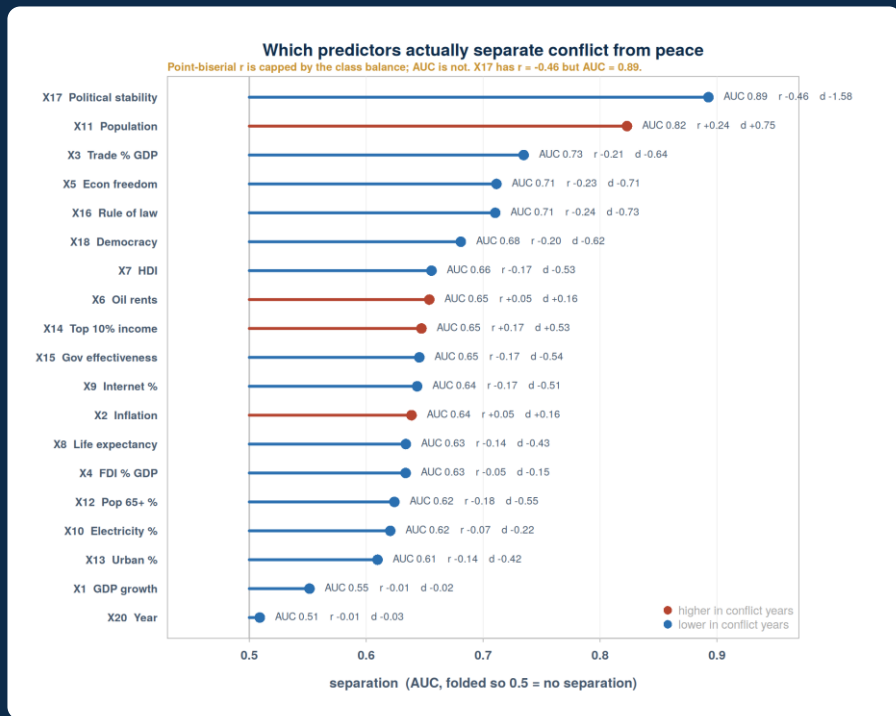
Everything from Sections 1–3 still applies: groups, aggregation, independence.

Start with the base rate, by group and over time



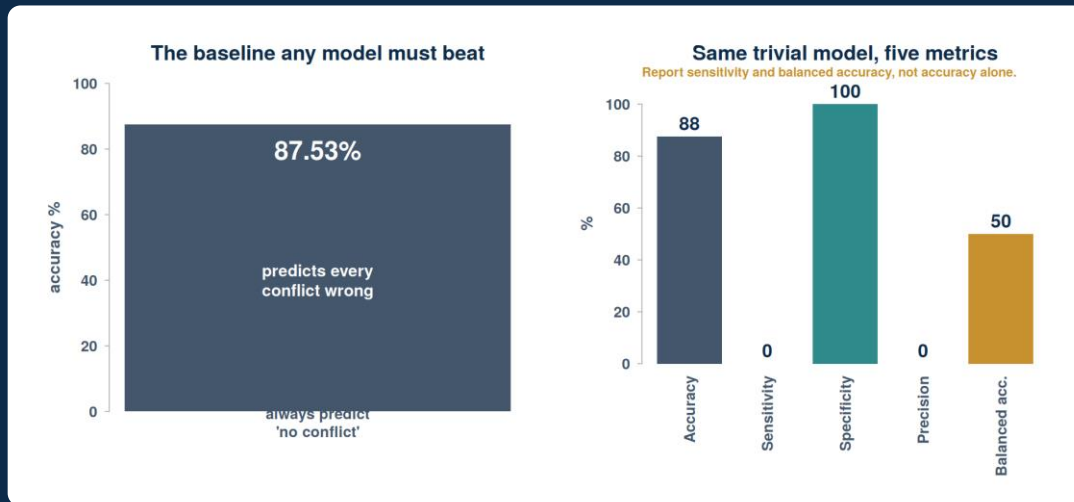
- 347 of 2,782 country-years (12.5%).
- 0% in the Pacific and Caribbean; 50.0% South Asia; 30.4% Southeast Asia; flat over time.
- Persistent: 78.2% of conflict years follow a conflict year (3.0% after peace).
- 100 countries never have conflict; 2 always do.

Which predictors separate conflict from peace?



- Predictors ranked by AUC (0.5 = no separation).
- Political stability: AUC 0.89 alone, but r only -0.46 (r is capped at ~ 0.59 with 12.5% positives).
- Screen binary outcomes with AUC or Cohen's d , not Pearson r .

The accuracy trap



- "No conflict" for every row: 87.53% accurate, 100% specific.
- Catches none of the 347 conflicts.
- Sensitivity 0, precision 0, balanced accuracy 50%.

The specification, before and after looking

Before

- $Y1 = \beta_0 + \beta_1 X1 + \dots + \beta_{18} X18 + \varepsilon$
- OLS on all 18 predictors, raw scale.
- $n = 2,782$, classical standard errors.
- Region ignored; random 80/20 split; outliers removed by $1.5 \times \text{IQR}$.

After

- $\log(Y1) = \beta_0 + f(X15) + \beta X + \text{region} + \varepsilon$
- Outcome logged: skew 2.04 $\rightarrow$ roughly symmetric.
- X15 or X16, not both; X7 or X8, not both; region dummies, tiny categories collapsed.
- SEs clustered by country; split by country; exclusions justified substantively.

Six specification changes. Zero regressions run.

Six ways EDA goes wrong

1

Only the outcome

Most specification problems are in the predictors.

2

Trusting the matrix

Linear, pairwise, drops categoricals.

3

EDA as testing

Testing a hypothesis on the data that suggested it is not a test.

4

Tuning the cleaning

Decide rules before seeing what they remove.

Also: stopping at summary statistics (Anscombe), and not writing anything down.

Your turn: do this to Y2

- Tabulate Y2 by region. Which regions carry the positive cases? Does any region have too few?
- Compare each X for $Y2 = 1$ vs. $Y2 = 0$: which separate the classes, and by how many SDs?
- Compute the point-biserial r of Y2 with each X, then again within region. Does anything change sign?
- Count countries that ever have $Y2 = 1$ and those that switch: the effective sample for a fixed-effects model.
- What is the accuracy of always predicting 0? Write it down before fitting anything.

Hand in

Four figures and half a page on what each changed about how you would model Y2.

All from GOV1, all before fitting anything

- 1 Four variables needing transformation; two predictor pairs that must not go in together ($r = 0.95, 0.92$).
- 2 A curved relationship inside the strongest linear correlate; a categorical variable invisible to the correlation matrix.
- 3 A pooled correlation of 0.01 that is -0.75 to $+0.27$ in every subgroup; an effective sample of ~ 149 , not 2,782.
- 4 Two region categories too small to estimate; a validation strategy that would have leaked.
- 5 For Y2: a correlation screen that misses the strongest predictor, and an 87.5% accuracy baseline to beat.

MODULE 3

Exploratory Analysis

None of this required a model. All of it changed the model. · Roumen Vesselinov