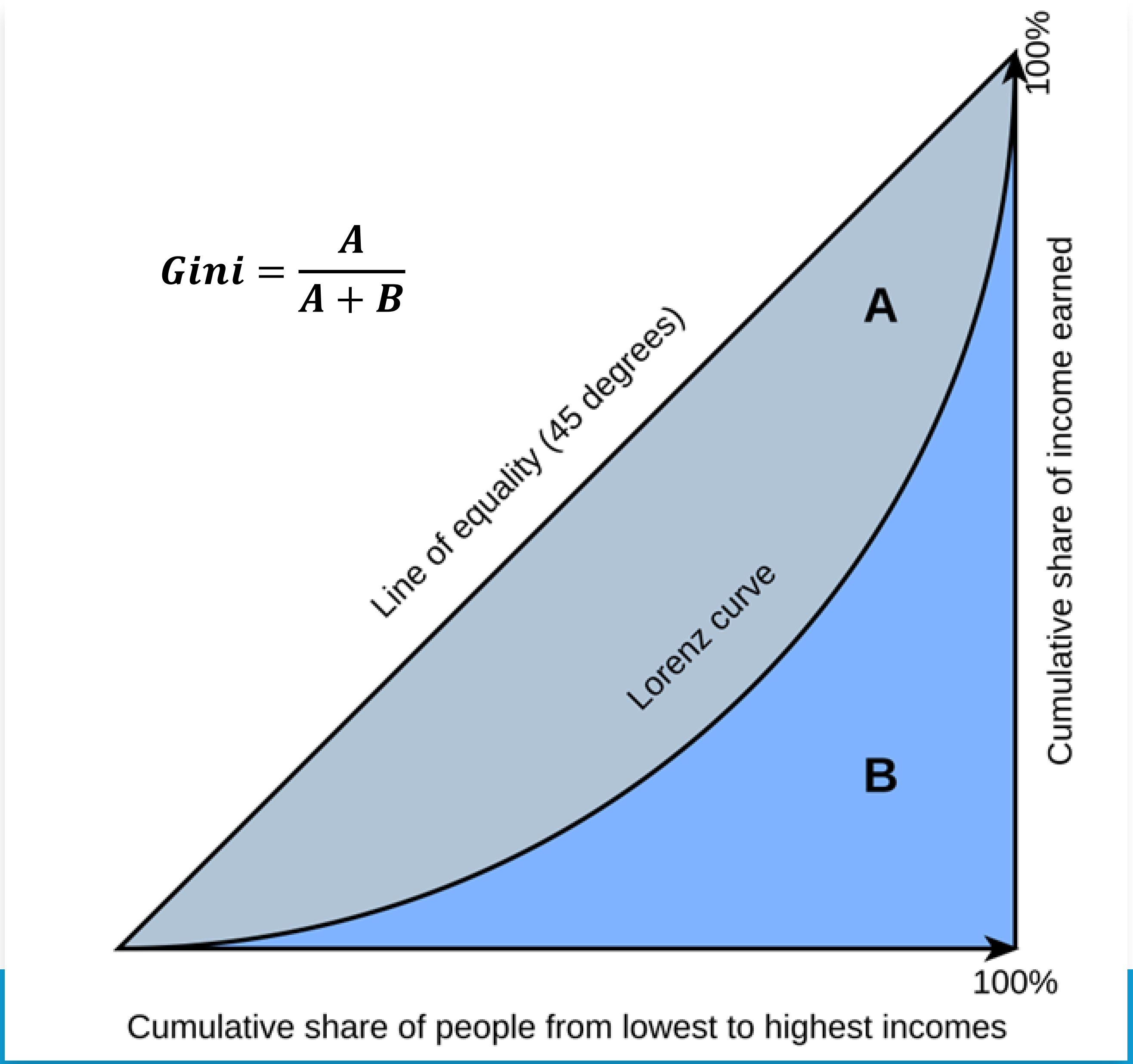
# World Bank Indicators - **Predicting The Gini Index**

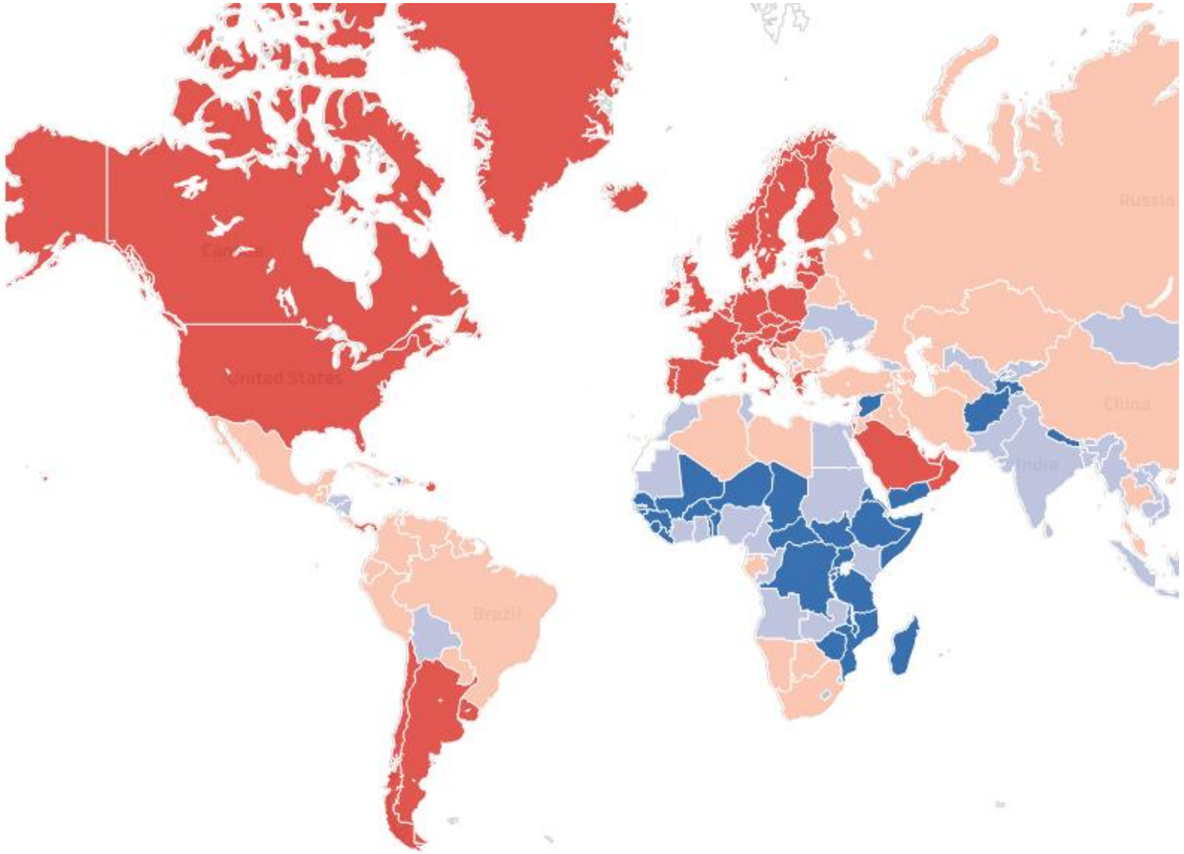

Team 5 - CSSE-415, Spring 2024-2025

Abdullah Islam, Agnay Srivastava, Parth Sundaram, Steven Johnson




# What is the Gini Index?

- It is a measure of income inequality:
  - 0 = perfect equality
  - 100 = perfect inequality
- Ratio of area between the Lorenz Curve & equality line
- Highest & Lowest
  South Africa at 63
  Slovakia at 23.2




# Our Kaggle Dataset: A Summary

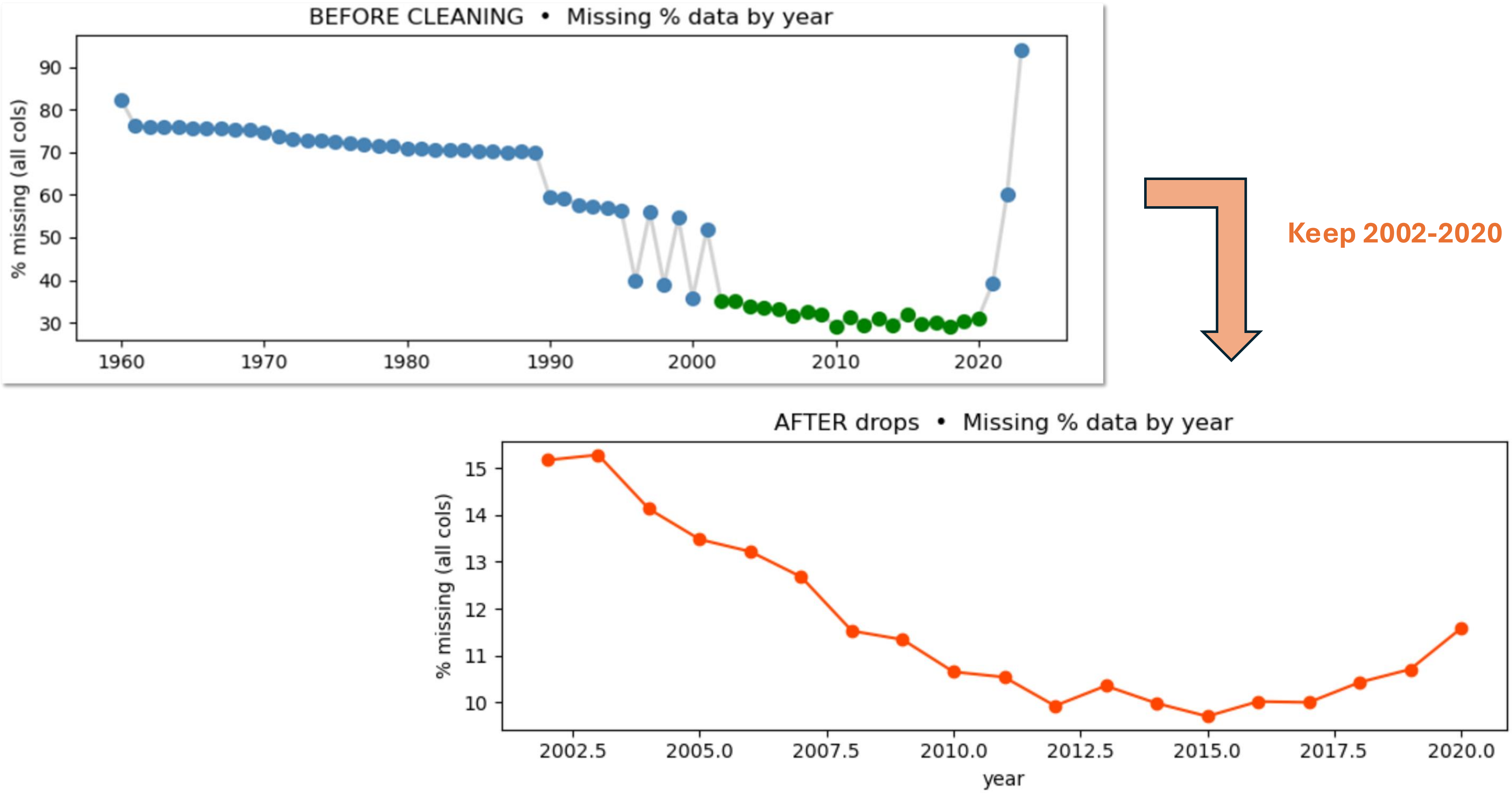
- Annual WB data since 1960
- 274 economies
- 50+ socio-economic, environmental, and institutional metrics
- Size: ~ 17,000 x 50.
- The main operations we perform are
    A. **Cleaning**
    B. **Imputation**
    C. **Time-Series Analysis**

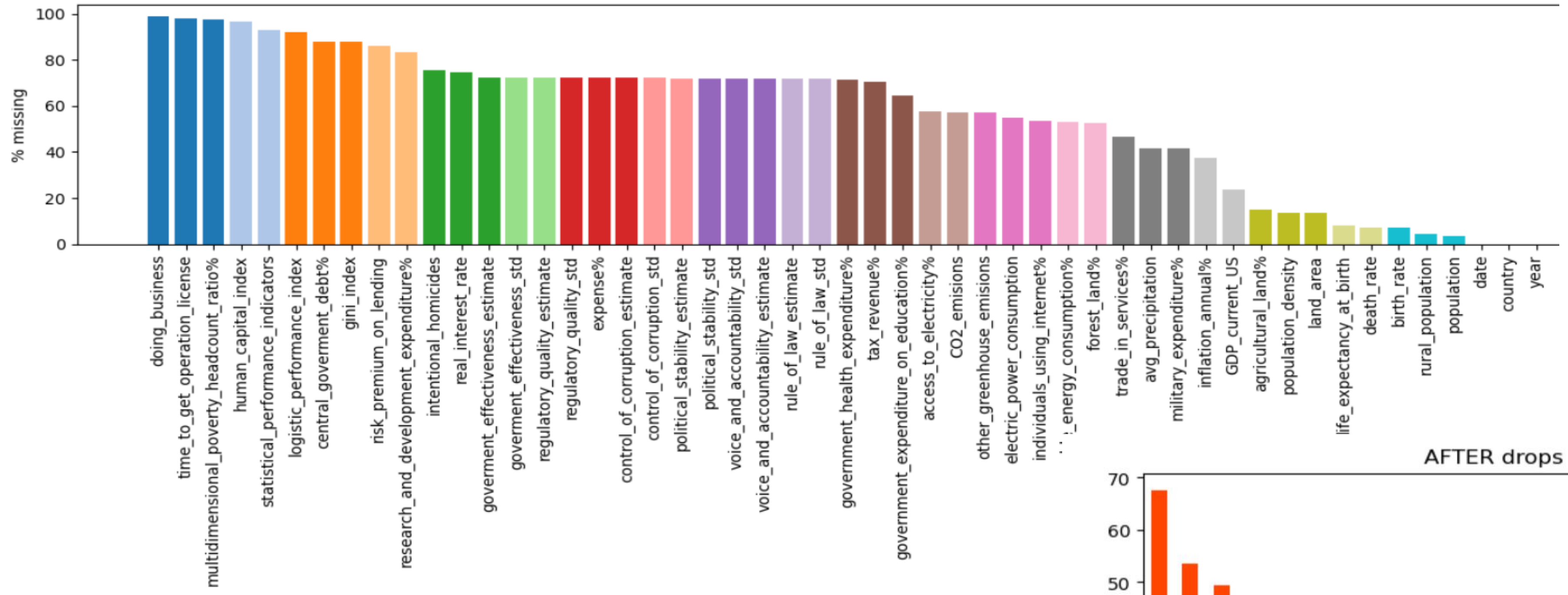
# 4 Core Data Cleaning Steps

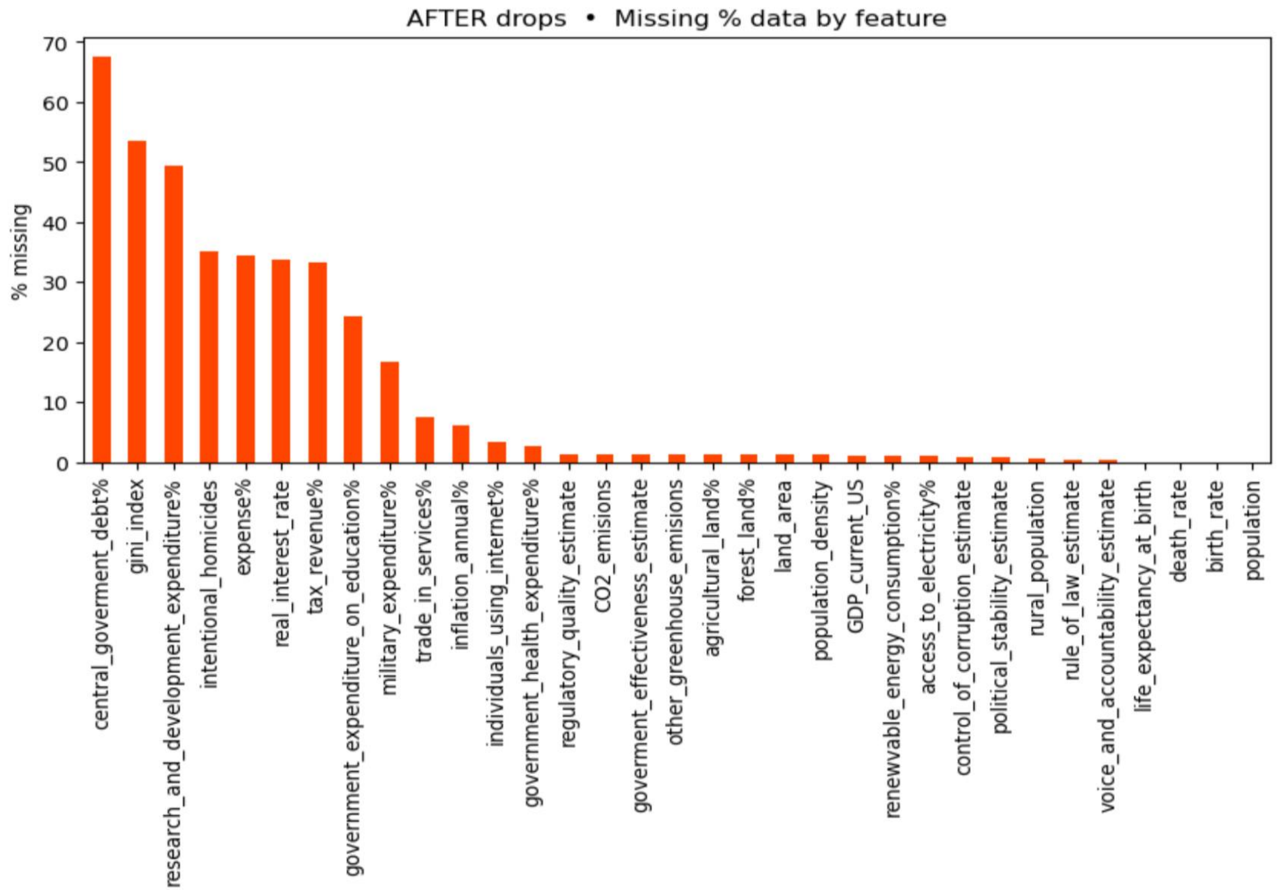
1. We filter years **2002 – 2020**
2. **Drop 15** Sparse, Irrelevant & Discontinued **Features**
3. Drop **countries ≥ 90 % empty** values
4. Finally, drop countries with **no Gini-Index**

- **Before: 274 Economies → After : 163**

# Before vs After Cleaning [1/3]:

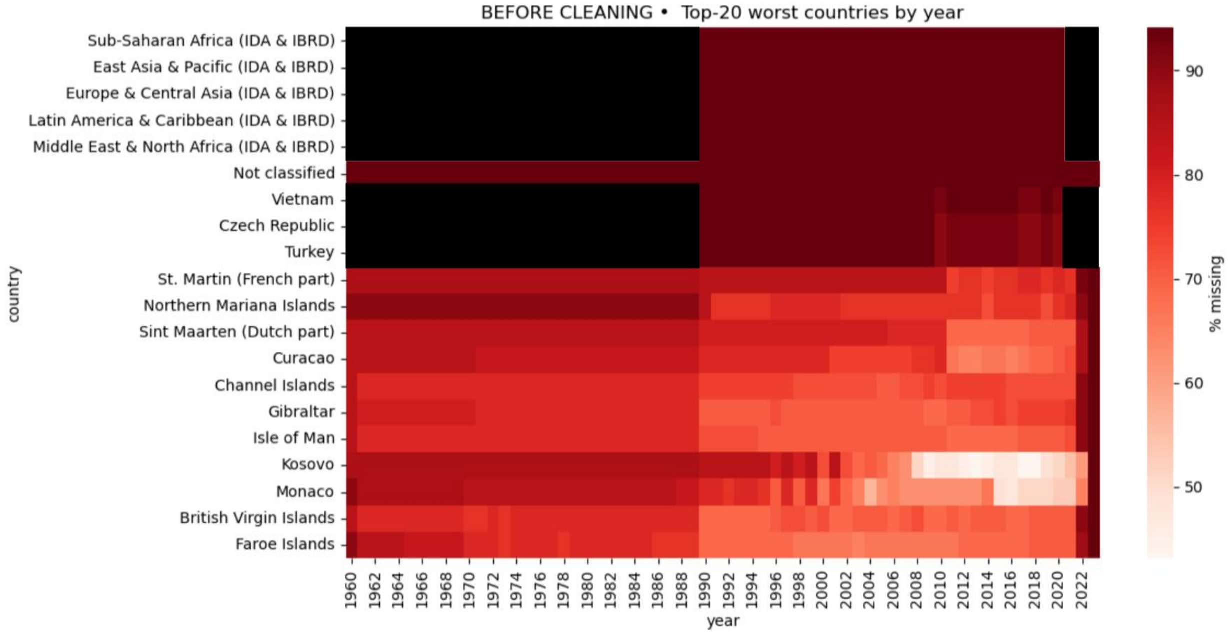


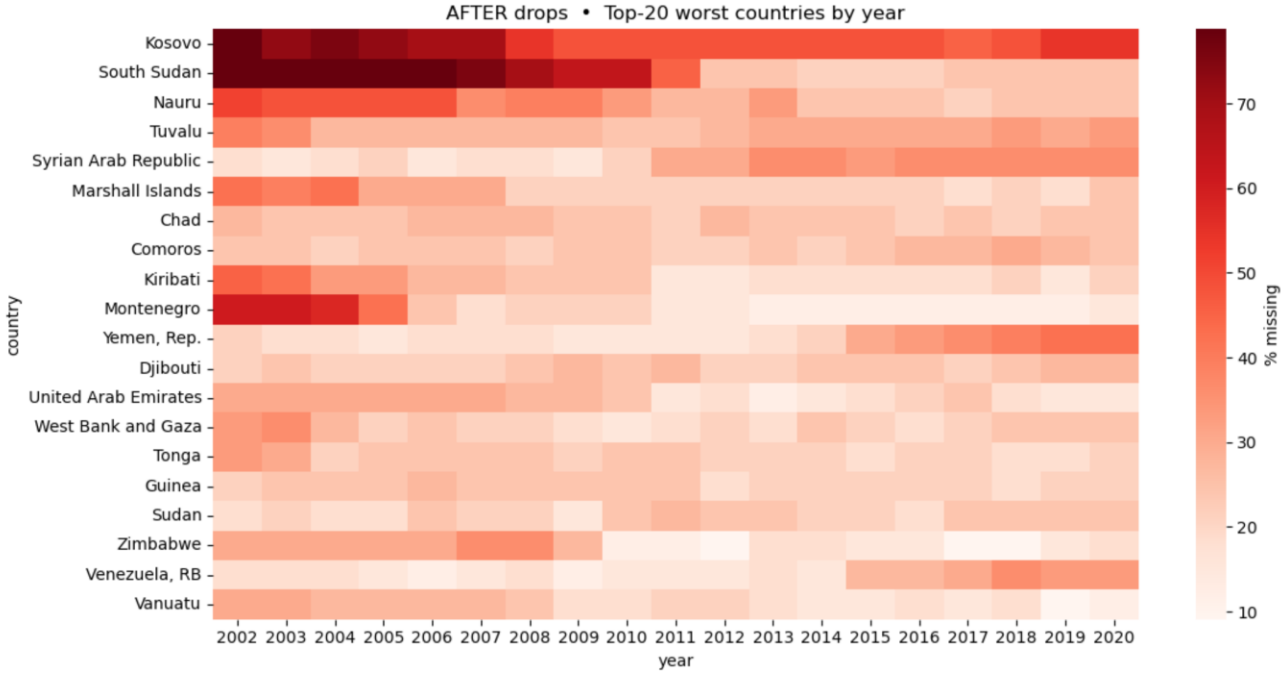



# Before vs After Cleaning [2/3]:

# Before vs After Cleaning [3/3]:
# Top 20 Countries With the Most Missing Data



# Mean Imputation Results

- **Dataset Size:** 3,097 rows × 36 columns

- KNN:
  - ⊗ K optimization
- MICE:
  - ⊗ Complex
- Miss-Forest:
  - ⊗ Slow, Poor Interpretability
- **Per-Country Mean:**
  - ✓ **Simple & Interpretable**
  - ⊗ **Doesn’t use other features**

❖ **Two-Stage Strategy**
  - **Country-Mean Fill**
  - **Global Fallback**

# Time Series Feature Engineering

- **Set 1: Year t**
  - All features from the **year before the Gini value**
  - Total of **32** features
- **Set 2: Year-t + Rolling Stats**
  - All features plus **5-year rolling mean** & **rolling STD**
  - Total of 32*3 = **96** features
- **Set 3: Five-Year Lags**
  - Which keeps all the raw indicators for **t – 1, t – 2, t – 3, t – 4.**
  - Total of 32*5 = **160** features

# Recap – The Big Picture So Far [1/2]

1. **Cleaning:** We removed features and countries with high degrees of missingness.

2. **Imputation:** Filled in what was still missing

3. **Time-Series:** Used data from past years for new feature columns and generated 3 datasets for each model to use.

❖ Our philosophy was to **pick interpretability** over complex procedures that promised accuracy.

⊗ **NO Polynomial Features, PCA**

# Recap – Dataset Numbers [2/2]

- Using each of the 3 data sets, we create **3 train/test sets**.
  - **80/20 Split by country**.
- **Y Train / Test stay the same**, as only features change.

| Feature Set | | X_train (rows × cols) | X_test (rows × cols) |
|---|---|---|---|
| **1:** | **1 Year lag** | 813 × 32 | 216 × 32 |
| **2:** | **Year 1 + Rolling stats** | 813 × 96 | 216 × 96 |
| **3:** | **5 Year lags** | 813 × 160 | 216 × 160 |

| Model | Data Set 1 (year 1) | Data Set 2 (year 1 + rolling mean + std) | Data Set 3 (year 1-5) | Avg $R^2$ % |
|---|---|---|---|---|
| | $R^2$ % | $R^2$ % | $R^2$ % | |
| Simple Bias | -2.2 | -2.2 | -2.2 | -2.2 |
| Linear Regression | 39.9 | 53.4 | 43.0 | 45.1 |
| Linear + Forward Feature Selection | 40.0 | 53.1 | 44.2 | 45.8 |
| Ridge Regression | 34.7 | 42.5 | 38.4 | 38.5 |
| Lasso Regression | 35.6 | 51.1 | 33.3 | 40.0 |
| Random Forest | 54.3 | 49.3 | 49.7 | 51.1 |
| Gradient Boosted Trees | 50.6 | 53.6 | 50.4 | 51.5 |
| K-Nearest Neighbors | 45.4 | 43.9 | 45.4 | 44.6 |
| Avg $R^2$ % | 42.9 | 49.6 | 43.5 | |

# Interpretation [1/3]: Best Dataset Type:

- Set 1 – **Year-t** only ($R^2 \approx 43$ %)
  - Inadequate temporal analysis -> **high bias**
- Set 3 – **Five-Year** Lags ($R^2 \approx 44$ %)
  - **Collinearity** and imputation noise -> **high variance**
- Set 2 – Year-t + **Rolling Stats** ($R^2 \approx 50$ %)
  - Captures temporal trends, avoiding noise

| Model | Set 1 | Set 2 | Set 3 | Avg |
|---|---|---|---|---|
| | $R^2$ % | $R^2$ % | $R^2$ % | $R^2$ |
| **SBR** | -2.2 | -2.2 | -2.2 | -2.2 |
| **LR** | 39.9 | 53.4 | 43.0 | 45.1 |
| **LRFS** | 40.0 | 53.1 | 44.2 | 45.8 |
| **Ridge** | 34.7 | 42.5 | 38.4 | 38.5 |
| **Lasso** | 35.6 | 51.1 | 33.3 | 40.0 |
| **RF** | 54.3 | 49.3 | 49.7 | 51.1 |
| **GBT** | 50.6 | 53.6 | 50.4 | 51.5 |
| **KNN** | 45.4 | 43.9 | 45.4 | 44.6 |
| **Avg $R^2$** | 42.9 | 49.6 | 43.5 | |

# Interpretation [2/3]: Best Model Types

- **Random Forests**

- Best Single $R^2$ -> on Set 1
- Performance stayed in the high 40% to low 50% range across datasets

**Why did they outperform the other models?**

- **Non-linear** relationships
- Built-in **feature selection**
- Robust to **noise**

**Gradient Boosted Trees**

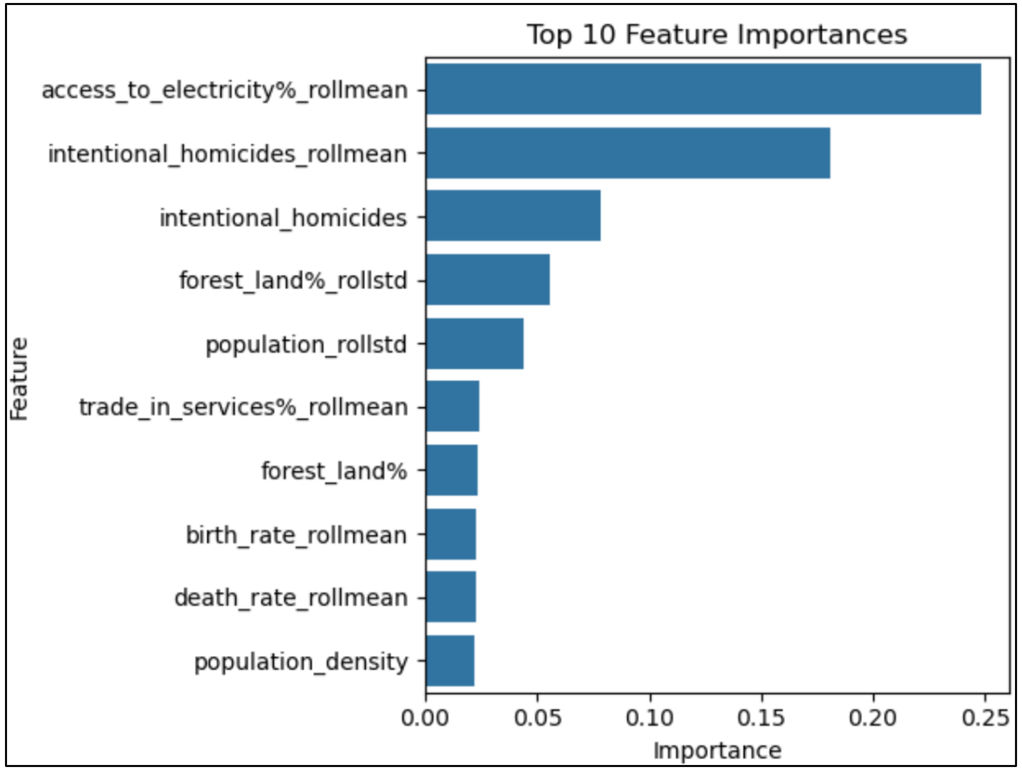
- Best $R^2$ on Set 2 (best performing dataset)
- Performance stayed at a 50-54% $R^2$ range across datasets

| Model | Set 1 $R^2$ % | Set 2 $R^2$ % | Set 3 $R^2$ % | Avg $R^2$ |
|---|---|---|---|---|
| RF | 54.3 | 49.3 | 49.7 | 51.1 |
| GBT | 50.6 | 53.6 | 50.4 | 51.5 |

**Gradient Boosted Trees** -> best model even though Random Forests had the highest $R^2$ of any model on Set 1!

# Interpretation [3/3]:
# Best Features from our Top Models

- Intentional Homicides (per 100,000 people)
  - **Interpretation:** Higher income inequality -> instability -> violence
- Forest Land (% of land area)
  - **Interpretation:** Less urbanization -> more dispersed, rural populations -> less ability to regulate those with power -> higher income inequality
- Access to Electricity (% of population)
  - **Interpretation:** electricity -> urbanization and machinery -> wider array of jobs-> lower income inequality
- Death Rate (per 1,000 people)
  - **Unexpected Insight:** Negative correlation




*Best Model – GBT on Set 2*

# Literature Review: Project Takeaways

❖ **From Li et al. (2022)[1]**

- They developed a two-stage ensemble (picking the best of many models) to predict state fragility using 100+ World Bank Development Indicators (WDI).
- We loosely adopt their concept, comparing several models.
- We also found that RFs and GBTs work the best.

❖ **From Koç & Akın (2021)[2]**

- They compared linear vs. tree-based methods for next-year Gini forecasting on OECD nations, using median imputation.
- Building on their analysis, we add more features, test more countries and imputers, and adopt their focus on next-year Gini forecasting.
- Similarly, we found that RF outperformed linear baselines.


1. X. Li, A. Vidmer, H. Liao, and K. Lu, "Data-Driven State Fragility Index Measurement Through Classification Methods," *Frontiers in Physics*, vol. 10, Art. no. 830774, Feb. 2022, https://doi.org/10.3389/fphy.2022.830774

2. T. Koç and P. Akın, “Comparison of Machine Learning Methods in Prediction of the Gini Coefficient for OECD Countries,” *Data Science and Applications*, vol. 4, no. 1, pp. 16–20, 2021. [Online]. Available: https://dergipark.org.tr/en/pub/datasci/issue/90860/1662110.

# Obstacles We Faced Throughout

- **Data Leakage in Cross-Validation:**
  - ⊗ Standard K-Fold mixes country–year rows.
  - ✓ Switched to GroupKFold by country - preventing inflated $R^2$.

- **Time-Series Feature Engineering**
  - ⊗ Few easy, accurate and interpretable methods.
  - ✓ Compared 2 strategies against our n-1 year baseline.

- **Excessive Missingness**
  - ⊗ Several empty series, risking biased imputations.
  - ✓ Dropped features and countries with large gaps.
  - ✓ Applied interpretable

Model Demo!

Pick a Country, Any Country.

# Strengths & Weaknesses of Best Model
# **GBT + Data Set 2**

**Strengths**

- **Top accuracy** of all models
- Captures **non-linear + interaction** effects automatically.
- Relies on **widely reported** metrics (electricity access, forest %, pop density, etc.) → no new surveys needed.
- Trained with **grouped CV** → proven to generalise to unseen countries.

**Weaknesses**

- Less interpretable – feature importance shows *weight*, not direction.
- 2nd-ranked variable *intentional homicides (35 % gaps)*
- Key policy levers (tax, education spend) absent from top-10 → model may lean on proxies.
- Only small gain over simpler linear + rolling (≈ 3 pp $R^2$) for higher complexity.

# Conclusion & Next Steps

- **Overall** found that trees shine on lagged data.
- **Surprisingly,** engineered simplicity is a close second (linear regressors)
- **Core Drivers of Inequality:** Homicide Rates, Forest Land, Death Rates
- **Next Steps:**
  - Implement Advanced Imputation.
  - Engineer Domain Driven Ratios (like GDP / pop)
  - Explore Dimensionality Reduction to reduce noise.
  - Check Model Fairness across income tiers and countries
  - Explore Bayesian Modelling

# Thank You!

| Model | Data Set 1 (year 1) | | Data Set 2 (year 1 + rolling mean + std) | | Data Set 3 (year 1-5) | | Avg $R^2$ % |
|---|---|---|---|---|---|---|---|
| | $R^2$ % | Hyper parameters | $R^2$ % | Hyper parameters | $R^2$ % | Hyper parameters | |
| Simple | -2.2 | | -2.2 | | -2.2 | | -2.2 |
| Linear Regression | 39.9 | N/A | 53.4 | N/A | 43.0 | N/A | 45.1 |
| Linear + Feature Selection | 40.0 | N/A | 53.1 | N/A | 44.2 | N/A | 45.8 |
| Ridge Regression | 34.7 | α = 100 | 42.5 | α = 1000 | 38.4 | α = 1000 | 38.5 |
| Lasso Regression | 35.6 | α = 0.01 | 51.1 | α = 0.1 | 33.3 | α = 1 | 40.0 |
| Random Forest | 54.3 | n_estimators = 50, max_depth = 14 | 49.3 | n_estimators =120 max_depth = 15 | 49.7 | n_estimators =260 max_depth = 14 | 51.1 |
| Gradient Boosted Trees | 50.6 | n_estimators = 20 lr = 0.16 max_depth = 3 | 53.6 | n_estimators = 50 lr = 0.06 max_depth = 3 | 50.4 | n_estimators = 90 lr = 0.16 max_depth = 3 | 51.5 |
| K-Nearest Neighbors | 45.4 | k = 42 P = 1 ‘distance’ | 43.9 | k = 42 P = 1 ‘distance’ | 45.4 | k = 39 P = 1 ‘distance’ | 44.6 |
| Avg $R^2$ % | 42.9 | | 49.6 | | 43.5 | | |