Global Corruption

Observations and Modeling

DSI course

Jan 2024

By:

JR Garvin

Tanner Zuleeg

Massoud ‘Massi’ Alfi

Corruption Perception Index (CPI)

- Corruption is defined as the abuse of entrusted power for private gain.

- **Corruption Perception Index (CPI):**

  A score introduced by Transparency International to measure how corrupt each country's public sector is perceived to be.

  - 0 to 100
  - High scores means less corrupt

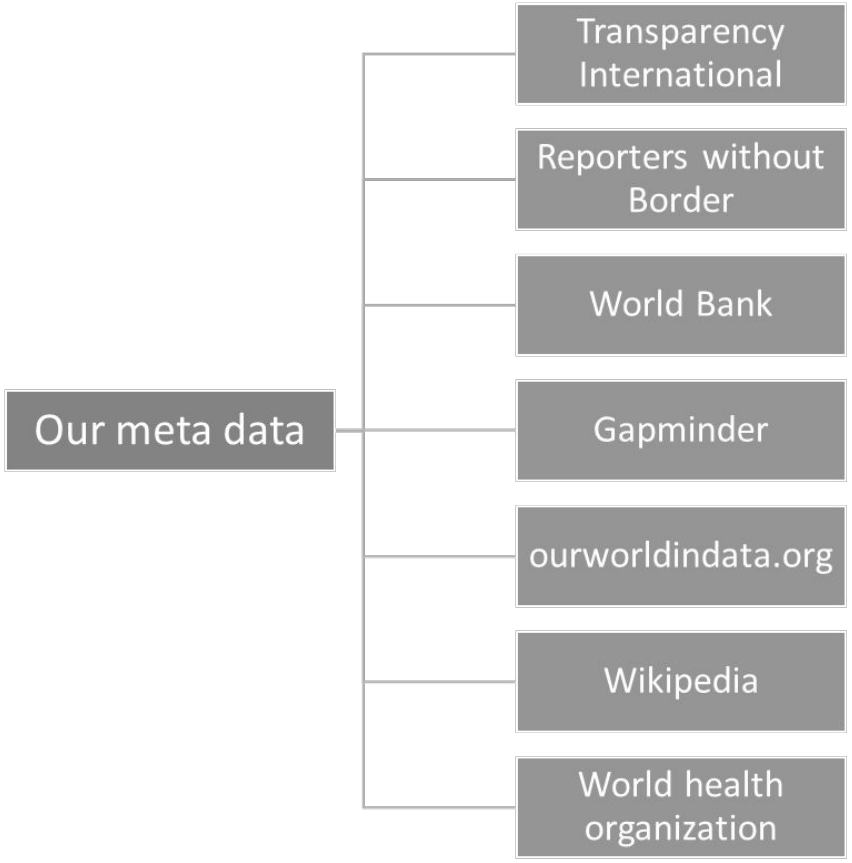
Problem Statement

- How are countries around the world fighting corruption?
- Is global corruption improving over time?
- Can we model/predict corruption?
- **What we did and why is it important?**

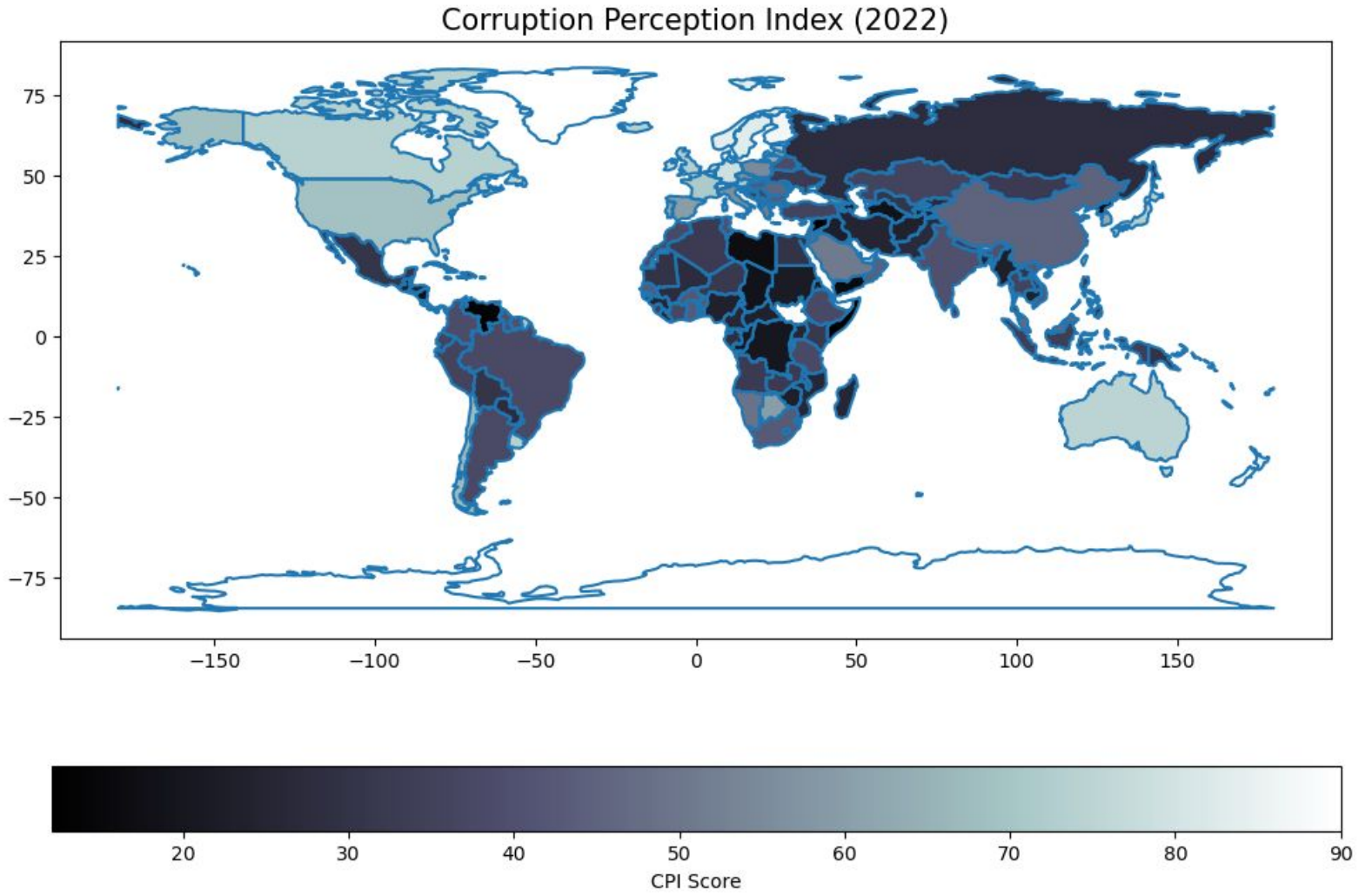
Our Data (to be fixed)

- What we used to predict global corruption?

  - Wealth of the country
  - Political indicators
  - Press freedom
  - Quality of life indicators
  - Ethnic and religious division
  - Social norms
  - Public health indicators

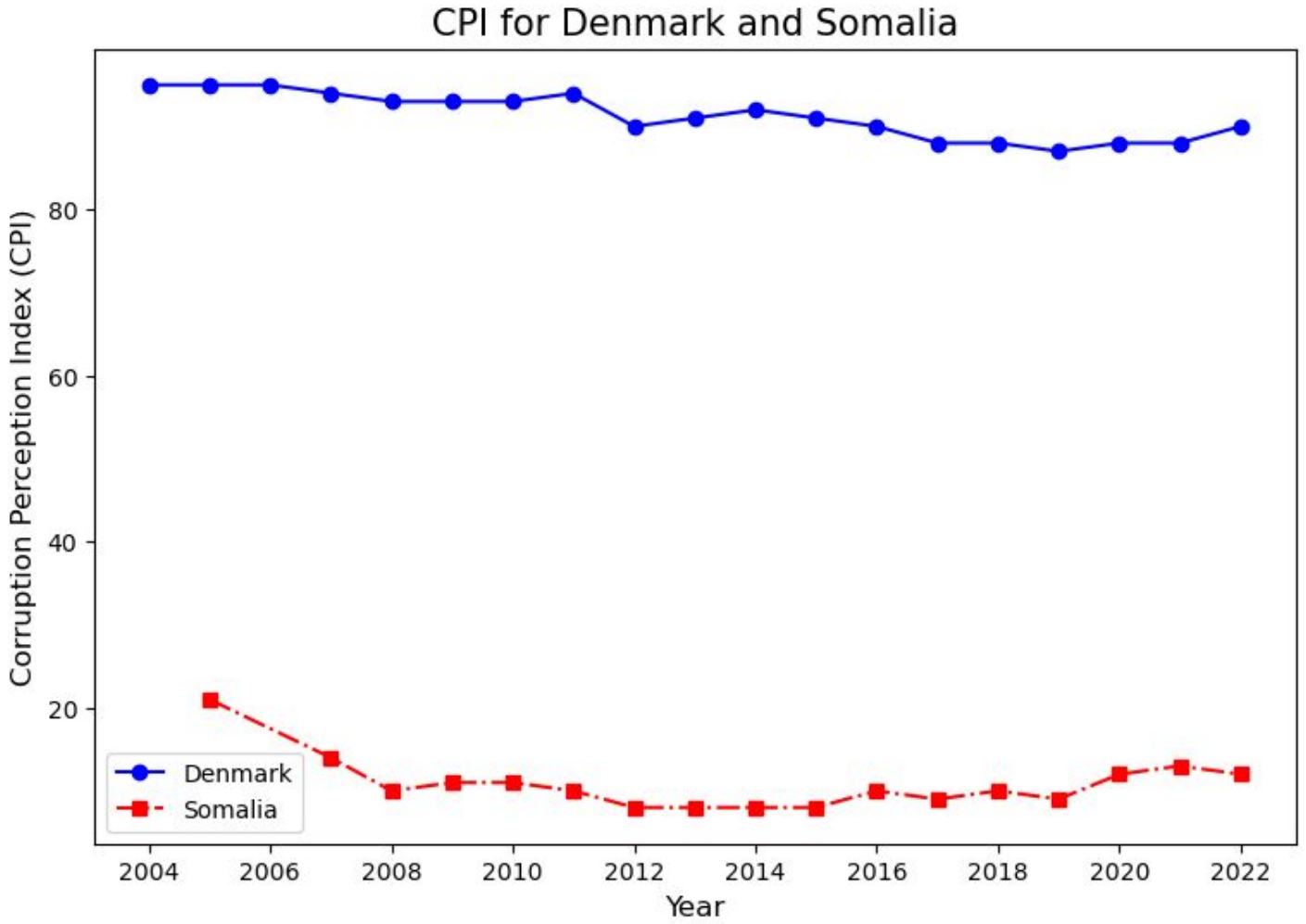


CPI in 2022



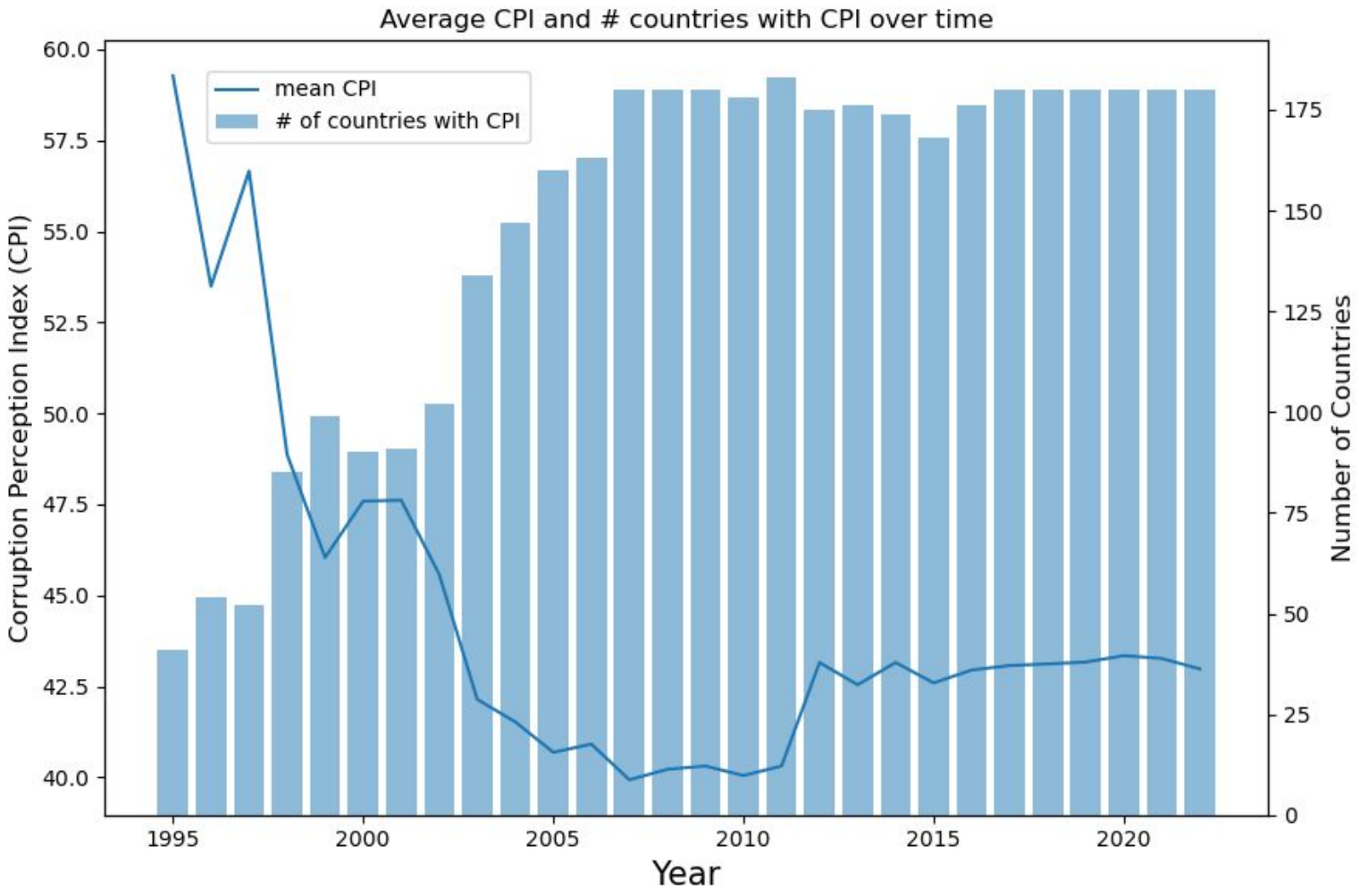
- Different regions

Goods and not Goods in 2022



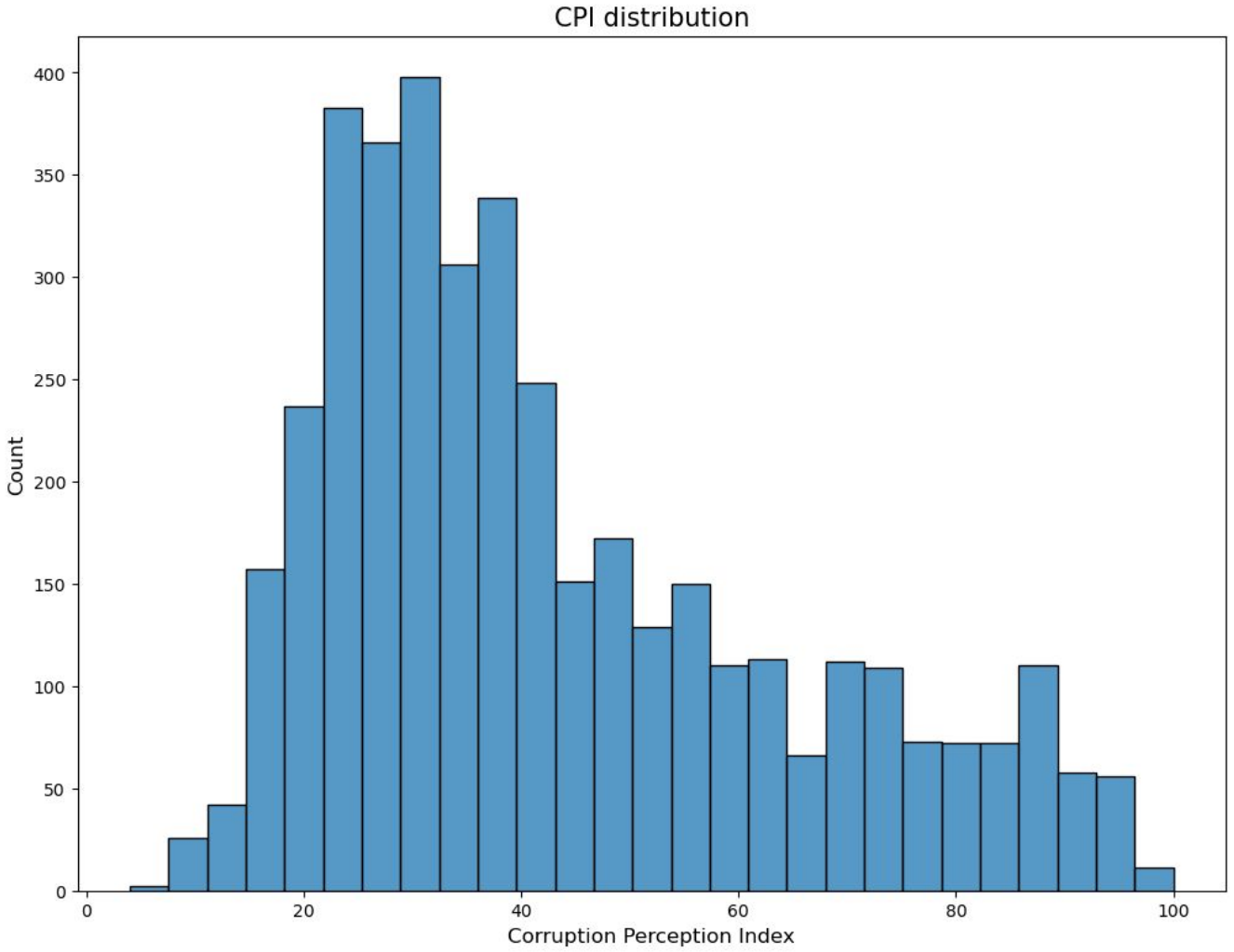
- Last 10 years for the first and last countries
- No improvements

How is the World Doing?



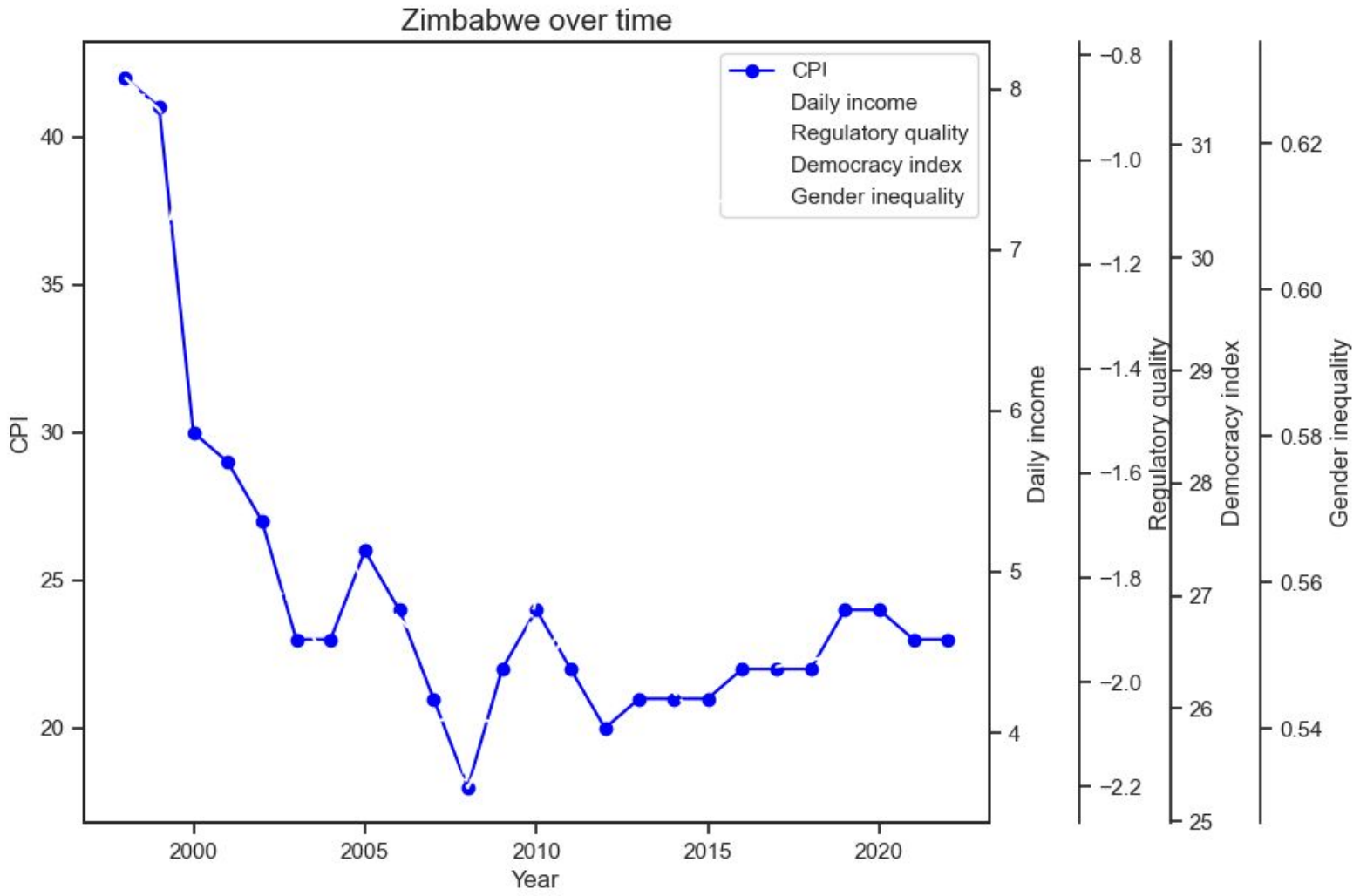
- World average is going down
- 55 countries had their decade's minimum in 2022

Overall CPI Distribution

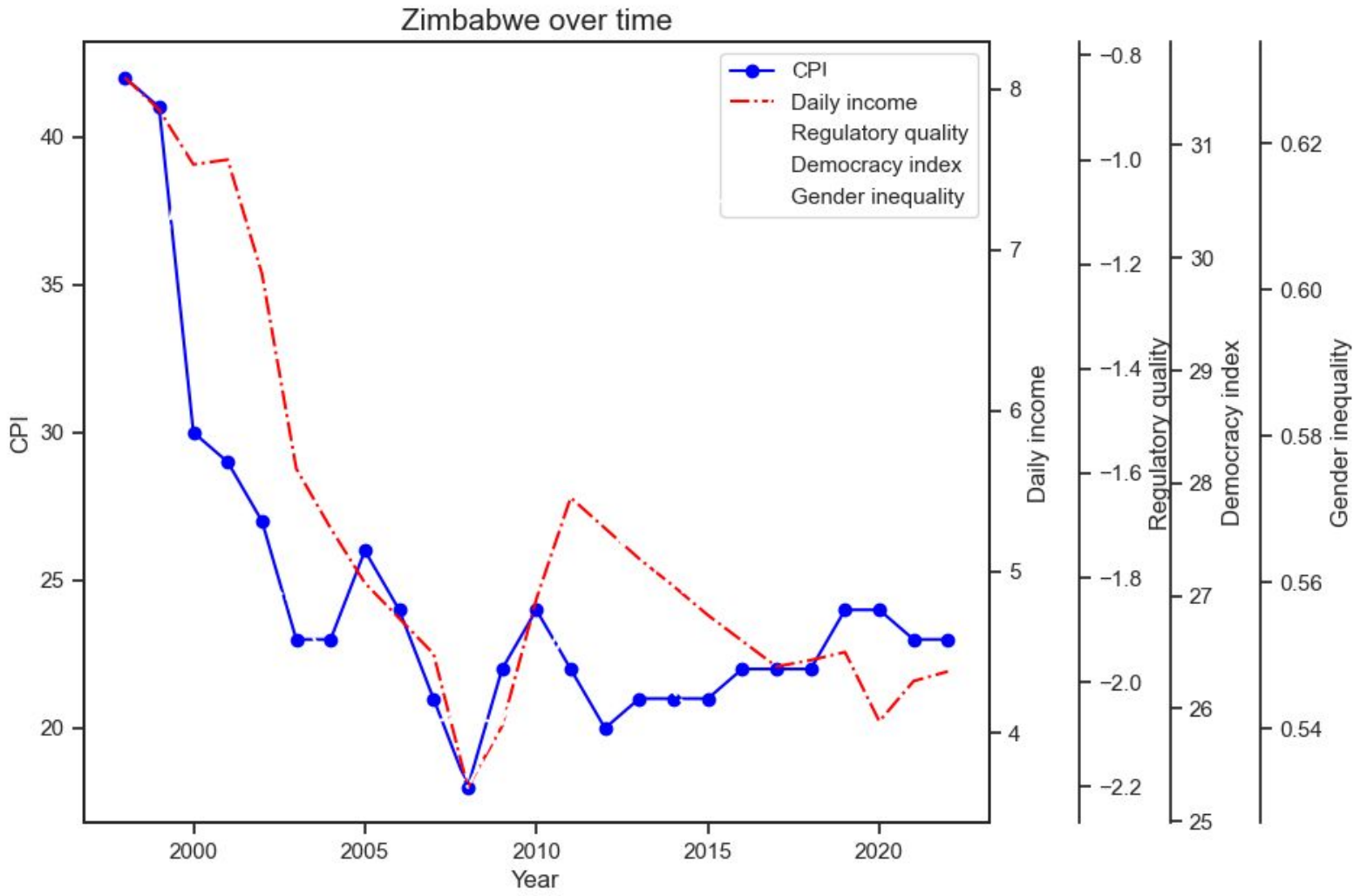


- Right skewed
- No outliers

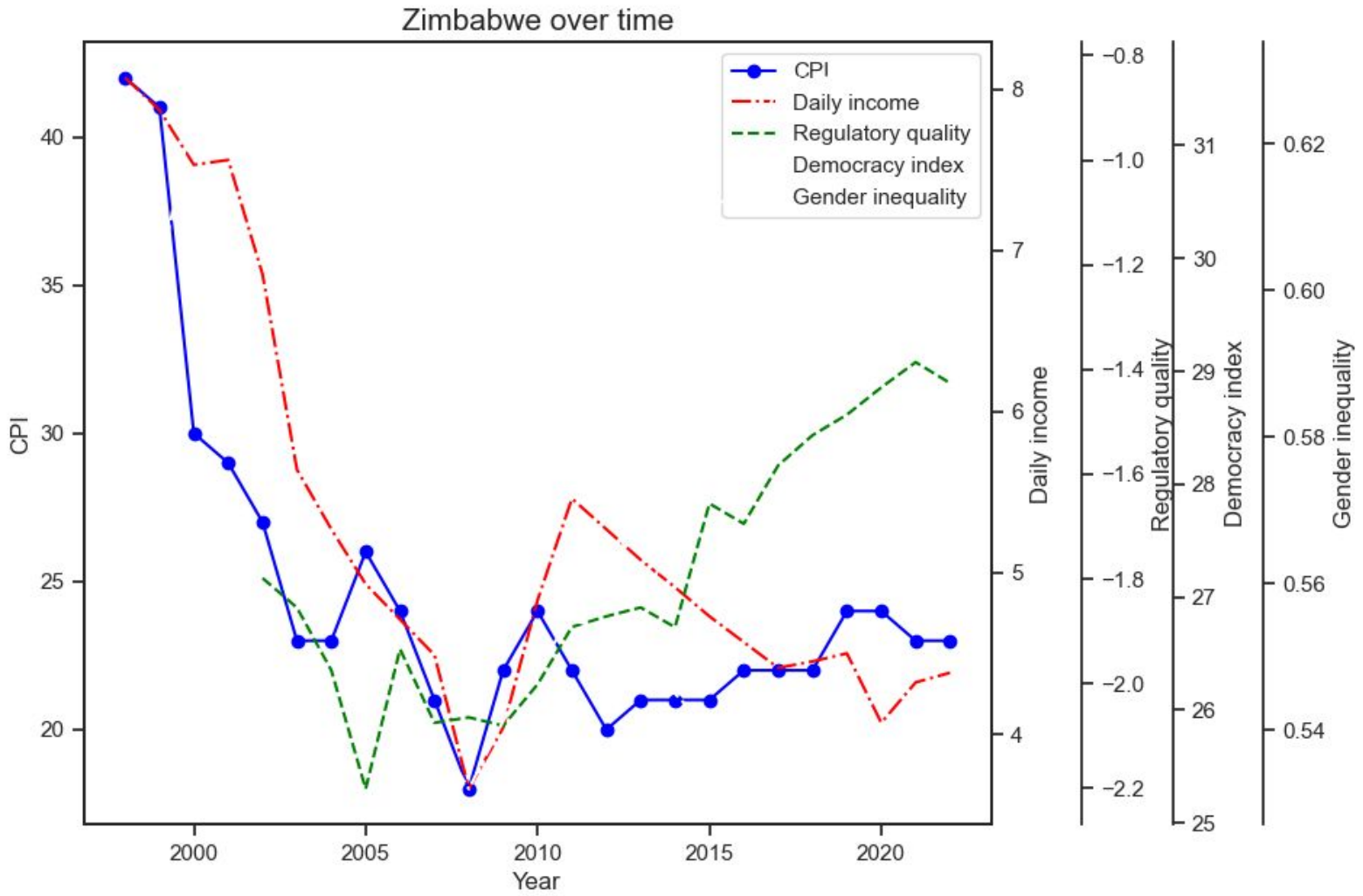
Zimbabwe over Time



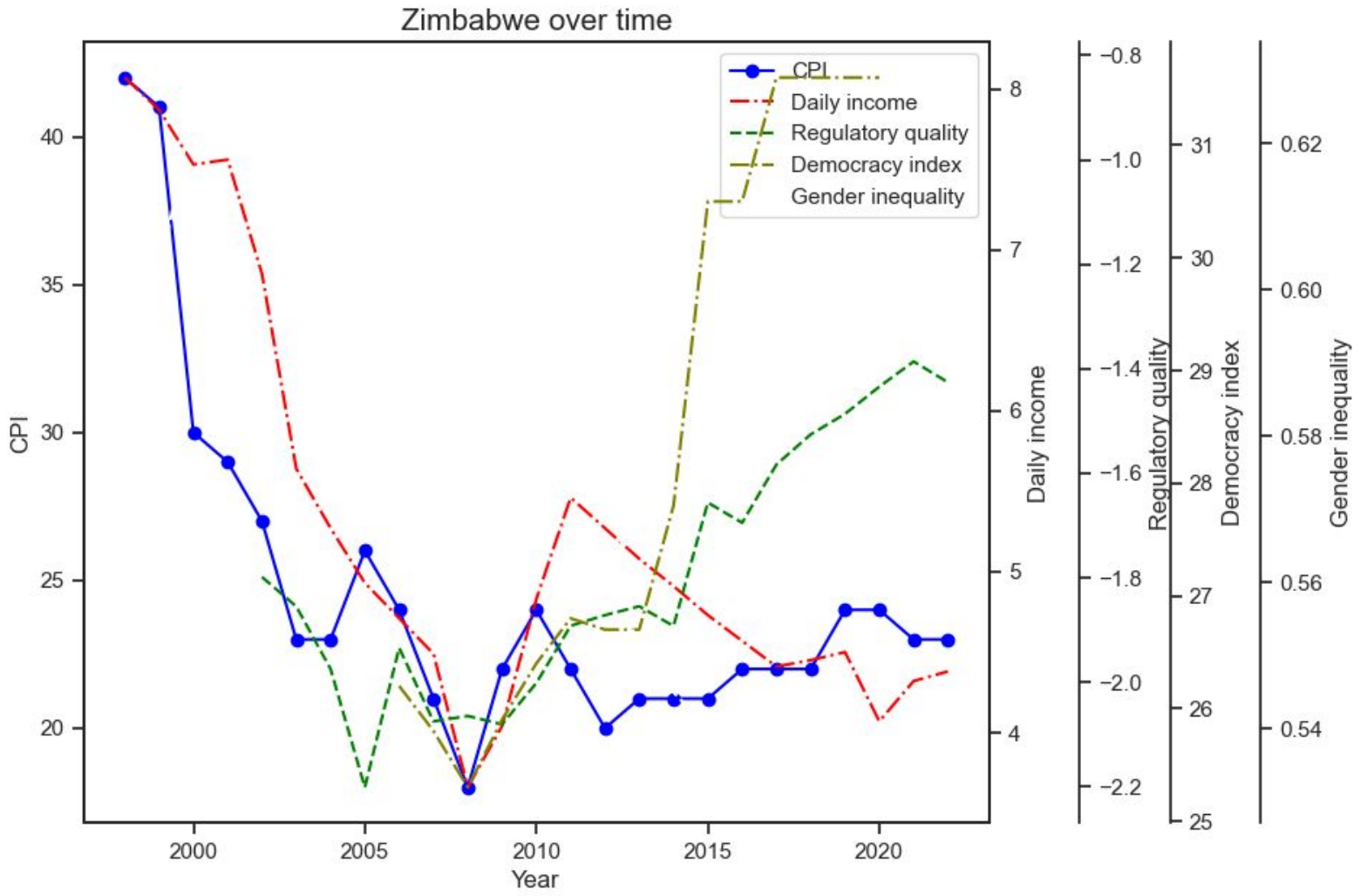
Zimbabwe over Time



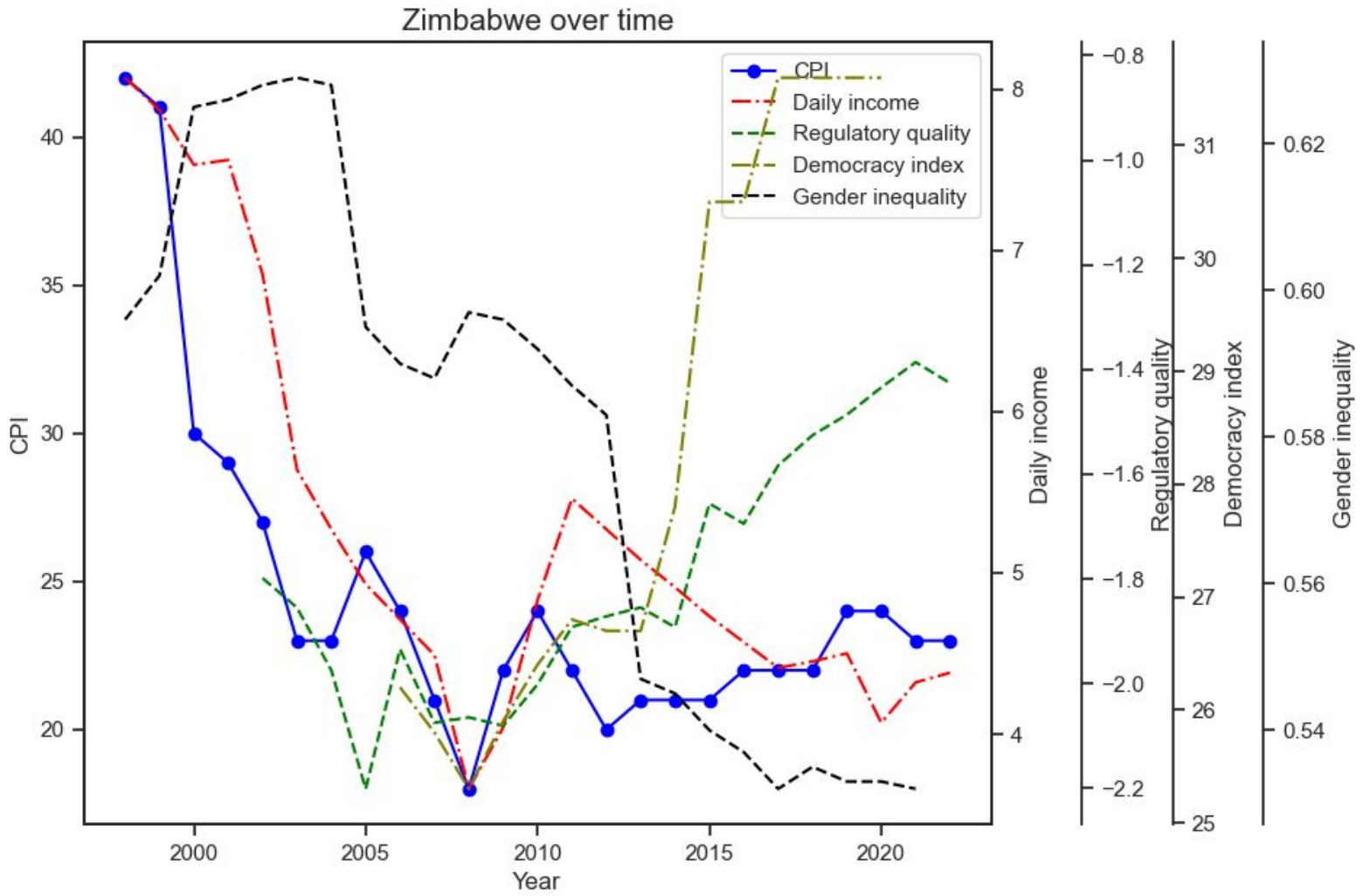
Zimbabwe over Time



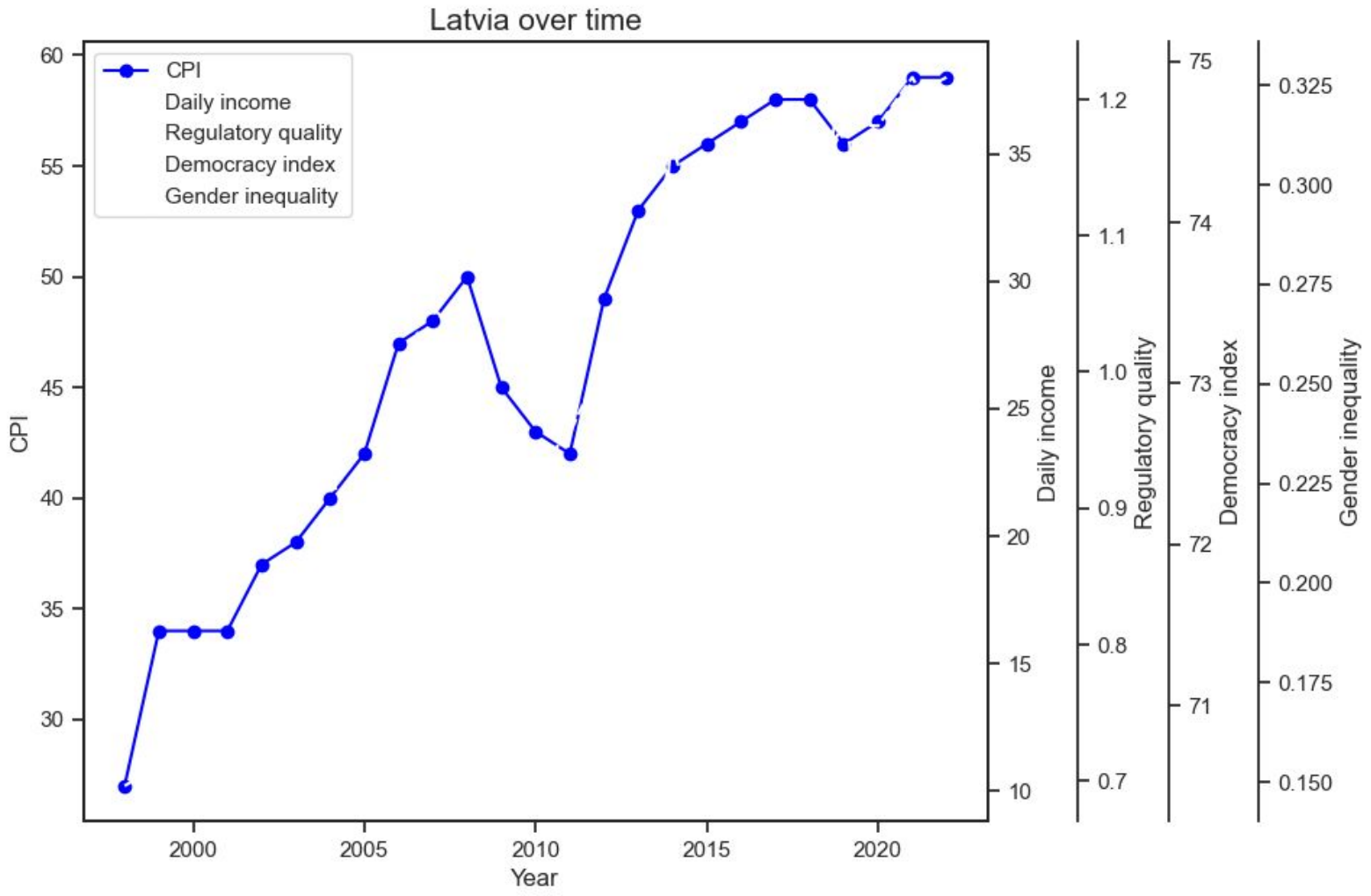
Zimbabwe over Time



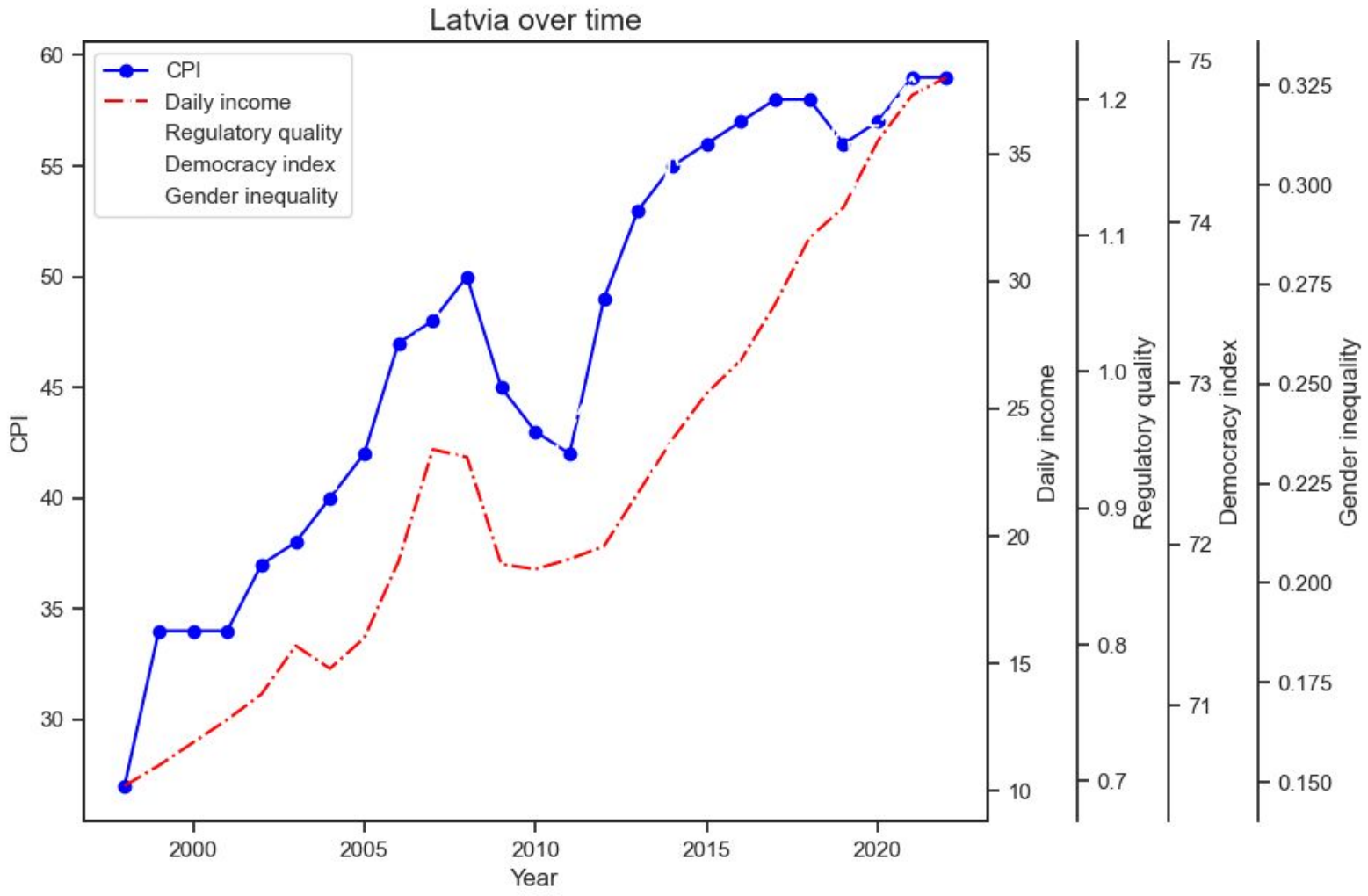
Zimbabwe over Time



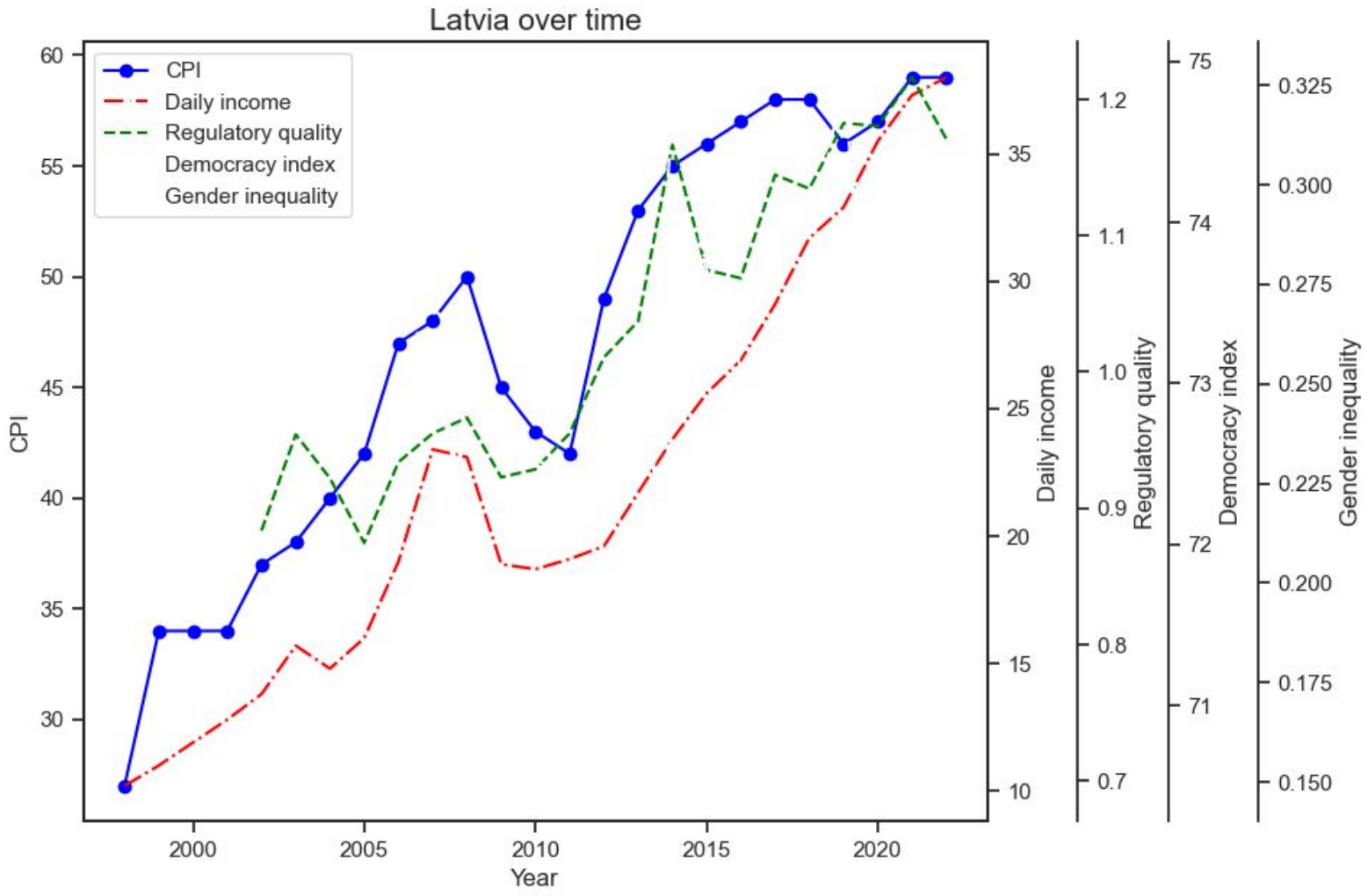
Latvia over Time



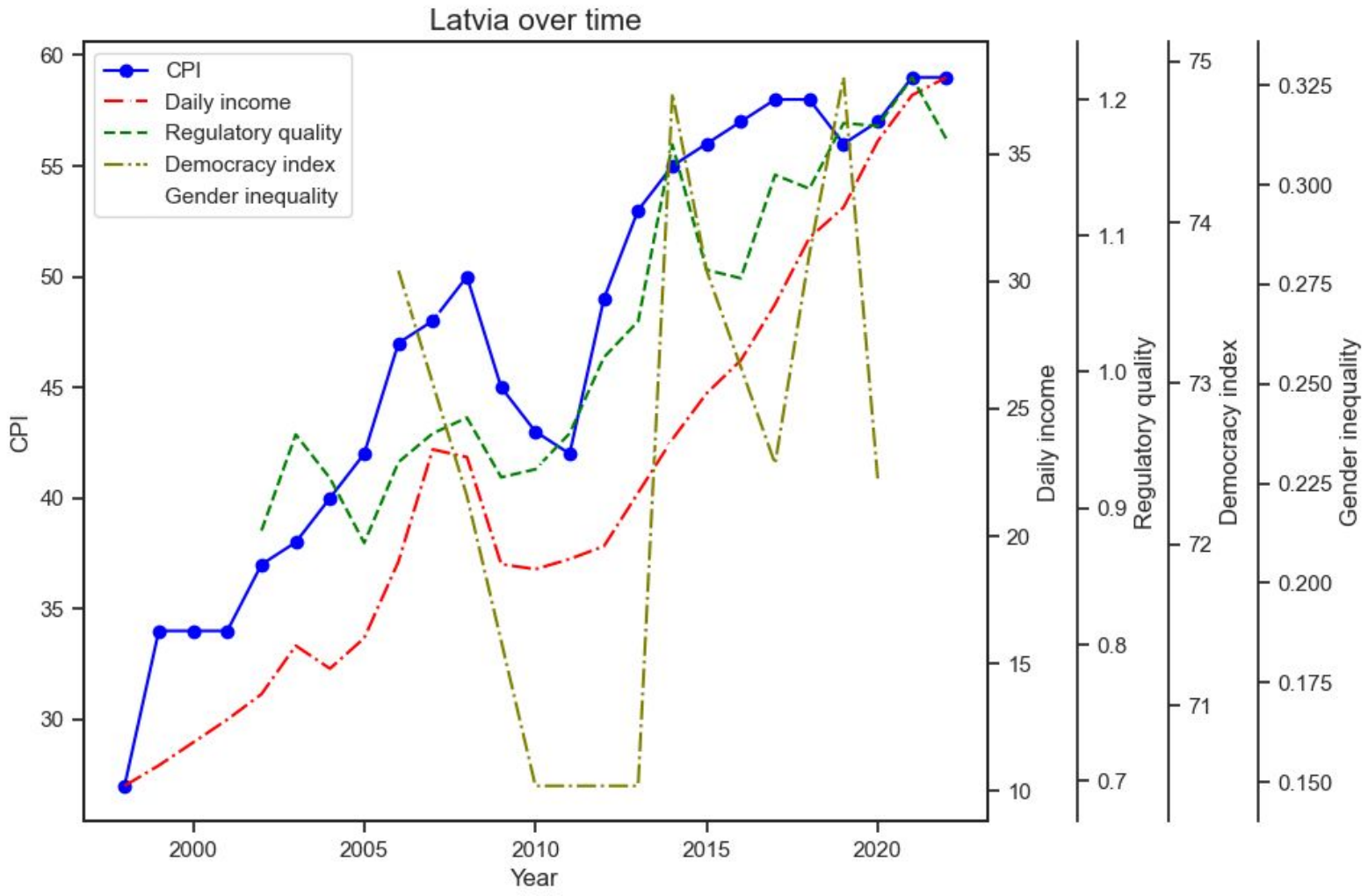
Latvia over Time



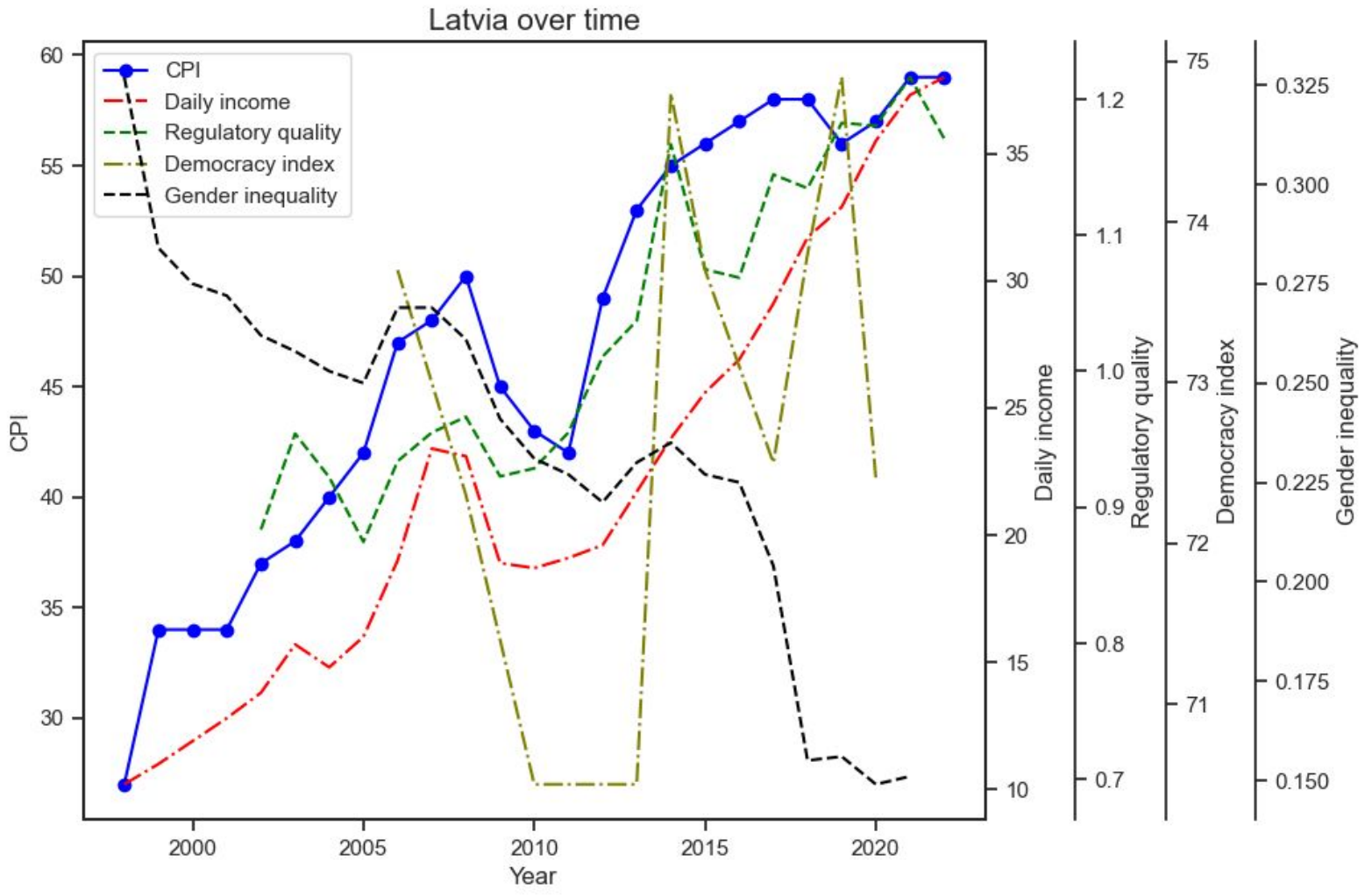
Latvia over Time



Latvia over Time

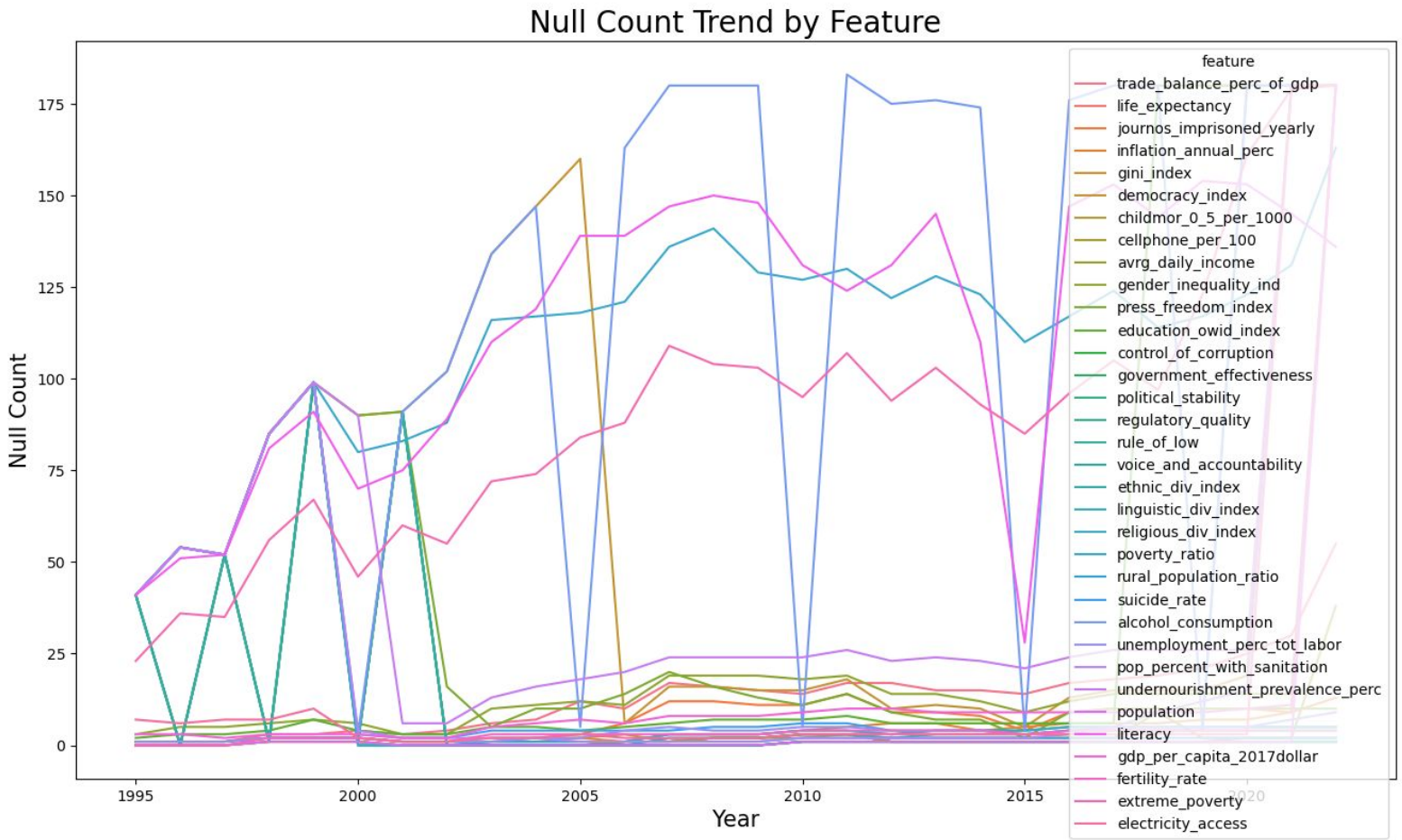


Latvia over Time



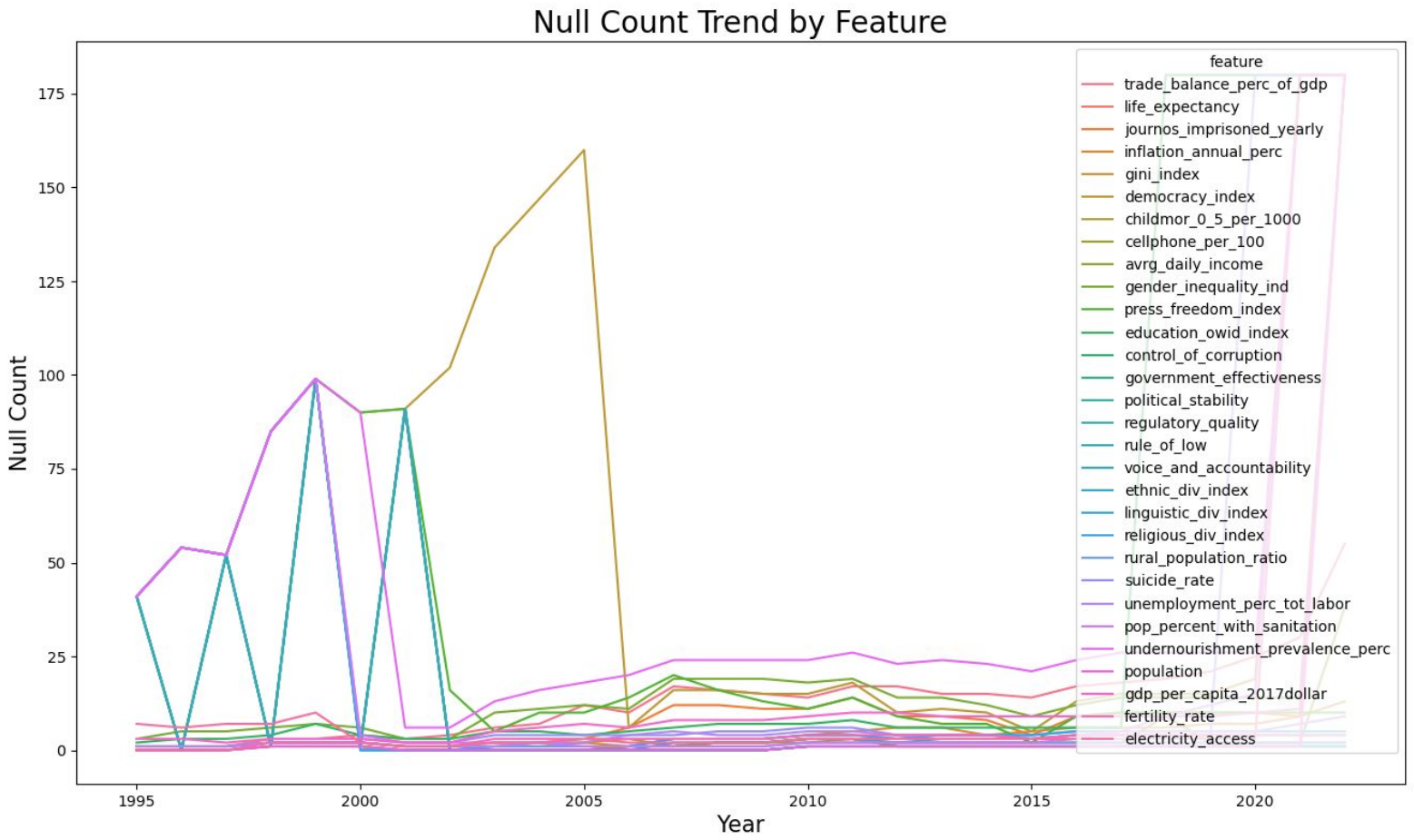
- Daily income
- Regulatory quality

Reconcile Nulls



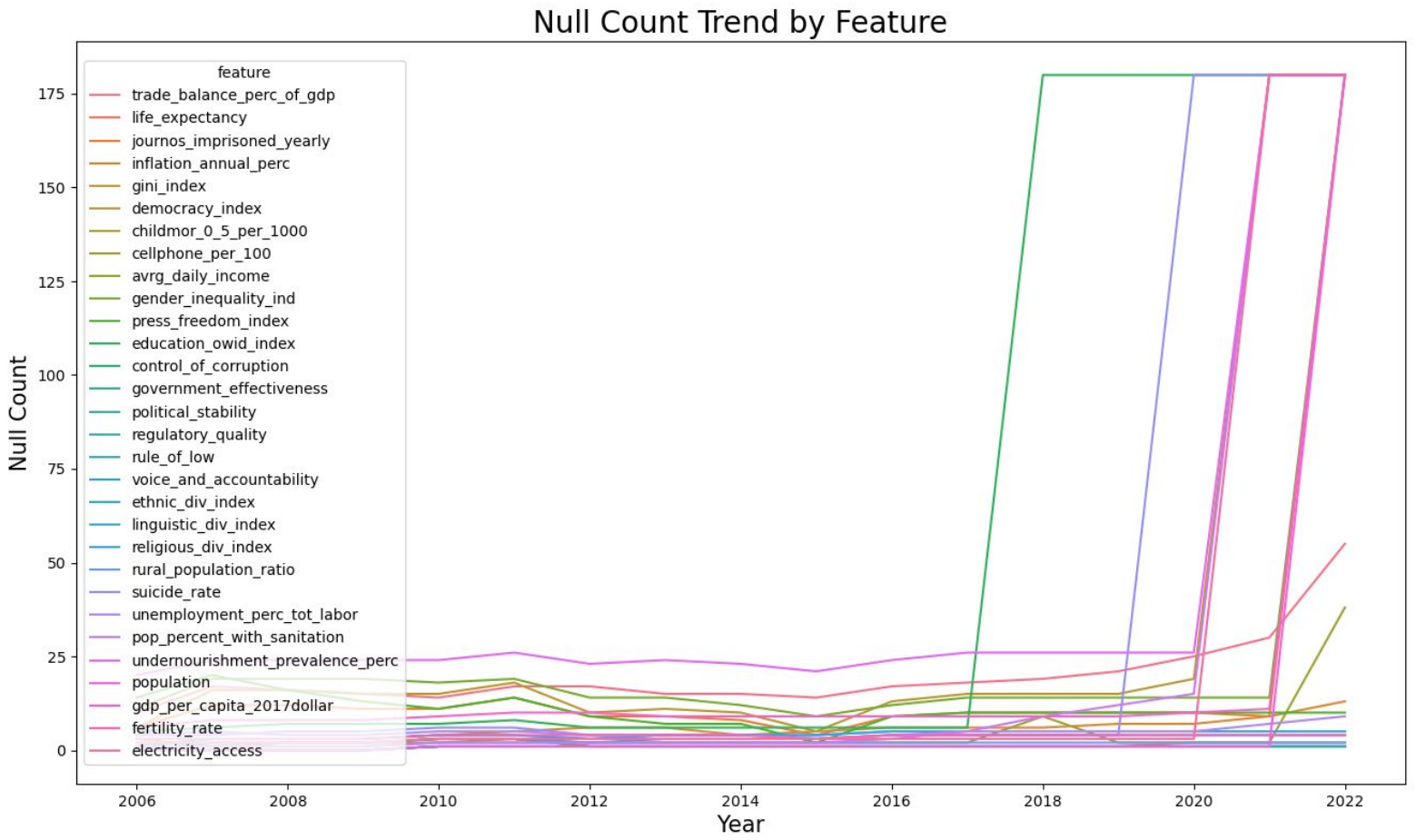
- Drop rows w/o CPI
- 34 features
- 4068 rows

Reconcile Nulls



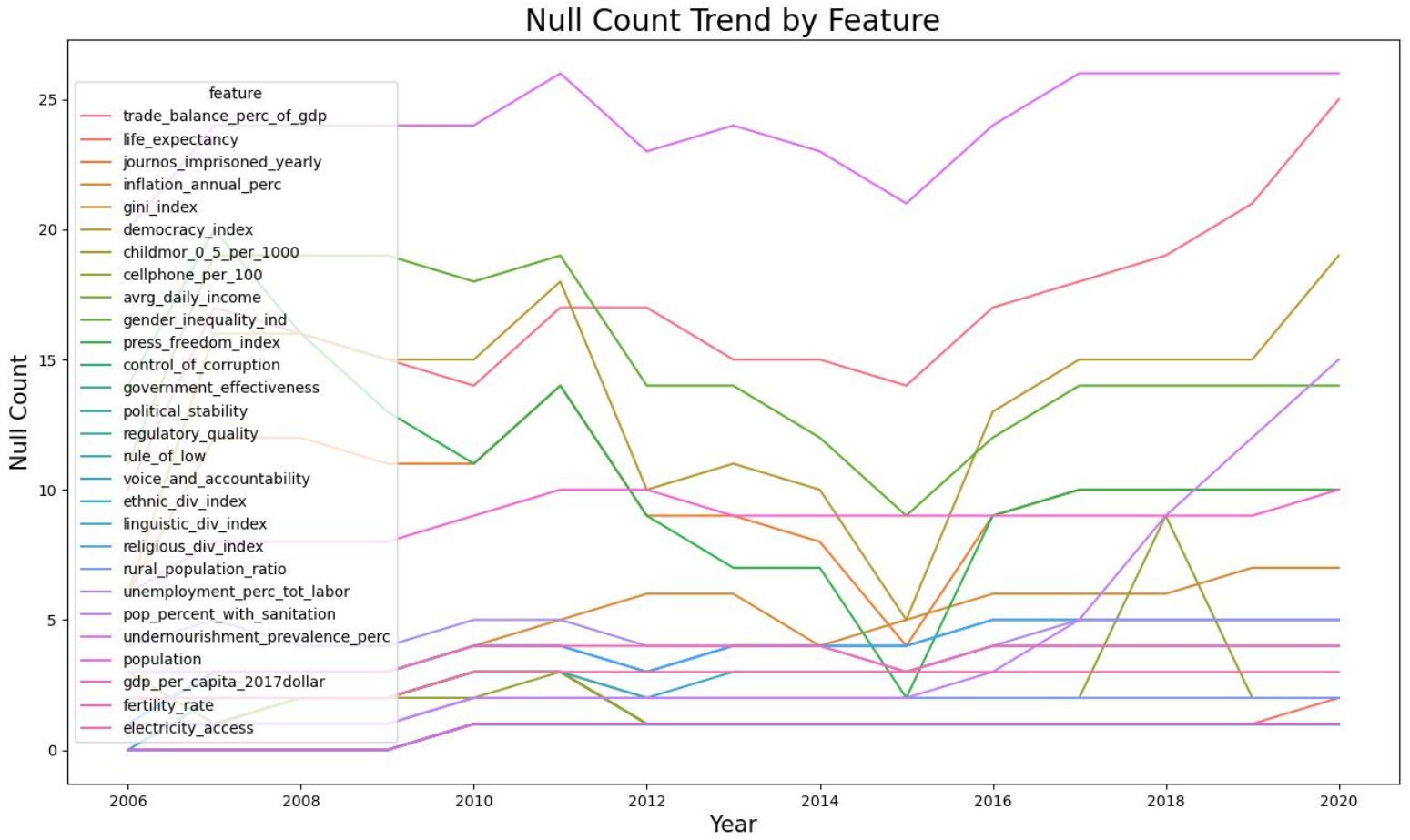
- Drop 4 columns
- 30 features
- 4068 rows

Reconcile Nulls



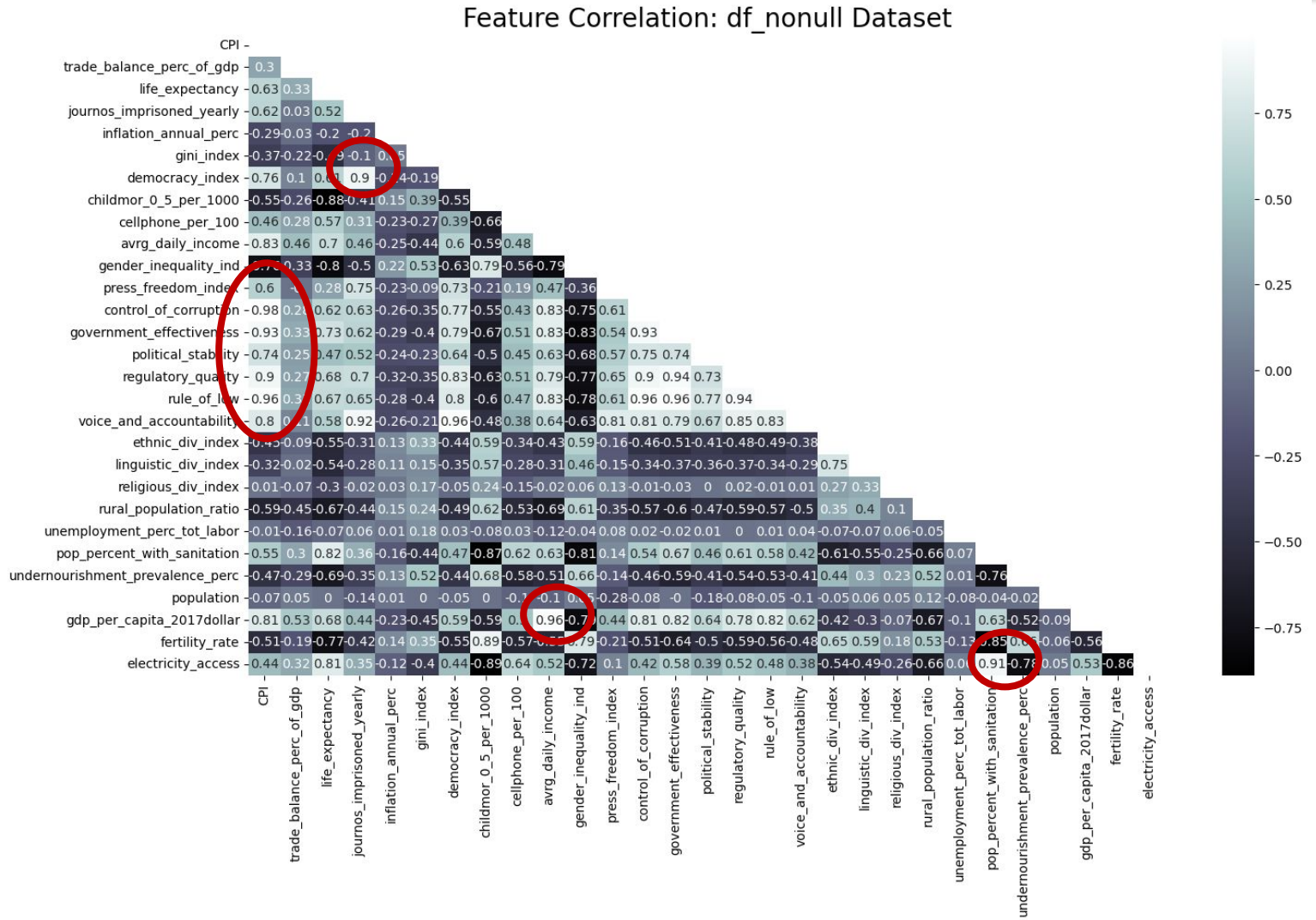
- Drop data prior to 2006
- 30 features
- 3013 rows

Reconcile Nulls



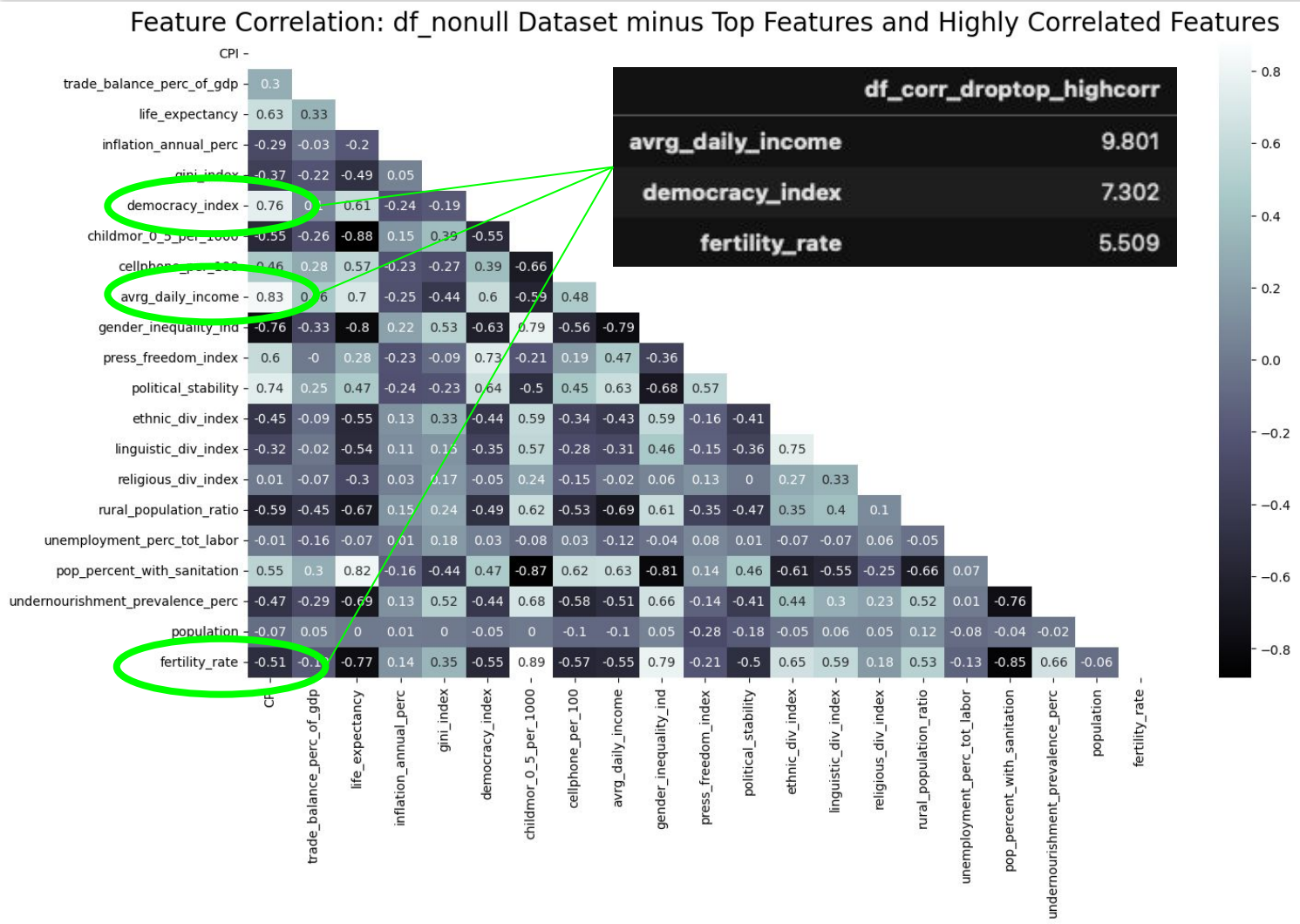
- Drop data after 2020
- Drop 2 columns
- 28 features
- 2653 rows

Collinearity Reduction



- Drop eight features, correlation issues
- 20 features
- 1859 rows after drop nulls

Linear Regression



- Training score 0.845
- Test score 0.835

Model Optimization

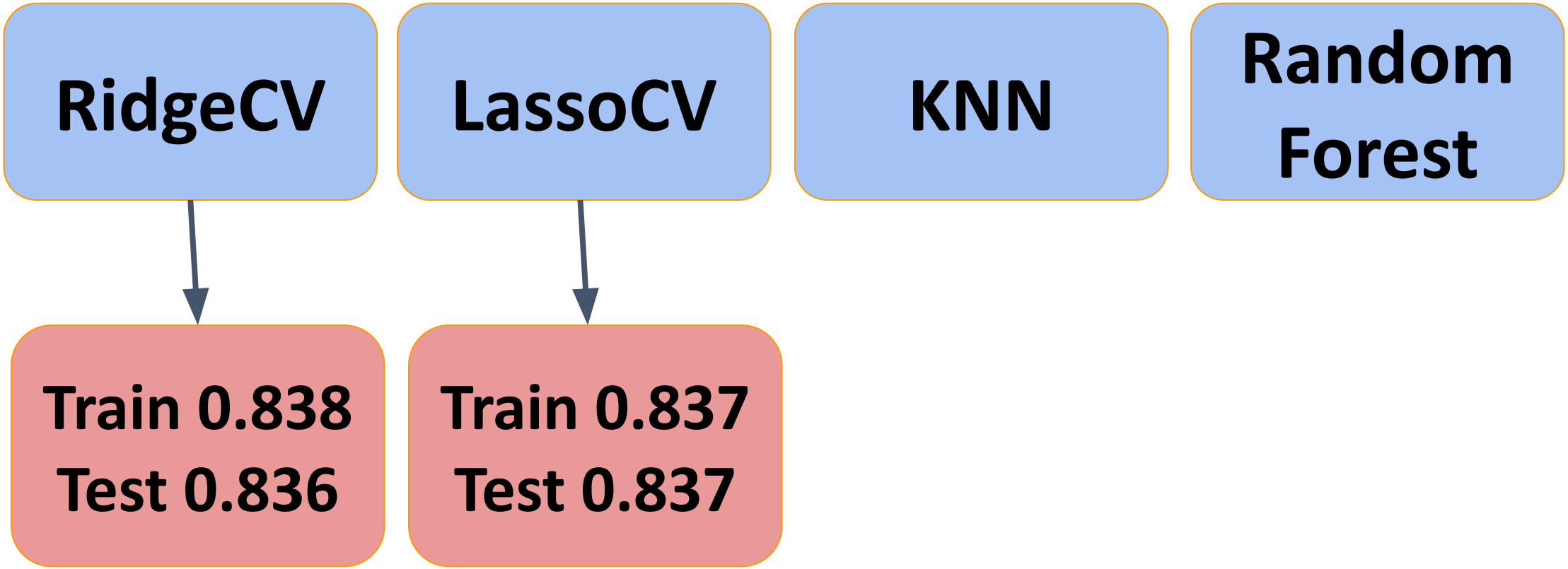
RidgeCV

LassoCV

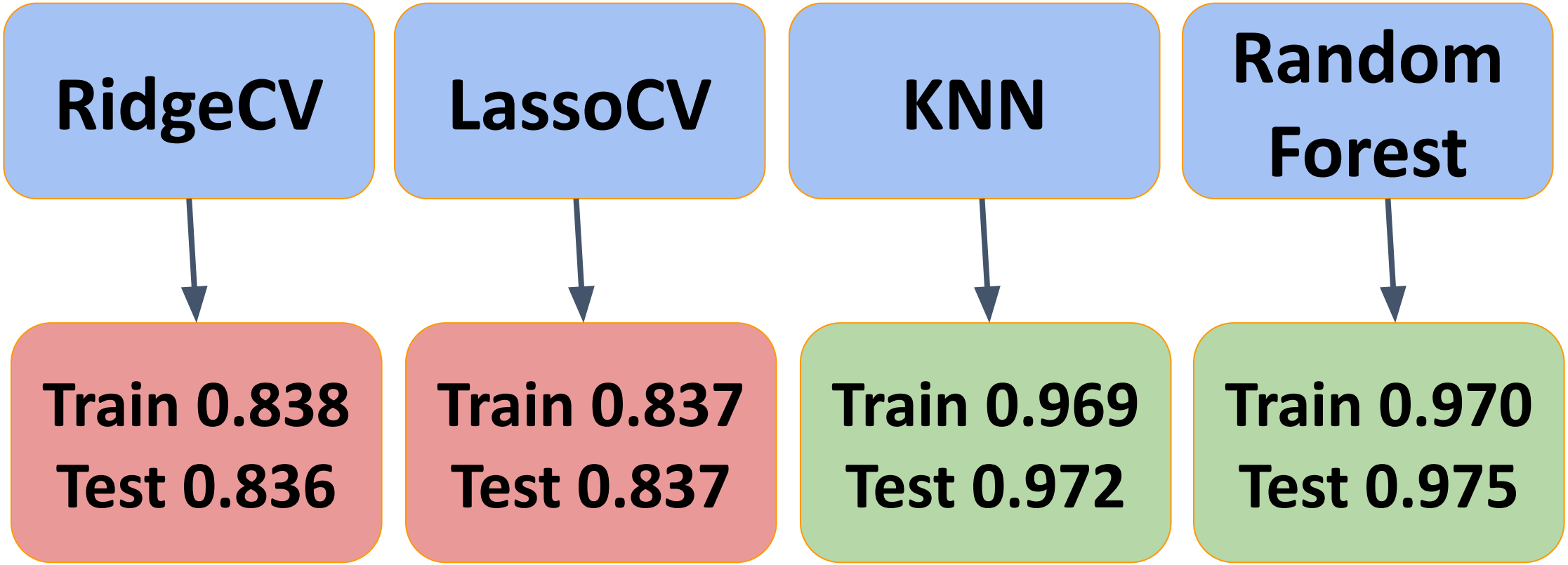
KNN

Random Forest

Model Optimization



Model Optimization



Model Optimization

Boosted	Boosted
KNN	Random Forest

Model Optimization

Boosted

KNN

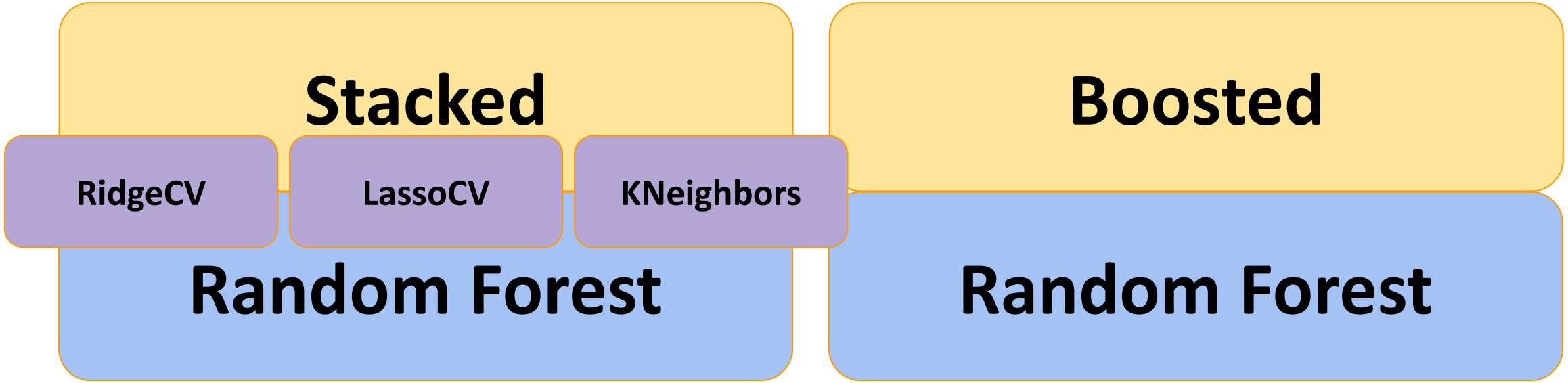
Train 0.966
Test 0.970

Boosted

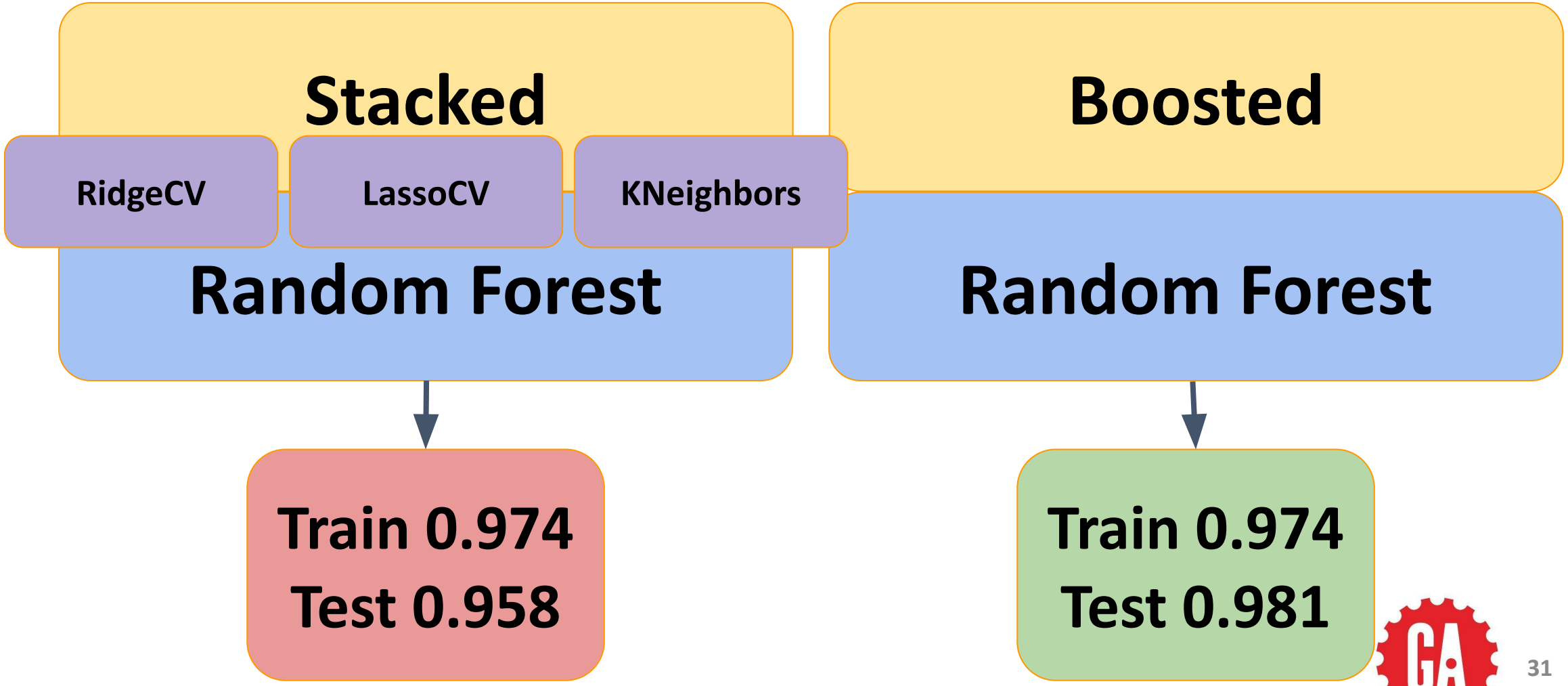
Random Forest

Train 0.974
Test 0.981

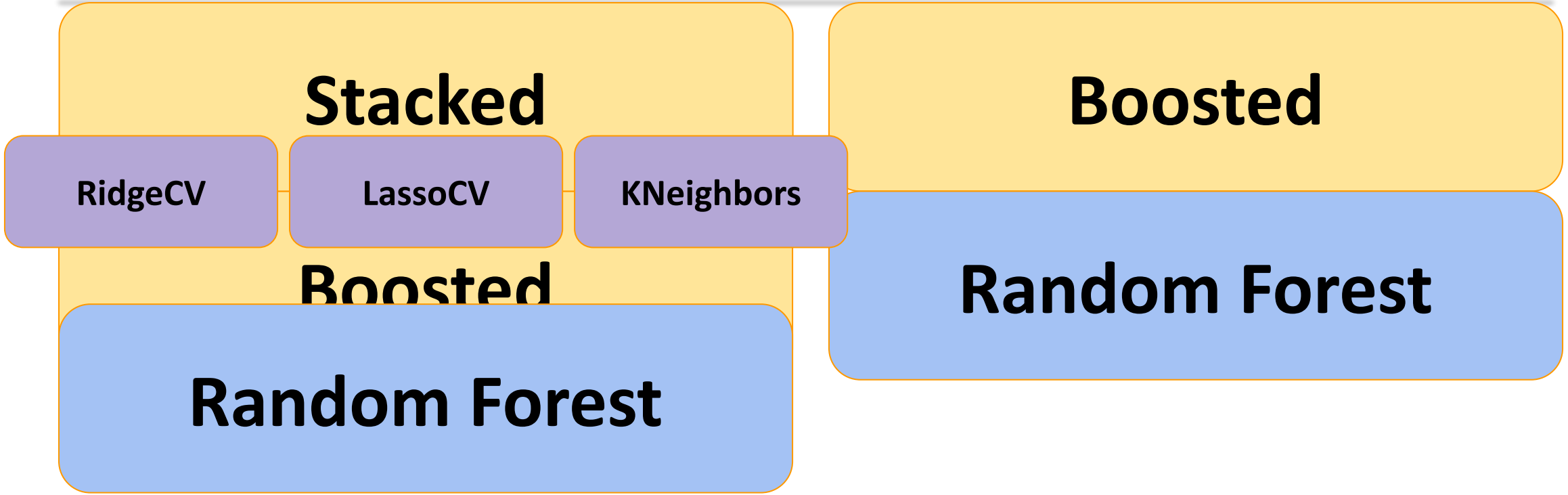
Model Optimization



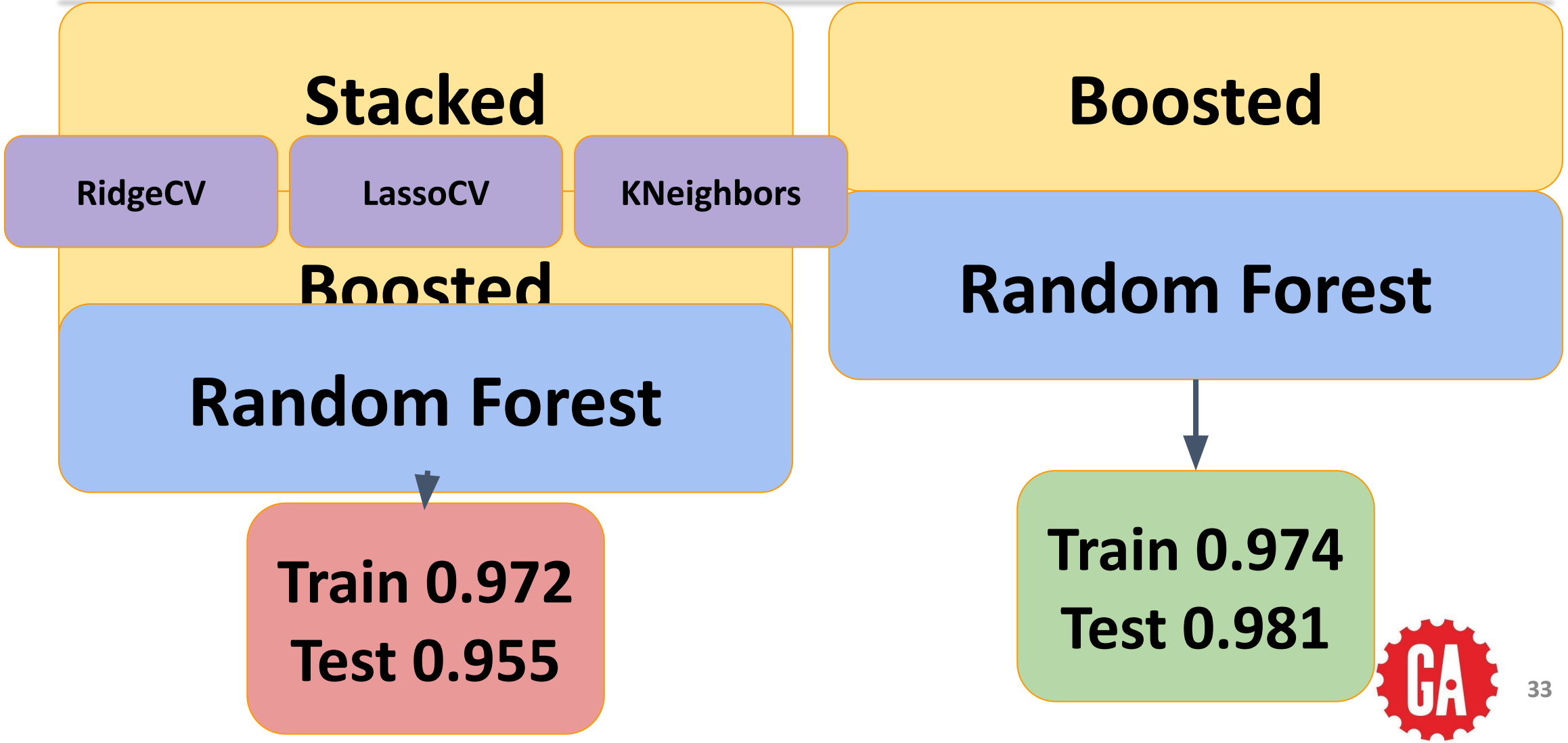
Model Optimization



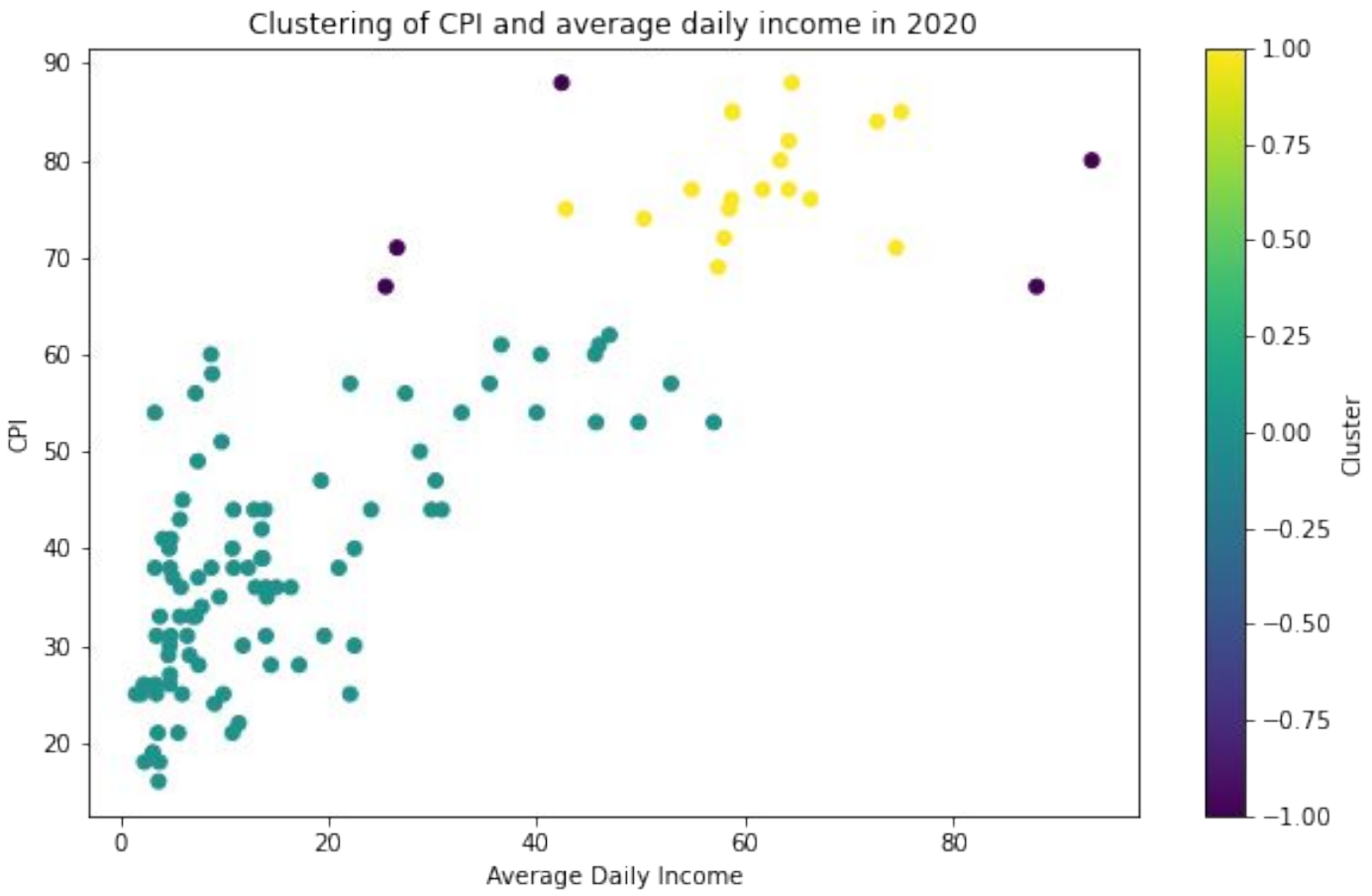
Model Optimization



Model Optimization



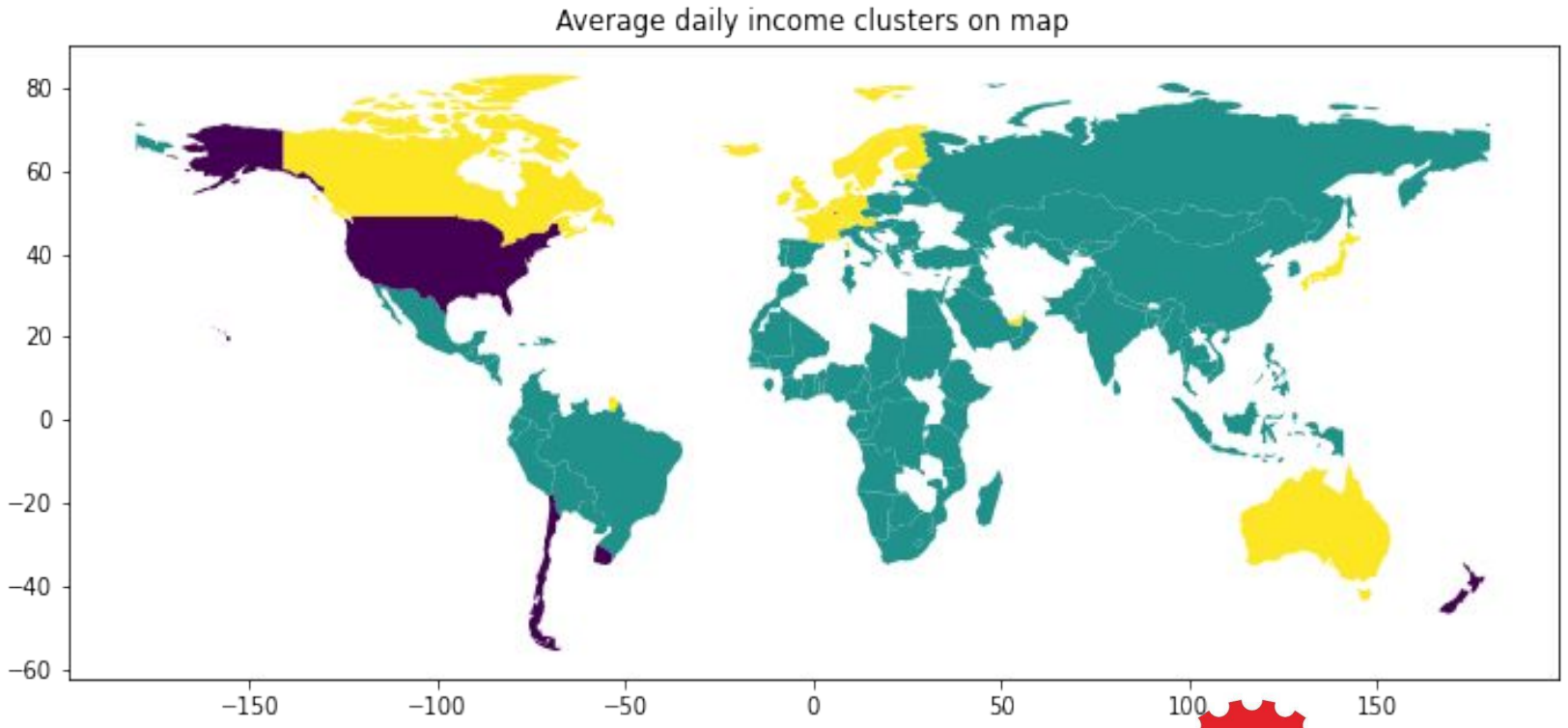
Analysis from clustering



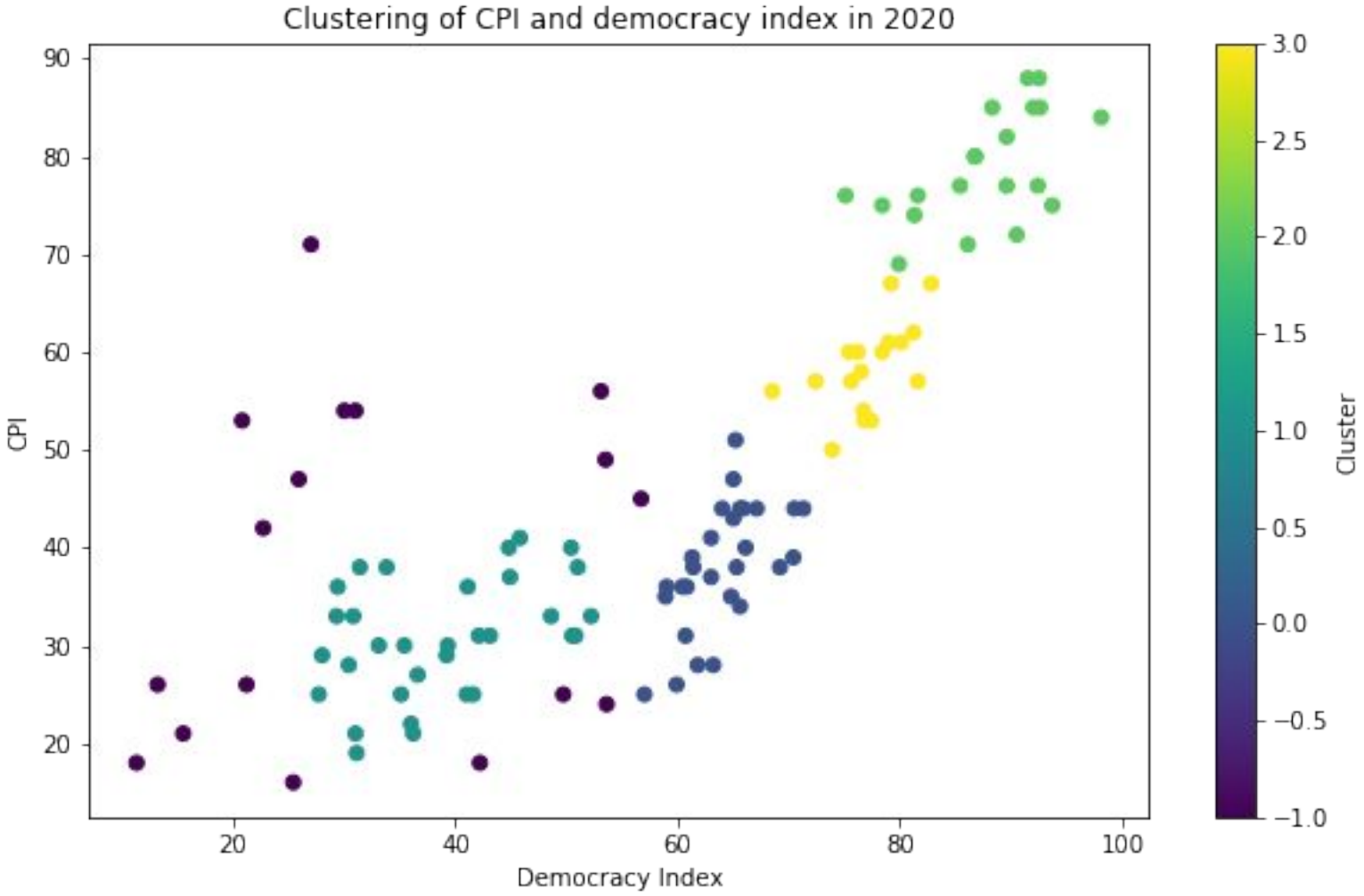
- Average daily income versus CPI in 2020
- Two clusters
- Roughly low and high CPI

Daily Income clusters visualized

- Most of northern Europe, Canada, Japan and Australia
- US, Chile, New Zealand (non clusters)
- Rest of world



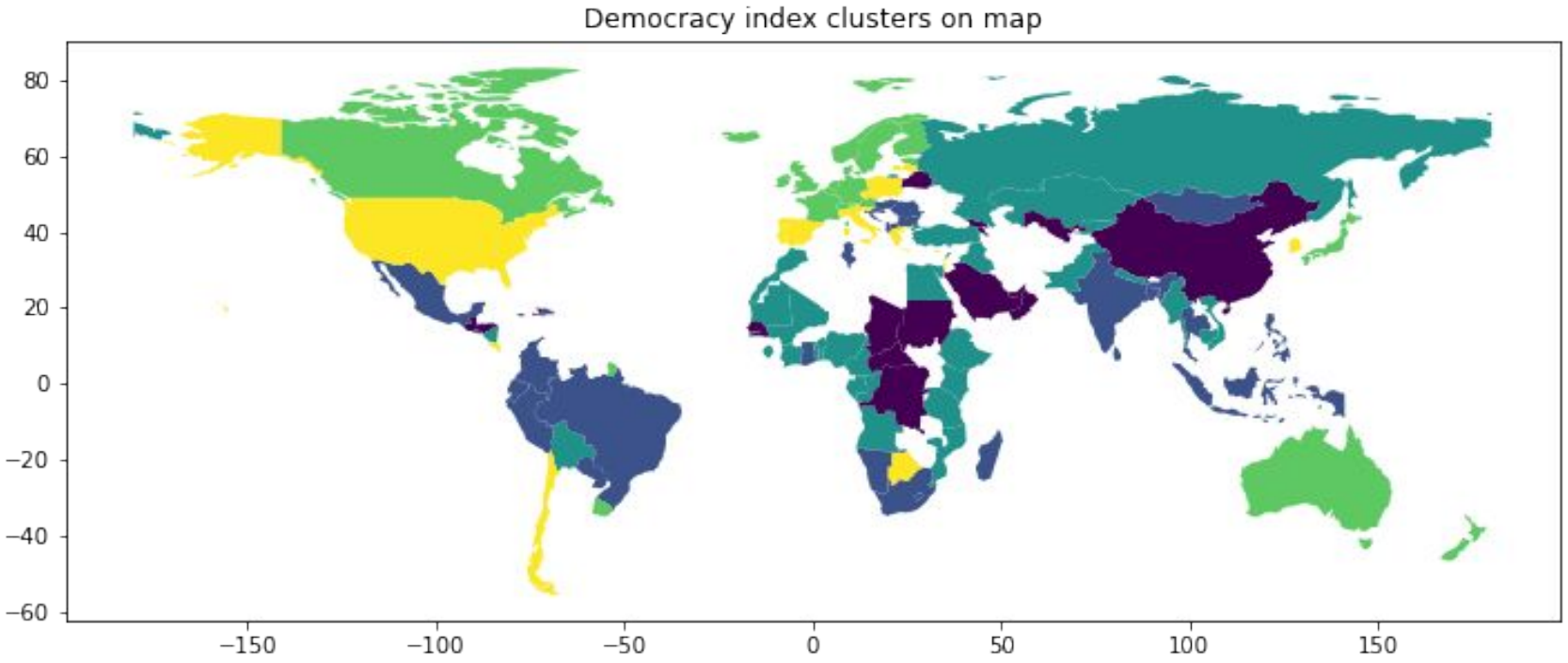
Democracy VS corruption



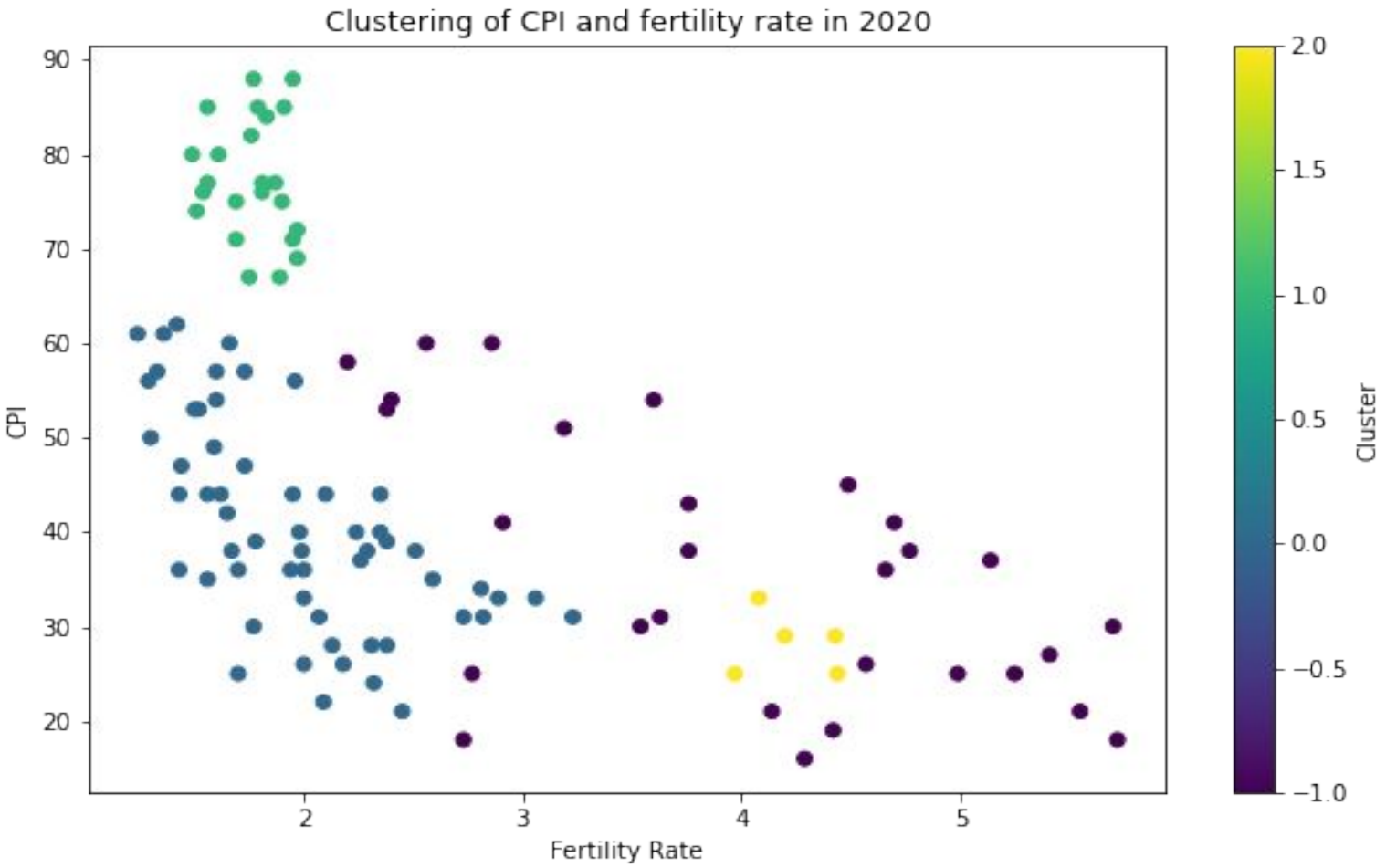
- Four clusters
- Ascending order (less corrupt)

The map for democracy and corruption

- Canada with northern Europe and Australia



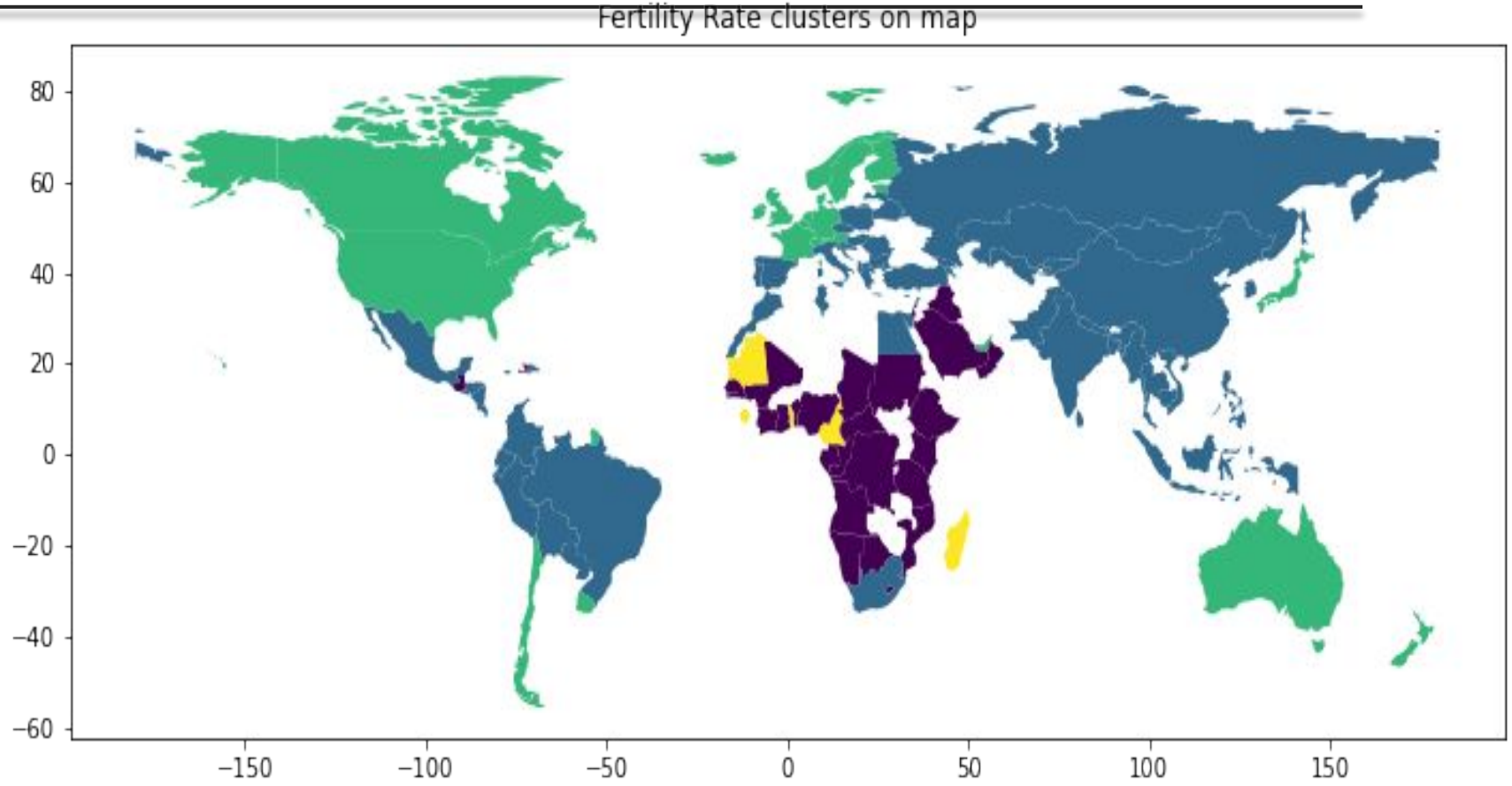
Fertility and Corruption



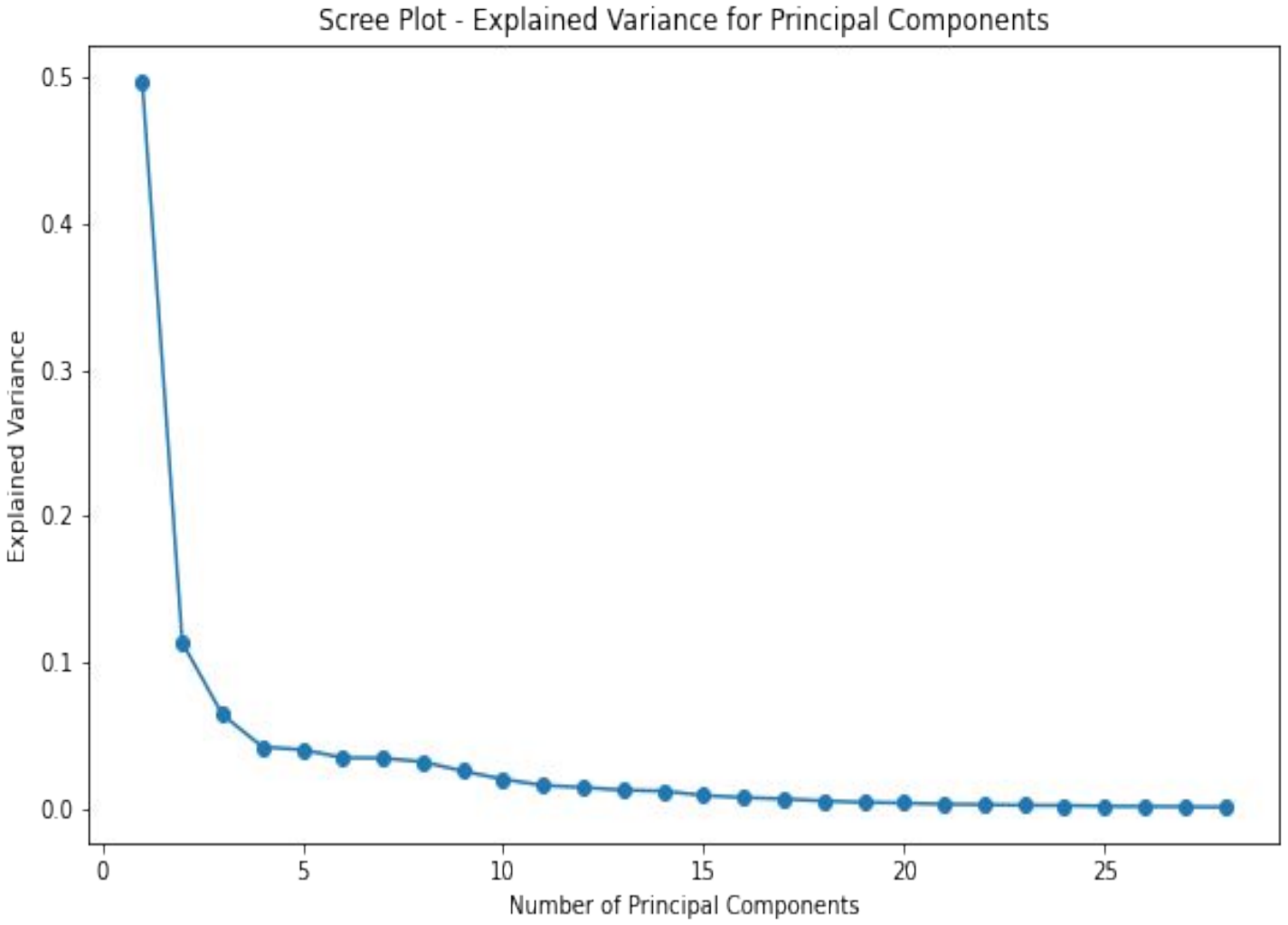
- Many non clusters
- A very large cluster

Now on the map

- US now with Canada, Northern Europe, Japan, New Zealand and Australia
- Huge variety for blue



Principle Component Analysis



- Elbow at four principal components
- Achieved a score of 0.73
- Using max amount of components only yields 0.83

Conclusions

- Corruption is not improving globally with the current trends!
- The three most heavily weighted features
  - were avrg_daily_income, democracy_index, fertility_rate.
- Features are geographically clustered as well meaning countries could impact each other
- **Boosted RandomForest** scored highest (0.981).