

Modeling European E-Waste Recycling Rates (2010–2020)

1. Problem Definition

Electronic waste (e-waste) is one of the fastest-growing waste streams globally. For European policymakers and environmental agencies, understanding which countries lead or lag in e-waste recycling — and why — is critical for designing effective interventions, targeting support, and tracking progress against EU policy goals.

This project focuses on:

- **Describing** how e-waste recycling rates have evolved across European countries.
- **Comparing** performance between countries and over time.
- **Modeling** country-level recycling rates using socioeconomic context and country identity, to understand which factors explain variation in performance.

The analysis is framed as if the client were a European environmental policy unit (e.g., an EU Directorate-General, regional NGO, or research group) interested in benchmarking countries, identifying patterns, and prioritizing where to investigate system design and policy.

Objective

Build an analysis-ready panel dataset of European e-waste recycling rates, perform exploratory data analysis (EDA), and develop regression models that predict each country's **e-waste recycling rate (%)** using:

- **year**
- **population**
- **log GDP per capita**
- **country indicators (dummy variables)**

The overall goal is **insight** rather than pure forecasting: explain patterns, test relationships (e.g., with GDP), and evaluate whether simple models can reasonably capture cross-country and temporal variation.

2. Data Sources and Wrangling

2.1 Data Sources

Two main public sources were used:

1. Our World in Data (OWID)

- Indicator: “12.5.1 – Proportion of electronic waste recycled (%)”
- Columns standardized to:
 - country
 - country_code (ISO3)
 - year

- ewaste_recycling_rate_pct

2. World Bank World Development Indicators (WDI)

- Population: SP.POP.TOTL → population
- GDP (current US\$): NY.GDP.MKTP.CD → gdp_current_usd
- Pulled via JSON API for years **2010–2022** and converted into tidy long format: country, country_code, year, value.

2.2 Geographic Scope

A broad list of **European ISO3 country codes** was defined (EU-27, EEA/EFTA, UK, Balkans, Eastern Europe, and microstates). The OWID and WDI tables were filtered to this list and restricted to rows with valid ISO3 codes.

2.3 Merging and Cleaning

Steps:

1. **Filter** OWID to European countries → eu_owid_recycling_rate.csv.
2. **Filter** WDI population and GDP to the same ISO3 list → eu_wdi_population.csv, eu_wdi_gdp.csv.
3. **Merge** on (country_code, year) to create a single panel:
 - Columns: country, country_code, year, ewaste_recycling_rate_pct, population, gdp_current_usd.
4. **Drop duplicates** and restrict to an aligned window focused on modern reporting:
 - Core analysis window: **2010–2020** (depending on availability per country).
5. **Handle missingness**
 - Rows with missing ewaste_recycling_rate_pct were dropped, since this is the target variable.
 - Population and GDP are complete within the retained rows in the final cleaned panel.

Key cleaned artifacts:

- data_clean/eu_panel_clean.csv
- data_clean/eu_country_year_coverage.csv
- data_clean/eu_recycling_rate_timeseries.csv

3. Exploratory Data Analysis (EDA)

3.1 Schema and Coverage

The cleaned panel (eu_panel_clean.csv) contains:

- **Rows:** 271 country-year observations
- **Columns:**
 - country
 - country_code
 - year (2010–2020)
 - ewaste_recycling_rate_pct
 - population
 - gdp_current_usd

Key facts:

- **31 European countries** in the panel.
- Year coverage varies by country. A coverage table (n_years per country) was created to track how many years of data each country has, and a bar chart visualizes this.

3.2 Distribution of Recycling Rates

A histogram of ewaste_recycling_rate_pct shows:

- Most observations lie between **~75–90%**.
- The **median** recycling rate is **~83%**.
- Some values exceed 100%; these likely reflect reporting conventions (e.g., treatment of previously stockpiled e-waste).

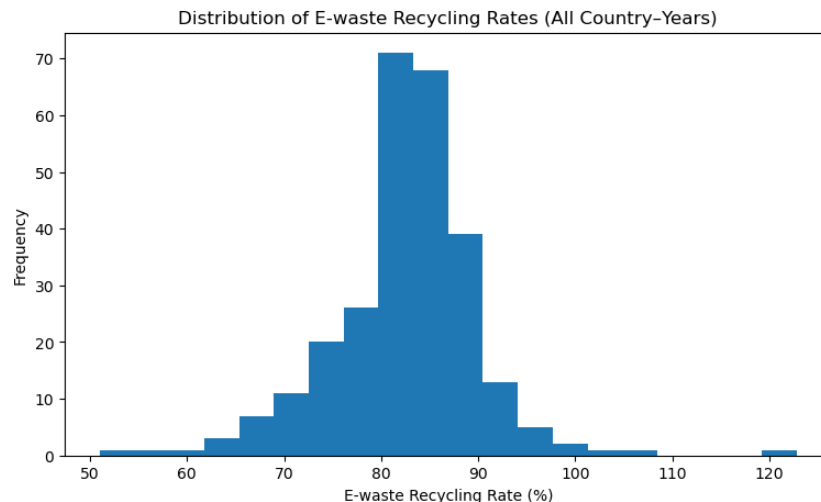


Figure 1. Distribution of E-waste Recycling Rates (All Country-Years).

Most observations fall between 75% and 90%, with a few outliers above 100%, likely due to reporting or classification differences. The distribution is slightly left-skewed, indicating many countries achieve relatively high recycling performance.

3.3 EU-Level Trend: Mean vs Population-Weighted

Using data_clean/eu_recycling_rate_timeseries.csv, two EU-level series were computed:

- **Simple mean:** each country counts equally.
- **Population-weighted mean:** larger populations have more weight.

Observations:

- From **2010 to ~2018**, both series stay relatively stable around **80–85%**.
- The **population-weighted rate is consistently below the simple mean**, indicating that larger countries (e.g., Germany, France, Italy, Poland) tend to recycle **slightly less** than smaller, high-performing countries (e.g., Nordic states).
- There is a **sharp dip around 2019**, followed by a partial rebound in 2020. This is more plausibly a **reporting or coverage artifact** rather than a collapse of recycling systems, highlighting the importance of data completeness when interpreting time series.

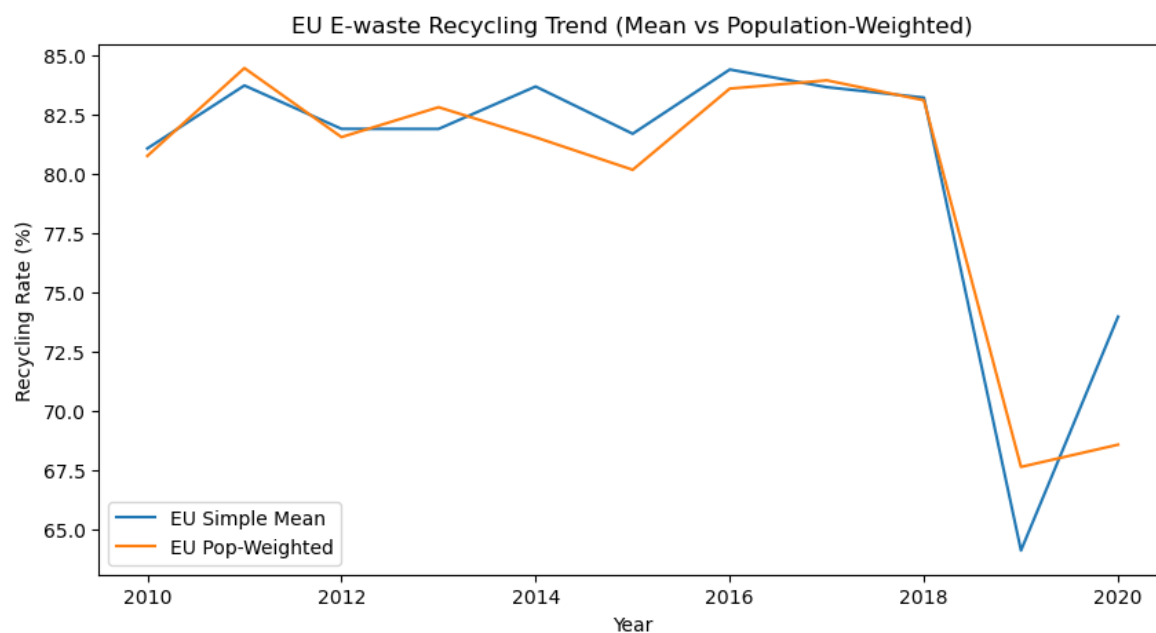


Figure 2. EU E-waste Recycling Trend (2010–2020).

3.4 Country Rankings and Trends

- A “**latest common year**” was chosen as **2018**, the last year with data for most countries.
- Countries were ranked by recycling rate in 2018 using a horizontal bar chart.

Patterns:

- **Leaders** (e.g., Austria, Finland, Croatia, Slovakia, Netherlands) maintain high rates (often **80–95%**) with relatively **small year-to-year variation**, indicating mature, stable systems.
- **Laggards** and more volatile reporters (e.g., Iceland, Luxembourg, Liechtenstein) show either:
 - Gradual improvement (Iceland), or
 - Large jumps and drops likely tied to reporting or small-population effects.

3.5 Relationship with GDP per Capita

A key question is whether **richer countries recycle more**.

Steps:

1. Created $\text{gdp_per_capita} = \text{gdp_current_usd} / \text{population}$.
2. Logged it: $\log_gdp_per_capita$ to reduce skew.
3. Plotted **scatter of recycling rate vs GDP per capita** (all country-years, log scale) and computed **Pearson correlations**:
 - All country-years: $r \approx -0.13$
 - Country means across years: $r \approx -0.25$

Interpretation:

- The relationship is **weak and slightly negative**.
- Wealth alone does **not** guarantee higher e-waste recycling performance.
- This supports the view that **policy design, collection infrastructure, and institutional maturity** are more important than GDP per capita.

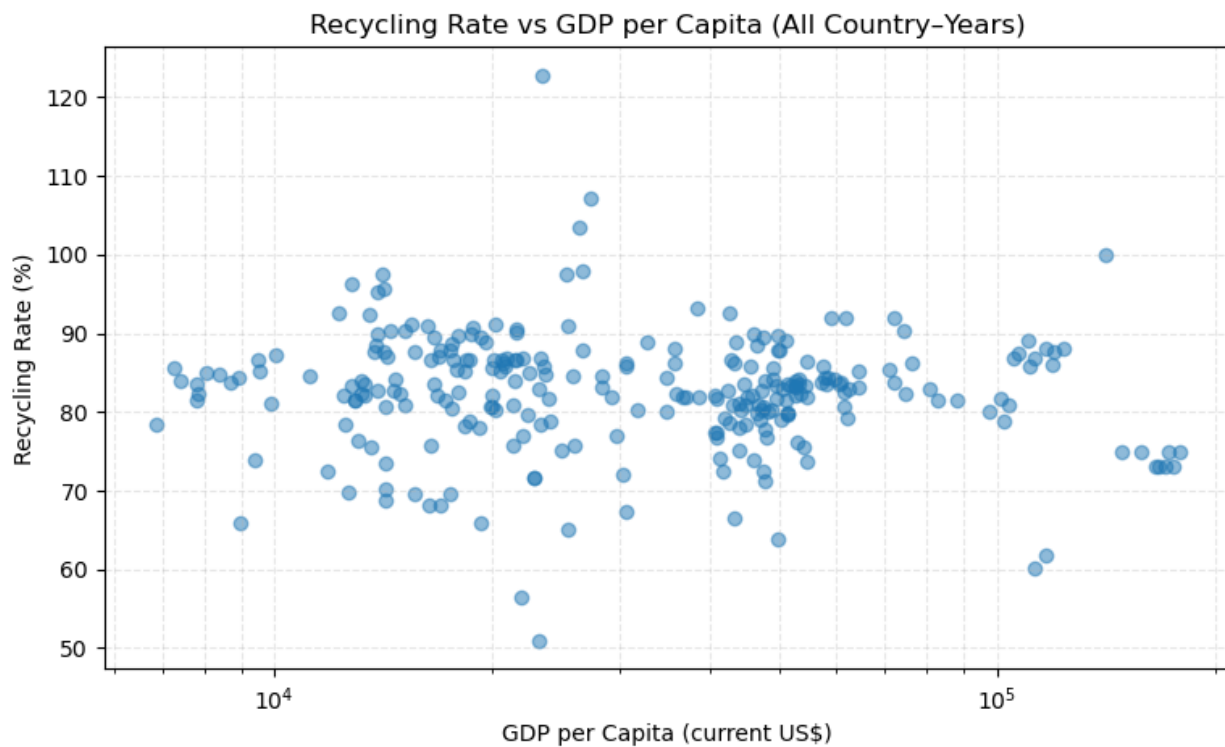


Figure 3. Recycling Rate vs GDP per Capita (All Country-Years).

There is no strong positive correlation between GDP per capita and recycling performance. Wealthier countries do not necessarily recycle more e-waste. This suggests that national policy design, collection systems, and reporting structures may be more influential than economic capacity alone.

4. Feature Engineering and Data Preparation for Modeling

4.1 Target and Features

- **Target variable:**
ewaste_recycling_rate_pct (continuous, percentage).
- **Numeric features:**
 - year
 - population
 - log_gdp_per_capita
- **Categorical feature:**
 - country_code (one-hot encoded)

Feature engineering:

- Generated gdp_per_capita and log_gdp_per_capita.
- Dropped rows with missing values in key fields.
- Final modeling dataset (model_df) contains **271 rows** and **8 columns** (including engineered features).

4.2 Train / Validation / Test Split

To evaluate generalization:

- **Total dataset:** 271 observations.
- Split into:
 - **Train:** 60%
 - **Validation:** 20%
 - **Test:** 20%

Using train_test_split with random_state=42:

- Train: 162 rows
- Validation: 54 rows
- Test: 55 rows

Histograms confirm that the target distribution is similar across splits.

4.3 Preprocessing Pipeline

A ColumnTransformer and Pipeline were used:

- **Numeric** (year, population, log_gdp_per_capita):
 - Standardized via StandardScaler (mean 0, std 1 based on **training set only**).
- **Categorical** (country_code):

- Encoded using OneHotEncoder(drop="first", handle_unknown="ignore") to avoid full dummy trap and gracefully handle unseen categories.

The fitted preprocessing pipeline was saved as:

- data_model/preprocessor.pkl

Transformed shapes (after preprocessing):

- X_train_proc: (162, 33) features
- X_val_proc: (54, 33)
- X_test_proc: (55, 33)

5. Modeling and Evaluation

Three models were evaluated:

1. **Baseline mean regressor**
2. **Linear Regression**
3. **Random Forest Regressor** (with hyperparameter tuning)

Performance was assessed using:

- **MAE** (Mean Absolute Error)
- **RMSE** (Root Mean Squared Error)
- **R²** (coefficient of determination)

5.1 Baseline – Mean Regressor

A trivial model that always predicts the **training-set mean** recycling rate.

- **Train:**
 - MAE \approx 5.38
 - RMSE \approx 7.84
 - R² = 0.00
- **Validation:**
 - MAE \approx 4.46
 - RMSE \approx 6.16
 - R² \approx -0.00

This provides a **floor** that real models should beat.

5.2 Linear Regression

A standard LinearRegression model was fit on the processed features.

- **Train:**

- MAE ≈ 3.53
- RMSE ≈ 5.86
- $R^2 \approx 0.44$
- **Validation:**
 - MAE ≈ 4.34
 - RMSE ≈ 6.45
 - $R^2 \approx -0.10$

Interpretation:

- The model fits the training data moderately well but fails to generalize (negative validation R^2).
- This suggests that **recycling patterns are not well captured by a simple linear relationship** between the standardized features and the target.

5.3 Random Forest Regressor (Tuned)

A RandomForestRegressor was tuned with GridSearchCV (5-fold CV) using the following grid:

- n_estimators: [100, 200]
- max_depth: [None, 3, 5, 7]
- min_samples_split: [2, 5]
- min_samples_leaf: [1, 2]

Best parameters:

- n_estimators = 100
- max_depth = None
- min_samples_split = 5
- min_samples_leaf = 1
- random_state = 42

Validation performance of best RF model:

- **Train:**
 - MAE ≈ 2.33
 - RMSE ≈ 4.26
 - $R^2 \approx 0.70$
- **Validation:**
 - MAE ≈ 3.68
 - RMSE ≈ 5.32
 - $R^2 \approx 0.25$

This is the **only model with positive validation R^2** , and it clearly improves on the baseline and Linear Regression in terms of MAE and RMSE.

5.4 Final Model and Test Performance

The tuned Random Forest was selected as the **final model**.

Procedure:

1. Concatenated **train + validation** sets.
2. Refit the final Random Forest on the combined training/validation data.
3. Evaluated on the **held-out test set**.

Test performance:

- **Test MAE:** ≈ 5.46
- **Test RMSE:** ≈ 9.57
- **Test R²:** ≈ -0.46

Interpretation:

- Negative test R² indicates the model does **not generalize well** to this small, heterogeneous test set.
- This is not surprising given:
 - The very small dataset (271 total rows).
 - Strong **country-specific effects** and reporting quirks.
 - Limited predictor set (no explicit policy variables, collection system types, etc.).

The Random Forest is still the **best model within this project scope**, and it helps quantify which features are most informative.

5.5 Feature Importance

Random Forest feature importances indicate:

Top features (approximate importances):

- population: ~ 0.26
- year: ~ 0.18
- log_gdp_per_capita: ~ 0.11
- Country dummies
(e.g., country_code_MLT, country_code_HRV, country_code_LTU, country_code_SVN):
collectively large importance.

Takeaways:

- **Population** and **year** are key numeric predictors.
- **GDP per capita** plays a secondary role.
- Several **country codes** are highly influential, reinforcing that **national systems, policies, and reporting practices** drive much of the variation.

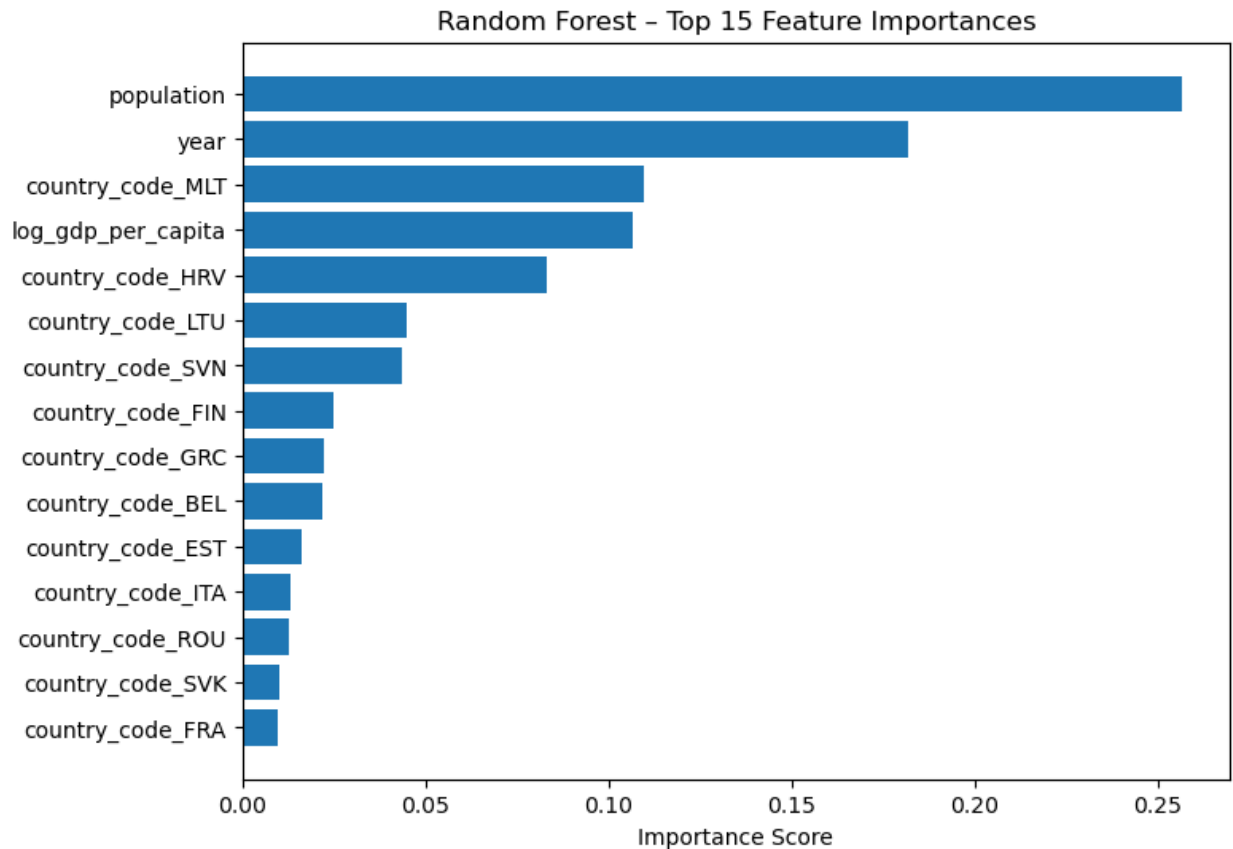


Figure 4. Random Forest – Top 15 Feature Importances.

Population and year are the strongest predictors of recycling performance, followed by several country-specific indicators. This suggests that national structural factors — not just economic variables — play a dominant role in shaping e-waste recycling rates across Europe.

6. Key Insights and Recommendations

6.1 Insights

1. Stable but Plateaued EU-Level Performance

- EU e-waste recycling rates are relatively high (around 80–85%) and stable from 2010–2018.
- There is no strong upward trend, suggesting that many countries have reached a plateau rather than continuing to improve.

2. Large Countries Lag Behind Smaller High Performers

- The population-weighted EU rate is consistently lower than the simple mean.
- This indicates that **larger countries tend to perform slightly worse** than many smaller, often Nordic or Western European countries.

3. **Wealth \neq High Recycling**

- The weak negative correlation between recycling rate and GDP per capita shows that **economic wealth alone does not predict better e-waste recycling**.
- Institutional design, enforcement, and system maturity are more important.

4. **Country Identity Matters**

- Country dummy variables are among the most important predictors in the Random Forest, reflecting the importance of **country-specific systems, regulations, and behaviors**.

6.2 Recommendations for the Client

Recommendation 1 – Focus on System Design Over Income Level

Do not assume that increasing GDP or income will automatically improve e-waste recycling. Instead, prioritize **policy and system benchmarking**: study high-performing smaller countries (e.g., Nordic countries, Austria, Netherlands) and identify replicable aspects of their collection schemes, producer responsibility systems, and enforcement mechanisms.

Recommendation 2 – Target Interventions in Large, Underperforming Countries

Because larger countries pull the population-weighted EU rate down, improving performance in these countries can have a **disproportionately large impact** on EU-level outcomes. Use this analysis to flag large countries with middling or declining rates and prioritize additional investigation, support, or pilot programs there.

Recommendation 3 – Improve Data Consistency and Coverage

The dip around 2019 and variation in coverage suggest that **data gaps and reporting changes** can obscure real trends. Investing in harmonized reporting standards, consistent year-to-year coverage, and transparent methodology would make evaluation and monitoring more robust and reduce misinterpretation of apparent “shocks” driven by data artifacts.

7. Limitations and Future Work

Limitations

- **Small sample size:** Only 271 country-year observations and 31 countries.
- **Limited feature set:** Lacks direct policy variables (e.g., WEEE Directive implementation details, collection targets, enforcement intensity).
- **Data quality & coverage issues:** Some years show sudden changes likely driven by reporting or classification changes.
- **Random Forest generalization:** Negative test R^2 highlights that current features are not sufficient to robustly predict unseen observations.

Future Work

- Incorporate **policy and system design features** (e.g., presence of deposit-refund systems, number of collection points, producer responsibility organization maturity).

- Use **time-series cross-validation** or country-level rolling windows to better reflect forecasting scenarios.
- Group countries into **regional or policy clusters** (Nordic, Western, Eastern, Southern Europe) to analyze model performance and effects within more homogeneous groups.
- Combine e-waste quantities (kg per capita) with recycling rates to measure **absolute volumes recovered vs generated**.

8. Conclusion

This project builds a European e-waste recycling panel dataset, explores differences across countries and over time, and evaluates regression models that attempt to explain recycling performance using socioeconomic features and country identity. While predictive performance on the test set is limited, the analysis clearly shows that:

- EU-level performance is high but largely flat.
- Large countries lag behind small high performers.
- Economic wealth is not the main driver of success.
- Country-specific systems and policies play a dominant role.

These findings provide a useful foundation for further policy-focused research and a roadmap for improving both data and models in future work.