Codebook

Santiago Taborga

2025-04-20

Overview of Data

This dataset consolidates multiple indicators across a variety of domains related to global development and infrastructure. The data spans multiple years (1960 to 2024) and includes a wide range of variables that capture technological, economic, environmental, and governance-related aspects of countries around the world. The dataset is organized by country and year, allowing for longitudinal analysis and cross-country comparisons.

Key Features:

- *Geographic Scope*: The dataset covers over 200 countries and territories, with geographic regions classified according to Kaggle's "Countries of the World" regional definitions.
- *Timeframe*: Data spans from 1960 to 2024, offering a long-term view of trends in internet penetration, mobile connectivity, trade, logistics performance, renewable energy adoption, and governance. Variable-based limitations arise due to the availability of published data.
- *Variables*: The dataset includes several variables, each representing a distinct facet of development, such as the share of the population using the internet, the mobile connectivity index, trade openness (as a percentage of GDP), the Logistics Performance Index (LPI), the share of modern renewable energy, and governance indicators.

The data is structured in a tidy format, with each row representing a unique observation for a given country and year, and each column representing a specific variable. Harmonization efforts were undertaken to standardize country names and regions across datasets. Missing values were handled thoughtfully, and NA values are explicitly coded as NA to maintain data integrity.

By merging these diverse datasets, this project aims to provide a comprehensive picture of how global indicators related to technology, trade, logistics, energy, and governance evolve over time and across regions.

Sources and Methodology

The data for this project was collected from a variety of reputable sources, including the International Telecommunication Union, the GSMA, the World Bank, and the International Energy Agency. Each dataset was merged into a unified panel with consistent identifiers (`country`, `year`) and cleaned for uniformity.

Source	**Variables**	**Notes**
International Telecommunication Union	Individuals using the Internet (% of population)	Reshaped to long format, retaining only country-level annual estimates.
GSMA Mobile Connectivity Index	mci_index, mci_infrastructure, mci_affordability, mci_cons_readiness, mci_content_services	Numeric indicators of mobile readiness across infrastructure, affordability, and content access.
World Bank (Trade)	Trade as % of GDP (trad_GDP)	Long format reshaping applied; harmonized country names.
World Bank (Governance)	Political stability indicators (psi_*)	Wide-format reshaped and renamed with psi_ prefix for clarity.
World Bank LPI	lpi_score	Combined sheets into a single panel; harmonized names and numeric years.
International Energy Agency	ren_energy	Cleaned and converted non-numeric entries ('..') to NA.
Kaggle	region, area, coastline	Region harmonization and gap-filling completed with a custom mapping.

All country names were harmonized using a custom recoding function to ensure uniform joins across sources. NA values were explicitly handled and coded using `NA` throughout.

Itemized Variable Descriptions

This section provides a reference guide to each variable in the dataset.

`country` *(character)*

Name of the country or territory. Harmonized to standardized names.

`region` *(character)*

Broad geographic region.

Asia (except Near East)
Eastern Europe
Northern Africa
Oceania
Western Europe
Sub-Saharan Africa
Latin America and Caribbean
Commonwealth of Independent States
Near East

Northern America
Baltics

year *(character)*

Year of the observation or data record ranging from 1960-2024.

int_pen *(numeric)*

Internet penetration rate – the percentage of individuals using the internet in a given country. This indicator is measured as a share of the total population, based on surveys and administrative data collected by the International Telecommunication Union.

Count	Min	Mean	Median	Max	Coverage
4949	0	39.41	33.82	100	2000—2024

mci_index *(numeric)*

Mobile Connectivity Index (MCI) – overall score, ranging from 0 to 100. This composite score measures the capacity of a country to support mobile internet adoption and usage, based on enablers like infrastructure, affordability, consumer readiness, and availability of content/services. A higher score indicates a stronger mobile internet ecosystem.

Count	Min	Mean	Median	Max	Coverage
1740	7.18	57.75	59.2	93.72	2014—2023

mci_infrastructure *(numeric)*

Infrastructure sub-index of the MCI, ranging from 0 to 100. This reflects the availability and quality of mobile network infrastructure (e.g., 3G/4G coverage, spectrum availability, international bandwidth).

Count	Min	Mean	Median	Max	Coverage
1740	12.43	57.4	58.96	98.78	2014—2023

mci_affordability *(numeric)*

Affordability sub-index of the MCI, ranging from 0 to 100. This captures the cost of mobile services relative to income, including handset and mobile data affordability.

Count	Min	Mean	Median	Max	Coverage
1740	1.36	56	56.28	100	2014—2023

`mci_cons_readiness` *(numeric)*

Consumer readiness sub-index of the MCI, ranging from 0 to 100. This reflects users' ability and willingness to use mobile internet, based on literacy levels, digital skills, gender equality, and device ownership.

Count	Min	Mean	Median	Max	Coverage
1740	11.23	67.43	73.28	96.33	2014—2023

`mci_content_services` *(numeric)*

Content and services sub-index of the MCI, ranging from 0 to 100. This measures the availability of relevant, local, and accessible content and services in local languages, including e-government, mobile apps, and social platforms.

Count	Min	Mean	Median	Max	Coverage
1740	2.92	53.86	54.77	94.4	2014—2023

`trad_GDP` *(numeric)*

Trade (% of GDP) – calculated as the sum of exports and imports of goods and services measured as a share of gross domestic product. Higher values indicate greater economic openness or dependence on international trade.

Count	Min	Mean	Median	Max	Coverage
8838	0.02	78.31	67.23	863.2	1960—2023

`lpi_score` *(numeric)*

Logistics Performance Index (LPI) – a score ranging from 1 to 5 that measures the quality of logistics services in a country, including customs procedures, infrastructure, shipment tracking, and timeliness. Higher scores reflect more efficient trade logistics environments.

Count	Min	Mean	Median	Max	Coverage
1092	1.21	2.89	2.75	4.3	2007—2023

`ren_energy` *(numeric)*

Modern renewable energy share – the percentage of total final energy consumption derived from modern renewable sources (e.g., solar, wind, biofuels, excluding traditional biomass). Based on data from the International Energy Agency.

Count	Min	Mean	Median	Max	Coverage
7064	0	10.43	5.25	82.79	1990—2021

`psi_cc` *(numeric)*

Governance indicator: Control of Corruption, ranging from approximately -2.5 (weak) to +2.5 (strong). Higher values reflect better control of corruption, defined as the extent to which public power is exercised for private gain.

Count	Min	Mean	Median	Max	Coverage
5153	-1.97	0.01	-0.2	2.46	1996—2023

`psi_ge` *(numeric)*

Governance indicator: Government Effectiveness, ranging from -2.5 to 2.5. This captures the quality of public services, civil service, and the credibility of government policy implementation.

Count	Min	Mean	Median	Max	Coverage
5129	-2.44	0.01	-0.12	2.47	1996—2023

`psi_pv` *(numeric)*

Governance indicator: Political Stability and Absence of Violence, ranging from -2.5 to 2.5. Higher values represent more politically stable environments with lower likelihood of violence, terrorism, or government instability.

Count	Min	Mean	Median	Max	Coverage
5188	-3.31	0.02	0.12	1.96	1996—2023

`psi_rl` *(numeric)*

Governance indicator: Rule of Law, ranging from -2.5 to 2.5. Reflects the extent to which agents have confidence in and abide by the rules of society, including property rights, judicial independence, and contract enforcement.

Count	Min	Mean	Median	Max	Coverage
5243	-2.59	0.01	-0.12	2.12	1996—2023

`psi_rq` *(numeric)*

Governance indicator: Regulatory Quality, ranging from -2.5 to 2.5. Measures the ability of the government to formulate and implement sound policies and regulations that promote private sector development.

Count	Min	Mean	Median	Max	Coverage
5131	-2.55	0.01	-0.1	2.31	1996—2023

`psi_va` *(numeric)*

Governance indicator: Voice and Accountability, ranging from -2.5 to 2.5. Reflects citizens' ability to participate in selecting their government, as well as freedom of expression, freedom of association, and a free media.

Count	Min	Mean	Median	Max	Coverage
5190	-2.31	0.01	0.06	1.8	1996—2023

`area` *(numeric)*

Total land area of the country or territory, measured in square kilometers. Represents the entire surface area including inland water bodies.

Count	Min	Mean	Median	Max	Coverage
13893	2	620613.6	92391	17075200	1960—2024

`coastline` *(numeric)*

Coastline-to-area ratio, calculated as coastline length divided by total land area (in miles per square mile). This metric is used to indicate geographic exposure to maritime environments, often linked to trade potential.

Count	Min	Mean	Median	Max	Coverage
13893	0	20.99	0.71	870.66	1960—2024

Data Sources

- GSMA. (2023). Mobile Connectivity Index. https://www.mobileconnectivityindex.com/assets/excelData/MCI_Data_2024.xlsx

- International Energy Agency. (2024). Share of Modern Renewables in Final Energy Consumption. https://www.iea.org/data-and-statistics

- International Telecommunication Union. (2025). ITU DataHub. https://datahub.itu.int/indicators/

- Lasso, F. (2018). Countries of the World [Dataset]. Kaggle. https://www.kaggle.com/datasets/fernandol/countries-of-the-world

- World Bank. (2023). Logistics Performance Index (2007–2023). https://lpi.worldbank.org/

- World Bank. (2023). Worldwide Governance Indicators. https://www.worldbank.org/en/publication/worldwide-governance-indicators

- World Bank. (2024). World Development Indicators – Trade. https://api.worldbank.org/