# **Dataset – Energy Data**

Owid-energy-data.csv → energy.csv

## **Selection of Variables**

In total, 40 variables were selected from the following areas:

· Demographics (5)

· Electricity Generation (15)

· Fossil Fuel Production (3)

· Energy Consumption (13)

· Energy Consumption Share (12)

· Other Metrics (4)

Aggregated variables are removed (e.g. low\_carbon, fossil, renewable) and teams may choose to aggregate these themselves as mentioned in the case brief. (-8)

## **Columns relocated for readability**

· Relocating Coal / Oil / Gas in same positions

## **Row Filtering**

Countries were selected on a **missing value count** basis:

· Countries that contained more than 3 missing fields across all relevant columns in the year 2000 were *excluded.*

· Energy data contained an aggregated “country” called **World**. This was removed.

· Due to missing data for **Taiwan** and **Venezuela** in the other datasets, these countries were also excluded despite having less than or equal to 3 missing fields.

Year Range: 2000 – 2016

## **Imputation / Cleaning of Missing Values**

· Fossil Fuel Production Metrics (e.g. coal production) had NA values transformed to 0, indicating that the particular country did not produce the fossil fuel. If consumption of the fossil fuel was non-zero, it likely indicates the fossil fuel was imported.

# **Dataset – WGI**

Wgidataset.xlsx → WGI.csv

**Recommended to not modify the cleaning for this dataset** (only estimate values should be used)

· Extracted only **estimate** values from WGI dataset in the specified year range.

· Column names were renamed to match the energy dataset and allow for ease in merging.

· Countries were renamed to match countries in energy dataset:

o "Czech Republic", "Czechia"

o "Slovak Republic", "Slovakia"

o "Egypt, Arab Rep.", "Egypt"

o "Hong Kong SAR, China", "Hong Kong"

o "Iran, Islamic Rep.", "Iran"

o "Russian Federation", "Russia"

o "Korea, Rep.", "South Korea"

o "Taiwan, China", "Taiwan"

o "Venezuela, RB", "Venezuela"

· Found incorrect country code which led to merging not working as intended:

o For Romania - "ROM", "ROU"

## **Imputation / Cleaning of Missing Values**

· Missing WGI indicator values for 2001. Values for every country were imputed using **Moving Average** of the preceding and subsequent terms (i.e. 2000 and 2002).

# **Exported Datasets**

There are three datasets:

· energy.csv (1292 x 40): Contains all energy metrics + GDP and **country, iso\_code** and **year**.

· WGI.csv (1292 x 9): Contains all 6 world governance indicators and **country, iso\_code** and **year.**

· energy\_WGI.csv (1292 x 46): Merged dataset of energy and WGI by **country, iso\_code** and **year**. **NOT INCLUDED IN THE DATASETS PROVIDED** - but can easily be created through one line of code (See final\_cleaning.R if code needed).

R Code for cleaning is called final\_cleaning.R