1. Summary

The production of agriculture and food involves the emission of greenhouse gases, such as carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O), which cause climate change.

2. Data source and collection

Data was obtained from FAOSTAT – the statistics division of Food and Agriculture Organization of the United Nations (FAO). It's an international agency, so the data they provide is reliable. FAO also provides sources for the data they collect and the methodology of how some variables were calculated.

Datasets:

- **emissions dataset**

The dataset contains emissions of greenhouse gases by pre- and post-agriculture production phase and by country. It's a yearly data, that covers the period from 2002 to 2021.

Column	Description
Domain Code	Two-letter domain code.
Domain	Name of the domain - missions from pre and post-agriculture production.
Area Code (M49)	Standard numerical country code.
Area	Country name.
Element Code	Code for greenhouse gas emission.
Element	Name of the greenhouse gas emission.
Item Code	Code for the agriculture production stage.
Item	Name of pre- or post-agriculture production stage.
Year Code	Year.
Year	Year.
Value	Emissions amount, in kilotonnes.

- **temperature_change**

The dataset contains yearly changes in the land temperature by country. The temperature changes are calculated using the baseline period (1951-1980). The dataset covers the period from 2002 to 2021.

Column	Description
Domain Code	Two-letter domain code.
Domain	Name of the domain - temperature change on land.
Area Code (M49)	Standard numerical country code.
Area	Country name.
Element Code	Code for the area of interest.
Element	Name of the area of interest - temperature change.
Months Code	Code for an observation period.
Months	Observation period (meteorological year in this case).
Year Code	Year.
Year	Year.
Value	Temperature change, in degrees Celsius.

- **population**

The dataset contains yearly population counts by country, from 2002 to 2021.

Column	Description
Domain Code	Two-letter domain code.
Domain	Name of the domain – annual population.
Area Code (M49)	Standard numerical country code.
Area	Country name.
Element Code	Code for the area of interest.
Element	Name of the area of interest – total population – both sexes.
Item Code	Code for an item.
Item	Item name - population estimated and projected.
Year Code	Year.
Year	Year.
Value	Population counts, in thousands.

3. Data cleaning

- **Data wrangling**

greenhouse_gas_emissions	
Columns dropped	**Reason**
Domain Code Domain Element Code Item Code Year Code	Not needed for analysis
Columns renamed	**Reason**
Area Code (M49) -> country_code Area -> country, Element -> emission_gas Year -> year Value -> emissions Item -> agriculture_prod_stage	For consistency and clarity

temperature_change	
Columns dropped	**Reason**
Domain Code Domain Element Code Element Months Code Months Year Code	Not needed for analysis

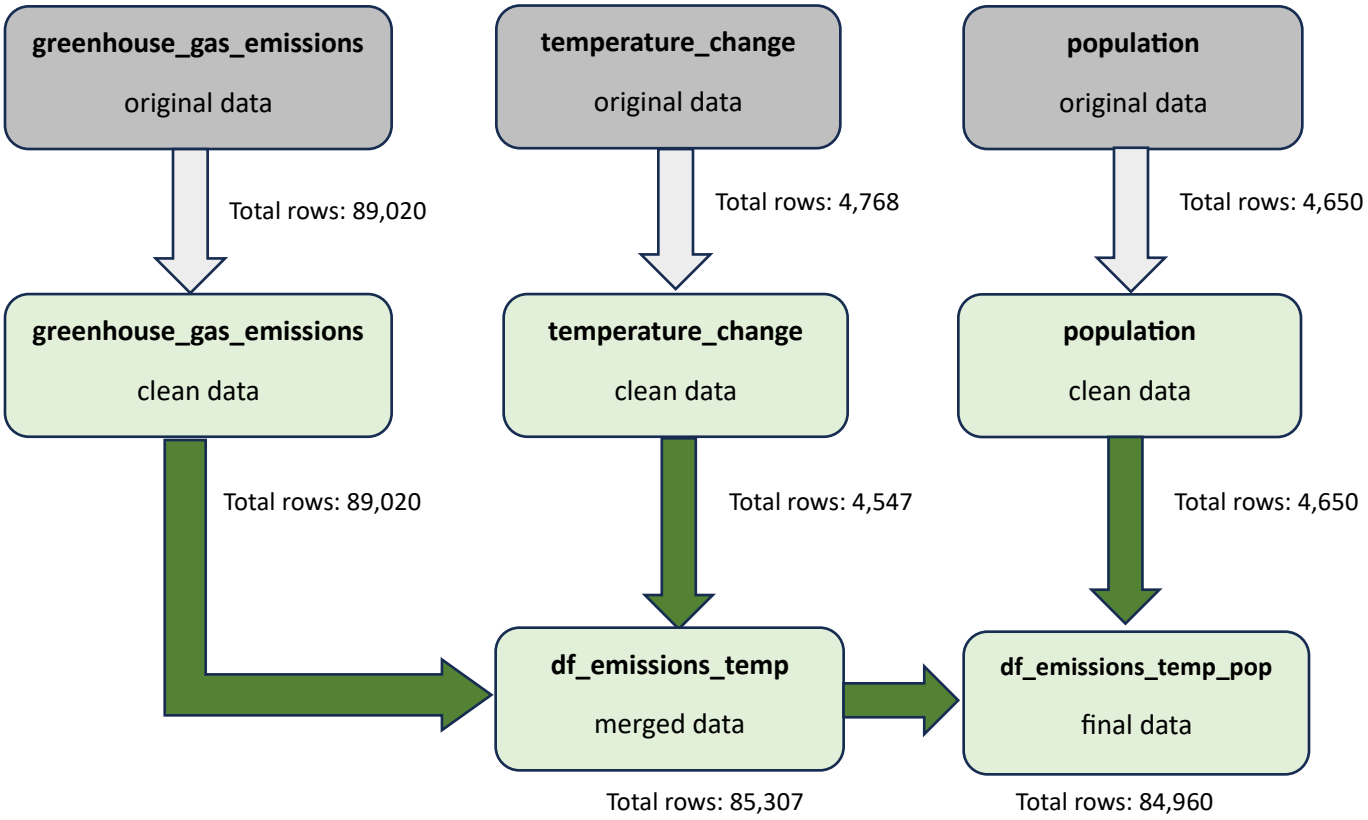
Columns renamed	Reason
Area Code (M49) -> country_code Area -> country Year -> year Value -> temp_change	For consistency and clarity

population	
Columns dropped	**Reason**
Domain Code Domain Element Code Element Item Code Item Year Code	Not needed for analysis
Columns renamed	**Reason**
Area Code (M49) -> country_code, Area -> country, Year -> year, Value -> total_population	For consistency and clarity

- **Consistency checks**

Dataset	Missing values	Missing values treatment	Duplicates
greenhouse_gas_emissions	No missing values	N/A	No duplicates
temperature_change	'temp_change' column had 221 missing values.	Removed rows with missing values (<5% of the dataset)	No duplicates
population	No missing values	N/A	No duplicate

4. Population flow



All datasets were merged based on the key columns: country_code, country, and year.

5. Understand the data

Descriptive statistics:

	year	emissions	temp_change	total_population
count	84960.000000	84960.000000	84960.000000	8.496000e+04
mean	2011.531733	924.271448	1.106345	4.889179e+04
std	5.747401	9847.718329	0.534788	1.823632e+05
min	2002.000000	0.000000	-0.505000	5.110000e-01
25%	2007.000000	0.003663	0.736000	1.983465e+03
50%	2012.000000	0.155964	1.041000	8.046828e+03
75%	2017.000000	22.442834	1.413000	2.822518e+04
max	2021.000000	485989.998950	3.691000	1.457935e+06

The max value for emissions seems to be very large. Such a big difference in values exists because some countries produce more emissions than others during each agriculture stage. Also, some production activities create higher emissions of certain types of gases than others.

Column	Qualitative/Quantitative	Qualitative: nominal/ordinal Quantitative: discrete/continuous
country_code	qualitative	nominal
country	qualitative	nominal
emission_gas	qualitative	nominal
agriculture_prod_stage	qualitative	nominal
year	qualitative	ordinal
emissions	quantitative	continuous
temp_change	quantitative	continuous
total_population	quantitative	continuous

6. Limitations and Ethics

Certain stages of agricultural production have fewer records than others. It's not specified in the data source if countries didn't provide relevant data or if they didn't produce significant amounts of greenhouse gases in some years. Also, only one aspect of climate change is included in the dataset, which is the temperature change.

The dataset doesn't contain any PII. The data is representative in terms of the countries included.

7. Key questions

- How do the gas emissions change over time (total and by agriculture production stage)?
- What countries produce the highest/lowest amount of greenhouse gases?
- Can countries be classified based on the gas emissions level?
- Do countries with higher GDP have higher values of greenhouse gas emissions? (For this question will need to add data on GDP by country)
- How does the emissions amount differ based on the agriculture production stage?
- Is there a relationship between the country population and gas emissions?
- Do gas emissions affect the temperature change?