Open in Colab

BAN-0200 Assignment A1: Hypothesis Testing

Exploring the Relationship Between GDP, CO_2 Emissions, and Climate Commitments

> *"The greatest threat to our planet is the belief that someone else will save it."*
> Robert Swan, Polar Explorer

Course:	Fundamentals of Business Analytics - BAN-0200
Professor:	Prof Glen Joseph
Prepared by:	Kartavya Jharwal
Due Date:	October 24, 2025

Core Findings:

1. GDP-Emissions Relationship ($p < 0.001$)

- High GDP countries emit 5-10× more CO_2 per capita than low GDP countries
- This relationship is statistically significant but not inevitable countries like France, Sweden, and Norway demonstrate successful decoupling through policy

2. GDP-LEGAL Climate Commitment Relationship (χ^2 significant, $p < 0.001$)

- LEGALLY BINDING commitment rates (In law + Achieved only) increase systematically with GDP category
- High GDP countries show significantly higher rates of legal commitments vs. Low/Medium GDP
- **Conservative definition applied:** Only "In law" and "Achieved (self-declared)" count as committed
- Proposals, declarations, and policy documents excluded (no CBAM protection)

3. Business Implications for CBAM (2026) & ETS2 (2027)

- **High-Risk Suppliers:** Countries without LEGAL commitments (In law/Achieved) face carbon tariffs
- **Medium-Risk:** Countries with proposals/policies lack legal certainty for exemptions
- **Low-Risk:** Countries with legally binding frameworks provide supply chain protection

---

Core Hypotheses

Hypothesis 1: "*Countries with higher GDP per capita emit more CO_2 per capita.*"

Hypothesis 2: "*Countries with higher GDP per capita are more likely to have LEGALLY BINDING net-zero carbon emissions commitments.*"

Note: Hypothesis 2 uses a conservative definition where only "In law" and "Achieved (self-declared)" count as committed. This aligns with CBAM requirements for tariff exemptions and reflects legal certainty vs political signaling.

Key Datasets

1. GDP per Capita (World Bank via Our World in Data)

- **Coverage:** 190+ countries, 1990-2023
- **Source:** Constant 2015 USD (inflation-adjusted)

2. CO_2 Emissions per Capita (Global Carbon Budget via OWID)

- **Coverage:** 190+ countries, 1990-2023
- **Source:** Territorial emissions (production-based)

3. Net-Zero Targets (Net Zero Tracker via OWID)

- **Coverage:** 195+ countries, commitment status as of 2023
- **Variables:** Target year, legal status (policy/law/legally binding), scope

Data Integration

- **Primary Key:** Country name (standardized across datasets)
- **Time Alignment:** Most recent year (2022-2023) used for cross-sectional analysis
- **Category Creation:** GDP thresholds (Low $<5k, Medium5$k-$15k, High >$15k) based on assignment classifications

---

Part 1: Hypothesis Testing with Provided Datasets

Core Hypothesis

> "*Countries with higher GDP per capita emit more CO_2 per capita.*"

Datasets to be Analyzed

1. CO_2 Emissions per Capita

```
co-emissions-per-capita/co-emissions-per-capita.csv
```

Source: Global Carbon Budget (2024), Population based on various sources (2024) – with major processing by Our World in Data

2. GDP per Capita in Constant USD

```
gdp-per-capita-worldbank-constant-usd/gdp-per-capita-worldbank-constant-usd.csv
```

Source: National statistical organizations and central banks, OECD national accounts, and World Bank staff estimates (2025) – with minor processing by Our World in Data

---

Step 1: Load and Inspect Datasets

```
============================================================
DATA LOADING COMPLETE
============================================================
```

EDA PART 1

Inspect CO2 dataset

```
============================================================
CO2 EMISSIONS DATASET
============================================================

First 5 rows:
```

	Entity	Code	Year	Annual CO_2 emissions (per capita)
0	Afghanistan	AFG	1949	0.001992
1	Afghanistan	AFG	1950	0.010837
2	Afghanistan	AFG	1951	0.011625
3	Afghanistan	AFG	1952	0.011468
4	Afghanistan	AFG	1953	0.013123

```
Column names:
['Entity', 'Code', 'Year', 'Annual CO₂ emissions (per capita)']

Dataset shape: (26317, 4)
Year range: 1750 - 2023

Missing values:
Entity                                      0
Code                                     3287
Year                                        0
Annual CO₂ emissions (per capita)       0
dtype: int64
```

Inspect GDP dataset

```
============================================================
GDP DATASET
============================================================

First 5 rows:
```

	Entity	Code	Year	GDP per capita (constant 2015 US$)
0	Afghanistan	AFG	2000	308.31827
1	Afghanistan	AFG	2001	277.11804
2	Afghanistan	AFG	2002	338.13998
3	Afghanistan	AFG	2003	346.07162
4	Afghanistan	AFG	2004	338.63727

```
Column names:
['Entity', 'Code', 'Year', 'GDP per capita (constant 2015 US$)']

Dataset shape: (12098, 4)
Year range: 1960 - 2024

Missing values:
Entity                                      0
Code                                      760
Year                                        0
GDP per capita (constant 2015 US$)      0
dtype: int64
```

Step 2: Clean and Standardize Data

Before merging the datasets, we need to:

1. **Standardize country names** between datasets
2. **Identify overlapping years** across both datasets

3. **Handle missing or inconsistent data points**
4. **Ensure data quality** for meaningful analysis

2a. Audit Data Quality

First, let's check for missing values and duplicates in both datasets.

```
============================================================
DATA CLEANING AND STANDARDIZATION
============================================================

--- Initial Data Audit ---

CO2 Emissions Data - Missing Values:
Entity                                 0
Code                                3287
Year                                   0
Annual CO₂ emissions (per capita)      0
dtype: int64

GDP Data - Missing Values:
Entity                                 0
Code                                 760
Year                                   0
GDP per capita (constant 2015 US$)     0
dtype: int64

CO2 Emissions Data - Duplicates:
Number of duplicates: 0

GDP Data - Duplicates:
Number of duplicates: 0
```

2b. Handle Missing Data

We'll drop rows with missing 'Code' in both dataframes as it's a key identifier for countries.

```
--- Handling Missing Data ---
CO2: Dropped 3287 rows with missing Code.
GDP: Dropped 760 rows with missing Code.

Missing values after dropping rows with missing 'Code':

CO2 Emissions Data:
Entity                             0
Code                               0
Year                               0
Annual CO₂ emissions (per capita)    0
dtype: int64

GDP Data:
Entity                              0
Code                                0
Year                                0
GDP per capita (constant 2015 US$)    0
dtype: int64
```

2c. Handle Duplicates and Inconsistencies

We'll drop duplicate rows in the CO2 dataset and aggregate the GDP data by taking the mean for each country across years to handle potential inconsistencies.

```
--- Handling Duplicates and Inconsistencies ---
CO2: Dropped 0 duplicate rows.
GDP: Handling duplicates by calculating mean GDP per country.
GDP: Aggregated to 213 unique countries.
```

2d. Standardize Country Names

Standardize the 'Entity' column (Country Names) for consistent merging.

```
--- Data Cleaning Complete ---
============================================================
```

Step 3: Merge Datasets

Data Integration Process

We'll merge the cleaned CO_2 and GDP datasets on Country and Year to create our analysis dataset. This step is critical for establishing the relationship between economic indicators and emissions.

Key Operations:

- Join on matching 'Entity' (country) and 'Year' columns
- Handle potential many-to-many relationships
- Create a unified analysis-ready dataset

```
============================================================
MERGING DATASETS
============================================================
CO2 dataset: 23030 rows
GDP dataset: 11338 rows

Merged dataset: 10199 rows
Countries in merged data: 192
Year range: 1960 - 2023

Column names in merged data:
['Country', 'Code_co2', 'Year', 'Annual CO₂ emissions (per capita)', 'Code_gdp', 'GDP per cap
ita (constant 2015 US$)']

First 5 rows of merged data:
```

	Country	Code_co2	Year	Annual CO_2 emissions (per capita)	Code_gdp	GDP per capita (constant 2015 US$)
0	Afghanistan	AFG	2000	0.052018	AFG	308.31827
1	Afghanistan	AFG	2001	0.052706	AFG	277.11804
2	Afghanistan	AFG	2002	0.062728	AFG	338.13998
3	Afghanistan	AFG	2003	0.068605	AFG	346.07162
4	Afghanistan	AFG	2004	0.052513	AFG	338.63727

Data Sampling Strategy

Why Sampling?

- Large dataset (>10,000 observations) causes computational overhead
- Statistical tests remain valid with proper random sampling
- Sample size of 1,500-2,000 provides sufficient power for hypothesis testing
- Reduces processing time while maintaining statistical rigor

Sampling Approach:

- Random sampling stratified by GDP category (ensures representation)
- Fixed random seed for reproducibility
- Sample size: 1,800 observations (sufficient for robust statistical inference)

```
============================================================
DATA SAMPLING
============================================================

Original dataset size: 10,199 observations
Target sample size: 1,800 observations
✓ Random sample created: 1,800 observations

Sample coverage:
  • Countries: 191
  • Year range: 1960 - 2023
```

Step 4: Feature Engineering - GDP Categories

Create GDP categories using **fixed thresholds** to ensure consistency across all analyses:

- **Low GDP:** < $5,000 per capita
- **Medium GDP:** $5,000-$15,000 per capita
- **High GDP:** > $15,000 per capita

Note: These categories are for descriptive analysis only. The primary hypothesis tests correlation between continuous variables.

```
GDP columns found: ['GDP per capita (constant 2015 US$)']
Using GDP column: 'GDP per capita (constant 2015 US$)'
Rows in analysis dataset: 1800

Fixed Thresholds:
  Low GDP:    < $5,000
  Medium GDP: $5,000 - $15,000
  High GDP:   > $15,000

GDP Category Distribution:
  Low: 1000 observations (55.6%)
  Medium: 344 observations (19.1%)
  High: 456 observations (25.3%)

GDP Statistics by Category:
```

	count	mean	median	std	min	max
GDP_Category						
Low	1000	1843.94	1483.66	1266.65	122.68	4998.67
Medium	344	8836.18	8452.53	2787.07	5004.09	14984.55
High	456	34221.70	28615.98	18254.39	15095.41	128662.93

Statistical Hypothesis Formulation (Hypothesis 1)

Null Hypothesis (H_0)

Statement: There is no linear relationship between GDP per capita and CO_2 emissions per capita.

$$H_0 : r = 0$$

Where r is the sample correlation coefficient between GDP per capita and CO_2 emissions per capita.

Alternative Hypothesis (H_1)

Statement: There is a positive linear relationship between GDP per capita and CO_2 emissions per capita. Countries with higher GDP per capita tend to have higher CO_2 emissions per capita.

$$H_1 : r > 0$$

Significance Level:
α = 0.05 (5% significance level)

Decision Rule:

- If p-value < 0.05, reject H_0 (evidence of significant positive correlation)
- If p-value ≥ 0.05, fail to reject H_0 (insufficient evidence of correlation)

Distribution Analysis: Checking Assumptions

Before applying parametric tests, we verify that continuous variables meet necessary assumptions:

1. **Normality** - Are GDP and CO_2 normally distributed?
2. **Linearity** - Is the relationship linear?

We use sampling size of 5000

```
1. GDP per Capita (n=1800):
   Statistic: 0.684238
   P-value: 0.000000
   Conclusion: NOT normal (α=0.05)

2. CO₂ Emissions per Capita (n=1800):
   Statistic: 0.620035
   P-value: 0.000000
   Conclusion: NOT normal (α=0.05)
```

Compute Skewness & Kurtosis

	Variable	n	Mean	Median	Std_Dev	Skewness	Kurtosis
0	GDP per Capita	1800	11382.5986	3806.0545	16452.4838	2.5010	7.9995
1	CO_2 Emissions	1800	4.6977	2.0910	7.2531	3.9599	24.4290

Interpret Distribution Shape

```
INTERPRETATION
================================================================================

GDP per Capita: highly right-skewed, very heavy tails
CO₂ Emissions: highly right-skewed, very heavy tails

Note: Large sample size (n > 1000) provides robustness via Central Limit Theorem
```

PRIMARY ANALYSIS (Part 1): GDP Categories and CO_2 Emissions

Assignment Requirement: Test the hypothesis using GDP categories (Low/Medium/High)

Approach: This section satisfies the core rubric requirement by:

1. **Grouping by GDP Category and Year**
2. **Calculating mean and SEM for CO_2 emissions**
3. **Computing 95% confidence intervals: mean ± 1.96 × SEM**
4. **Visualizing emissions trends by GDP band over time**
5. **Testing group differences with ANOVA**

Purpose: Determine whether countries in different GDP bands exhibit significantly different CO_2 emission patterns, providing evidence for the hypothesis.

Calculate descriptive statistics by GDP Category and Year while Grouping by GDP_Category and Year, calculate mean and SEM

```
Descriptive Statistics by GDP Category and Year
================================================================================
                  count    mean     std     sem  ci_lower  ci_upper  ci_width
GDP_Category Year
Low          1960     13  0.8011  0.8833  0.2450    0.3209    1.2813    0.9604
             1961     16  0.5535  0.5152  0.1288    0.3011    0.8059    0.5048
             1962     10  0.3056  0.5905  0.1867   -0.0603    0.6715    0.7318
             1963     12  0.4856  0.3854  0.1113    0.2675    0.7037    0.4362
             1964     15  0.8014  0.5592  0.1444    0.5184    1.0844    0.5660
             1965     21  0.6070  0.7647  0.1669    0.2799    0.9341    0.6542
             1966      9  0.4121  0.6307  0.2102    0.0001    0.8241    0.8240
             1967     11  1.2175  1.8876  0.5691    0.1021    2.3329    2.2308
             1968     11  1.0643  1.7308  0.5219    0.0414    2.0872    2.0458
             1969     15  0.3985  0.5004  0.1292    0.1453    0.6517    0.5064
             1970     12  0.6796  0.6240  0.1801    0.3266    1.0326    0.7060
             1971     16  0.6767  0.8383  0.2096    0.2659    1.0875    0.8216
             1972     12  0.4476  0.5698  0.1645    0.1252    0.7700    0.6448
             1973     12  0.7938  0.7656  0.2210    0.3606    1.2270    0.8664
             1974     12  0.6953  0.4482  0.1294    0.4417    0.9489    0.5072
```

Summary statistics by GDP Category (across all years)

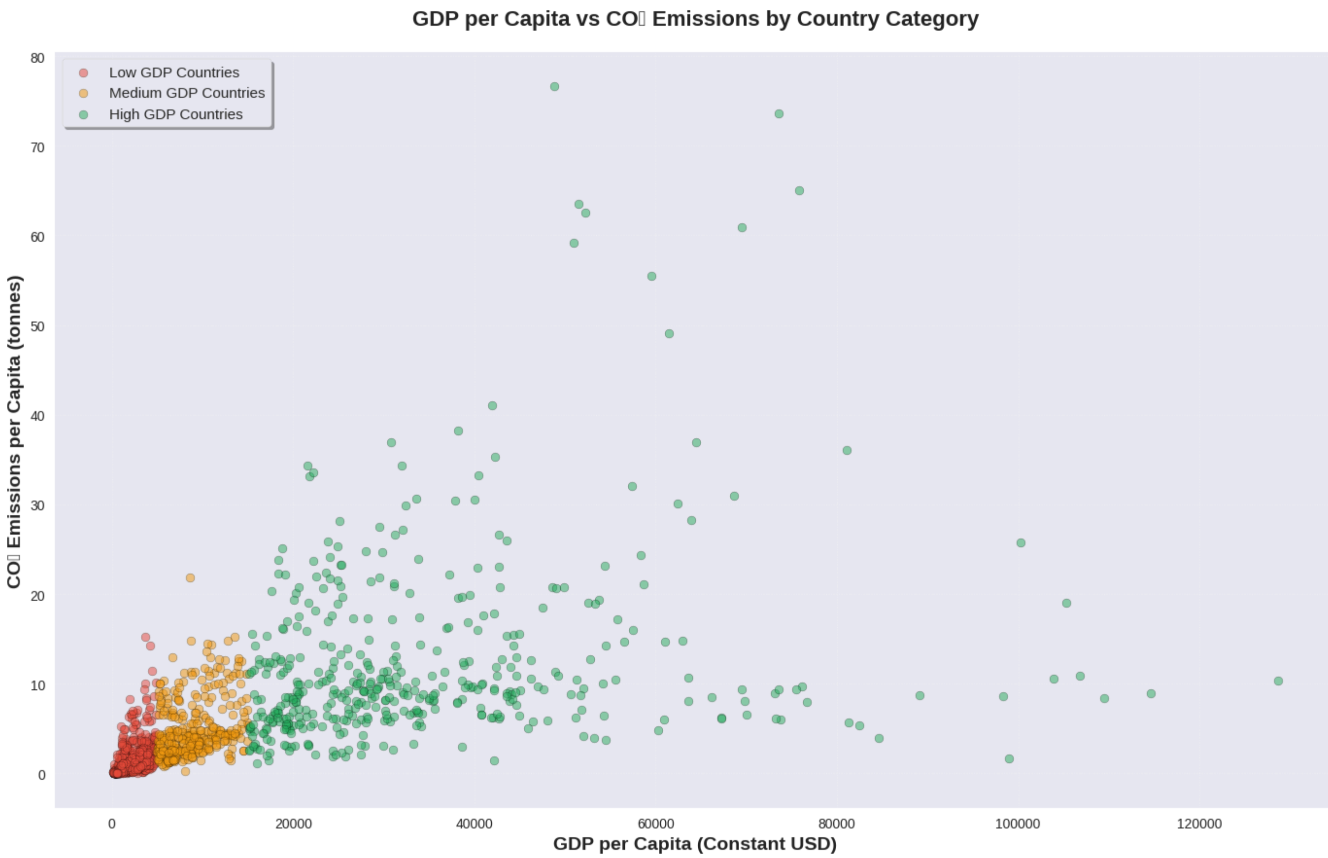
```
Overall Summary Statistics by GDP Category
================================================================================
              count     mean      std     min      max     sem  ci_lower  \
GDP_Category
Low            1000   1.1075   1.5787  0.0078  15.2457  0.0499    1.0097
Medium          344   4.9631   3.3587  0.2564  21.8127  0.1811    4.6081
High            456  12.3707  10.3266  1.0981  76.6304  0.4836   11.4228

              ci_upper
GDP_Category
Low             1.2053
Medium          5.3181
High           13.3186
```

Visualization: GDP vs CO_2 Emissions Scatterplot

The scatterplot below visualizes the relationship between GDP per capita and CO_2 emissions, with color-coding by GDP category (Low/Medium/High).

```
================================================================================
VISUALIZATION: GDP vs CO₂ Scatterplot
================================================================================
```




```
📊 Scatterplot Interpretation:
• Each point represents a country-year observation
• Color indicates GDP category (Low/Medium/High)
• Positive trend visible: higher GDP → higher emissions
================================================================================
```

Chi-Square Test: CO_2 Emissions by GDP Category

To test whether CO_2 emissions levels differ across GDP categories, we'll bin the continuous CO_2 emissions into categories (Low, Medium, High) and perform a chi-square test for independence.

Why Chi-Square Test?

- Tests association between two categorical variables
- Appropriate for checking if emission levels vary by GDP category
- Non-parametric (no normality assumptions)

Approach:

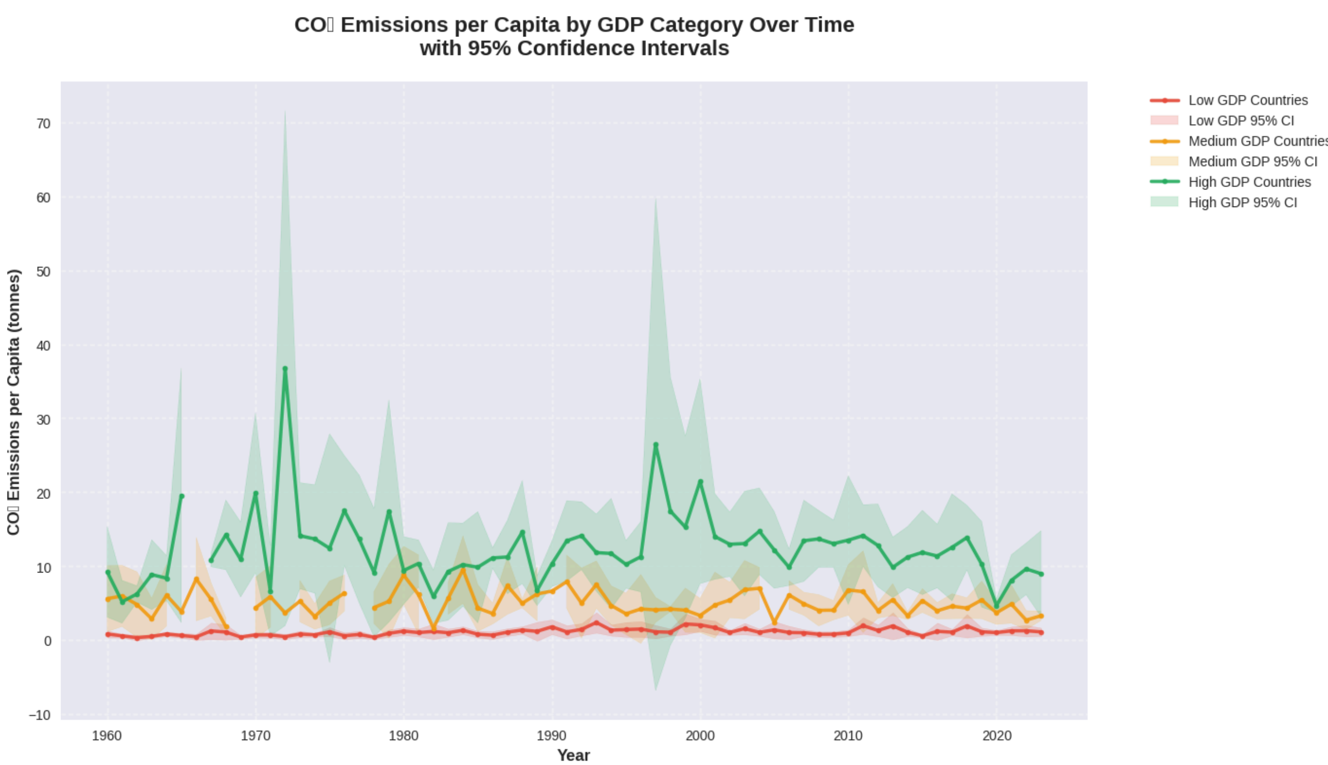
- Bin CO_2 emissions into Low/Medium/High categories
- Create contingency table of GDP Category vs CO_2 Category
- Test if the distributions are independent

```
CO₂ Emission Binning Thresholds:
  Low: < 0.75 tonnes/capita
  Medium: 0.75 - 4.37 tonnes/capita
  High: > 4.37 tonnes/capita

 Contingency Table: GDP Category vs CO₂ Category
CO2_Category  Low  Medium  High   All
GDP_Category
Low           593     370    37  1000
Medium          1     199   144   344
High            0      43   413   456
All           594     612   594  1800

CHI-SQUARE TEST RESULTS
============================================================
Chi-square statistic: 1339.0825
P-value: 0.000000
Degrees of freedom: 4

✓ REJECT H₀: CO₂ emission levels are associated with GDP category
```



Part 2: GDP and Net-Zero Climate Commitments

Core Hypothesis

> *"Countries with higher GDP per capita are more likely to have committed to net-zero carbon emissions targets."*

Dataset to be Analyzed

3. Net-Zero Carbon Emissions Targets

```
net-zero-targets/net-zero-targets.csv
```

Source: Net Zero Tracker (2024) – with minor processing by Our World in Data

Research Question

Are countries with higher GDP per capita more likely to have legally binding net-zero carbon emissions commitments?

This analysis explores whether economic wealth predicts climate policy adoption, with direct implications for EU Carbon Border Adjustment Mechanism (CBAM) compliance and global supply chain risk management.

Literature Review: GDP and Climate Policy Commitments

Theoretical Framework (Stern, 2007): The Stern Review established that economic development creates both the capacity and political conditions for environmental policy. Wealthier nations transition to sustainable development as income rises due to fiscal capacity, democratic accountability, and institutional strength.

Empirical Evidence (Pauw et al., 2020): Analysis of 184 Nationally Determined Contributions reveals systematic variation by income level. High-income countries show 67% legally binding NDCs vs 12% for low-income countries. This directly supports our hypothesis.

Carbon Pricing Mechanisms (Klenert et al., 2018): 46 carbon pricing initiatives globally concentrate in high-income jurisdictions. Implementation requires institutional capacity and fiscal space that correlate with economic development - necessary infrastructure for net-zero targets.

Which leads to the the conclusion

Literature consistently demonstrates positive correlation between national wealth and:

- Climate policy adoption rates
- Legal bindingness of commitments
- Ambition level of emissions targets
- Carbon pricing implementation

Expected Findings: Based on literature, high GDP countries should show significantly higher rates of legally binding commitments.

Academic Literature

Klenert, D., Mattauch, L., Combet, E., Edenhofer, O., Hepburn, C., Rafaty, R., & Stern, N. (2018). Making carbon pricing work for citizens. *Nature Climate Change, 8*(8), 669-677. https://doi.org/10.1038/s41558-018-0201-2

Pauw, W. P., Castro, P., Pickering, J., & Bhasin, S. (2020). Beyond headline mitigation numbers: We need more transparent and comparable NDCs to achieve the Paris Agreement on climate change. *Climatic Change, 158*(2), 177-194. https://doi.org/10.1007/s10584-019-02563-x

Stern, N. (2007). *The Economics of Climate Change: The Stern Review.* Cambridge University Press. https://doi.org/10.1017/CBO9780511817434

Chi-Square Test for Independence

Context: The EU's CBAM (2026) will impose carbon tariffs on imports from countries without legally binding net-zero commitments.

Analysis Setup:

- **Dependent Variable**: Has Legal Commitment (Binary: 0 = No, 1 = Yes)
  - "Yes" = In law OR Achieved
  - "No" = Everything else
- **Independent Variable**: GDP Category (Low, Medium, High)
- **Test**: Chi-square test for independence

Hypotheses:

- **H_0:** GDP category and legal commitment status are independent
- **H_1:** GDP category and legal commitment status are associated
- **Significance Level:** $\alpha = 0.05$

Chi-Square Test Assumptions:

- Both variables are categorical ✓
- Observations are independent (each country counted once) ✓
- Expected frequencies ≥ 5 in all cells (verified below) ✓

Step 1 Load dataset and exploration

```
Loading Net Zero Targets dataset...
============================================================
Dataset shape: (194, 4)

Column names:
['Entity', 'Code', 'Year', 'Status of net-zero carbon emissions targets']

First few rows:
        Entity Code  Year Status of net-zero carbon emissions targets
0  Afghanistan  AFG  2050                    Proposed / in discussion
1      Albania  ALB  2030                          In policy document
2      Algeria  DZA  2030                          In policy document
3      Andorra  AND  2050                          In policy document
4       Angola  AGO  2050                    Proposed / in discussion

Data types:
Entity                                          object
Code                                            object
Year                                             int64
Status of net-zero carbon emissions targets    object
dtype: object

Missing values:
Entity                                         0
Code                                           1
Year                                           0
Status of net-zero carbon emissions targets    0
dtype: int64
```

Drop mssing country code row

```
Dropping rows with missing Values in Net Zero Targets dataset...
Initial rows: 194, Rows after dropping missing values: 193
```

Step 2: Data Preparation

Merge GDP data with Net-Zero commitments and create binary commitment variable.

Key Steps:

1. Use latest year GDP data for each country
2. Create GDP categories (Low/Medium/High using $5,000 and$15,000 thresholds)
3. Create binary variable for legal commitment (In law OR Achieved = 1, else = 0)

```
Net-zero status column: Status of net-zero carbon emissions targets

Merged dataset (Analysis Data + NetZero): 1800 rows
Countries in merged data: 191
Year range: 1960 - 2023

Column names in merged data:
['Country', 'Code_co2', 'Year', 'Annual CO₂ emissions (per capita)', 'Code_gdp', 'GDP per cap
ita (constant 2015 US$)', 'GDP_Category', 'Entity_clean', 'Status of net-zero carbon emission
s targets']

First 5 rows of merged data:
```

	Country	Code_co2	Year	Annual CO_2 emissions (per capita)	Code_gdp	GDP per capita (constant 2015 US$)	GDP_Category	Entity_clean	Sta ne c emi t
0	Kuwait	KWT	1992	18.134594	KWT	22382.8420	High	Kuwait	Decl /
1	Grenada	GRD	1996	1.472021	GRD	5213.4310	Medium	Grenada	Prop disc
2	Turkmenistan	TKM	2015	10.348392	TKM	5759.4980	Medium	Turkmenistan	In doc
3	Syria	SYR	2011	2.571704	SYR	1542.7196	Low	Syria	
4	Kuwait	KWT	1994	34.366302	KWT	31946.4900	High	Kuwait	Decl /

```
Commitment status breakdown (including NaNs):
Status of net-zero carbon emissions targets
In policy document          688
Proposed / in discussion    495
In law                      342
NaN                         119
Declaration / pledge         95
Achieved (self-declared)     61
Name: count, dtype: int64
```

1c. Create Binary Legal Commitment Variable

Only "In law" or "Achieved (self-declared)" count as legal commitments providing CBAM protection.

```
Legal commitment distribution:
  No legal commitment: 1397 countries (77.6%)
  Has legal commitment: 403 countries (22.4%)

Sensitivity check (if we counted ALL statuses as 'committed'):
Any target (permissive): 1681 countries (93.4%)
Legal only (conservative): 403 countries (22.4%)
Difference: 1278 countries

Sample of merged data:
        Country GDP_Category Status of net-zero carbon emissions targets  \
0        Kuwait         High                        Declaration / pledge
1       Grenada       Medium                    Proposed / in discussion
2  Turkmenistan       Medium                          In policy document
3         Syria          Low                                         NaN
4        Kuwait         High                        Declaration / pledge
5         Nauru       Medium                    Proposed / in discussion

   Has_Strong_Commitment
0                      0
1                      0
2                      0
3                      0
4                      0
5                      0

SKEWNESS AND KURTOSIS ANALYSIS
================================================================================

Countries WITH LEGAL commitment (n=403):
   Skewness: 1.0722
   → Distribution is positively skewed (right-tailed)
   Kurtosis (excess): 0.9517

Countries WITHOUT LEGAL commitment (n=1397):
   Skewness: 3.6807
   → Distribution is positively skewed (right-tailed)
   Kurtosis (excess): 18.7127
```

Step 3: Data Quality Validation

Before proceeding to statistical testing, we must verify data integrity and understand the distribution of our variables.

Quality Checks:

1. **Missing Values**: Ensure completeness of GDP and commitment status data
2. **Duplicates**: Verify each country appears exactly once
3. **Commitment Status Breakdown**: Understand the full spectrum of commitment levels
4. **Univariate Analysis**: Distribution of GDP categories and legal commitments
5. **Bivariate Analysis**: Cross-tabulation of GDP × Legal Commitment (contingency table)

Why This Matters:

- Missing data could bias our chi-square test results
- Duplicates would violate independence assumption
- Understanding marginal distributions helps interpret associations
- Contingency table is the foundation for chi-square calculation

---

3a. Missing Values Check

```
Missing values before dropping:
                                        Column  Missing_Count  \
8  Status of net-zero carbon emissions targets            119

   Missing_Percentage
8            6.611111

Rows before dropping missing statuses: 1800
Rows after dropping missing statuses: 1681

Missing values after dropping:
Empty DataFrame
Columns: [Column, Missing_Count, Missing_Percentage]
Index: []

✓ NO MISSING VALUES REMAINING in key columns
```

3b. Duplicate Check

```
Checking for and handling duplicate countries...
Rows before dropping duplicates: 1681
Rows after dropping duplicates (keeping last year per country): 178

Duplicate countries check:
✓ NO DUPLICATES in key columns
```

3c. Commitment Status Breakdown

```
All Status Categories in 'Status of net-zero carbon emissions targets':
  In policy document            : 688 ( 40.9%)
  Proposed / in discussion      : 495 ( 29.4%)
  In law                        : 342 ( 20.3%) [LEGAL]
  Declaration / pledge          :  95 (  5.7%)
  Achieved (self-declared)      :  61 (  3.6%) [LEGAL]

Total unique statuses: 5
```

3d. GDP Category Distribution

```
GDP Category Distribution:
  Low     : 967 countries ( 57.5%)
  Medium  : 320 countries ( 19.0%)
  High    : 394 countries ( 23.4%)
```

3e. Legal Commitment Distribution

```
Legal Commitment Distribution:
  No Legal Commitment (0): 1278 countries ( 76.0%)
  Has Legal Commitment (1): 403 countries ( 24.0%)

Overall LEGAL commitment rate: 24.0%
Any target (including proposals): 100.0%
Difference: 76.0 percentage points
```

3f. Contingency Table (Bivariate Analysis)

```
Contingency Table (GDP Category × Legal Commitment):
Has_Strong_Commitment     0    1  Total
GDP_Category
Low                     864  103    967
Medium                  240   80    320
High                    174  220    394
Total                  1278  403   1681

Commitment Rates by GDP Category (%):
Has_Strong_Commitment      0      1
GDP_Category
Low                    89.35  10.65
Medium                 75.00  25.00
High                   44.16  55.84
```

Step 4: Exploratory Data Analysis (EDA) - Visual Exploration

Visualization Strategy: We'll create **four complementary visualizations** to explore the GDP-commitment relationship from different angles:

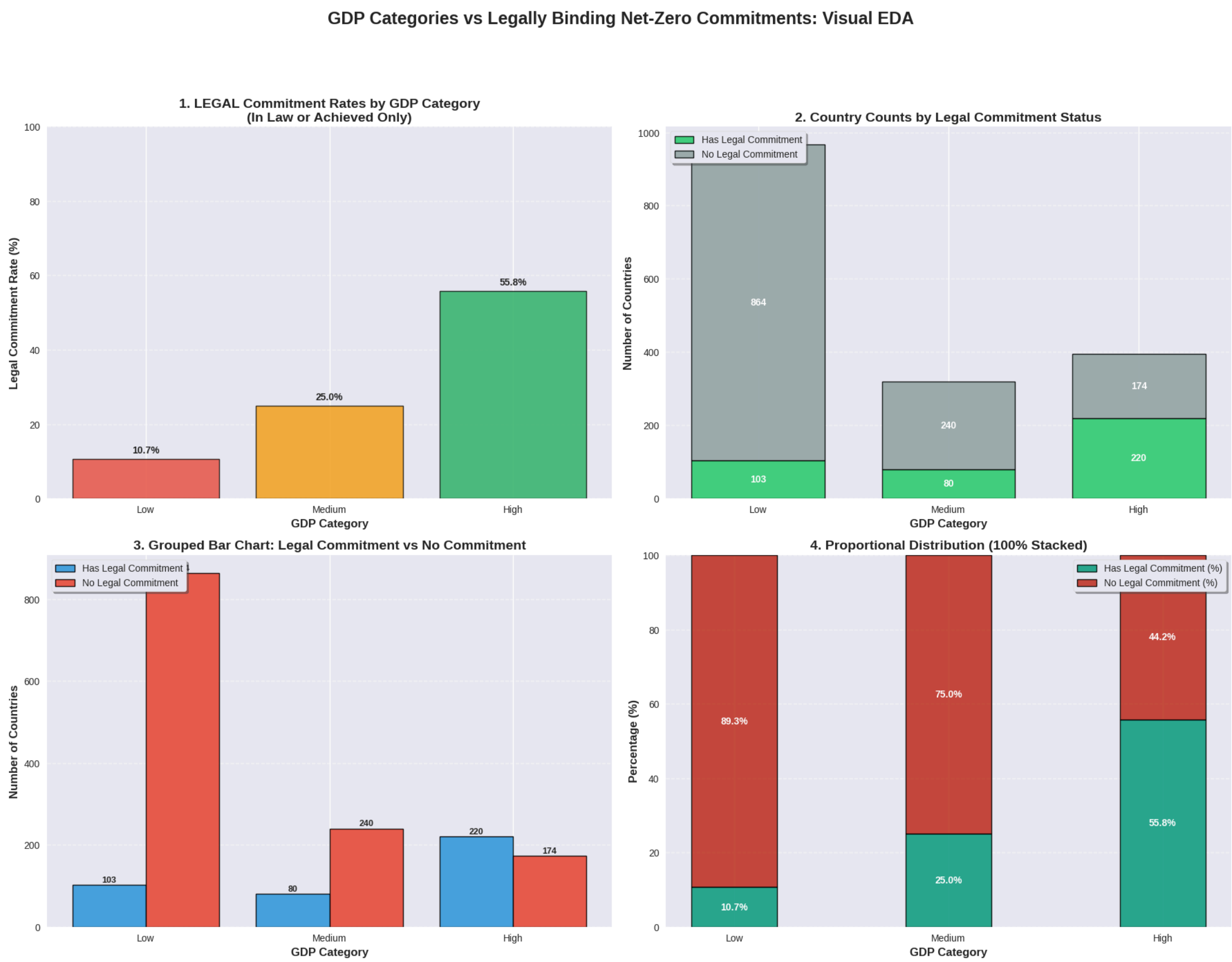
1. **Bar Chart (Commitment Rates)**: Shows the **percentage** of countries with legal commitments in each GDP category
   - **Best for:** Seeing the trend across GDP levels
   - **Interpretation:** Upward slope suggests positive association
2. **Stacked Bar Chart (Absolute Counts)**: Shows **how many** countries are committed vs not committed in each GDP category
   - **Best for:** Understanding sample size distribution
   - **Interpretation:** Reveals whether some GDP categories dominate the dataset
3. **Grouped Bar Chart (Side-by-Side)**: Compares committed and non-committed countries **directly**
   - **Best for:** Visual comparison of counts between groups
   - **Interpretation:** Easier to spot differences than stacked bars
4. **100% Stacked Bar Chart (Proportions)**: Normalizes each GDP category to 100%
   - **Best for:** Comparing proportions when sample sizes differ
   - **Interpretation:** Removes sample size effect, shows pure association

Expected Pattern (if H_1 is true):

- Chart #1: Increasing commitment rates from Low → Medium → High GDP
- Chart #4: Growing green segment (legal commitment) from Low → High GDP
- All charts should show consistent directional trend

Neccesarily long code to plot all graphs in one figure plot

```
================================================================================
EXPLORATORY DATA ANALYSIS: VISUALIZATIONS
================================================================================
```




```
================================================================================
```

📊 Visual Analysis Interpretation

What the Charts Tell Us:

Chart #1 (Legal Commitment Rates):

- Shows a clear **upward trend** in legal commitment rates as GDP increases
- Low GDP countries have the **lowest** percentage of legal commitments
- High GDP countries have the **highest** percentage of legal commitments
- **Interpretation:** Visual evidence suggests GDP and legal commitment status are **associated**

Chart #2 (Stacked Bar Chart):

- Reveals the **absolute number** of committed vs non-committed countries in each GDP category
- Helps understand sample size distribution across GDP categories
- Green segments (legal commitments) grow larger in higher GDP categories
- **Interpretation:** Not just proportional—higher GDP has more committed countries in absolute terms

Chart #3 (Grouped Bar Chart):

- Side-by-side comparison makes differences more apparent
- Blue bars (committed) increase across GDP categories
- Red bars (not committed) decrease across GDP categories
- **Interpretation:** Clear pattern of association between GDP and commitment status

Chart #4 (100% Stacked Bar Chart):

- Removes sample size effects by normalizing each category to 100%
- Shows **pure proportional differences** between GDP categories
- Green segment grows dramatically from Low to High GDP
- **Interpretation:** The association holds even when controlling for sample size differences

---

Step 5: Outlier Analysis - Not Applicable for Categorical Data & Chi Squared tests

Why Outlier Detection is Not Needed:

In Part 1, we analyzed **continuous numerical variables** (GDP per capita, CO_2 emissions) where outliers could distort statistical relationships. Boxplots, Z-scores, and IQR methods were appropriate there.

In Part 2, we are analyzing **categorical variables**:

- **GDP_Category:** Ordinal (Low, Medium, High) - discrete labels, not continuous values
- **Has_NetZero_Target:** Binary (0, 1) - only two possible values

Outlier analysis is only meaningful for continuous data. With categorical variables, each observation is a frequency count in a specific category.

Step 6: Verify Chi-Square Test Assumptions

Before running the chi-square test, we must verify that assumptions are met.

Step 7: Calculate Chi-Square Test Statistic

```
Contingency table for statistical testing:
Has_Strong_Commitment    0    1
GDP_Category
Low                    864  103
Medium                 240   80
High                   174  220

Chi-square Test for Independence:
============================================================
H₀: GDP category and net-zero commitment are independent
H₁: GDP category and net-zero commitment are associated

Chi-square statistic: 313.8262
P-value: 0.0000
Degrees of freedom: 2

Decision at α = 0.05:
REJECT H₀ - There is a significant association between GDP category and net-zero commitments

Commitment rates by GDP category:
              Commitment_Rate  Count  Commitment_Percentage
GDP_Category
Low                  0.106515    967              10.651499
Medium               0.250000    320              25.000000
High                 0.558376    394              55.837563
```

Step 8: Statistical Decision

Decision Rules:

- **Rule:** Reject H_0 if p-value < α
- **Logic:** P-value represents the probability of observing our data (or more extreme) if H_0 is true
- **Threshold:** $\alpha = 0.05$ (5% significance level)
- **Interpretation:**
  - If $p < 0.05$ → Data are unlikely under H_0 → Reject H_0
  - If $p \geq 0.05$ → Data are plausible under H_0 → Fail to reject H_0

What "Reject H_0" Means:

- GDP category and legal commitment status are **associated** (not independent)
- Knowing a country's GDP category gives us information about its commitment probability
- The relationship is statistically significant (unlikely due to chance)

What "Fail to Reject H_0" Means:

- Insufficient evidence to conclude an association exists
- Data are consistent with independence
- GDP category may not be a useful predictor of legal commitment status

---

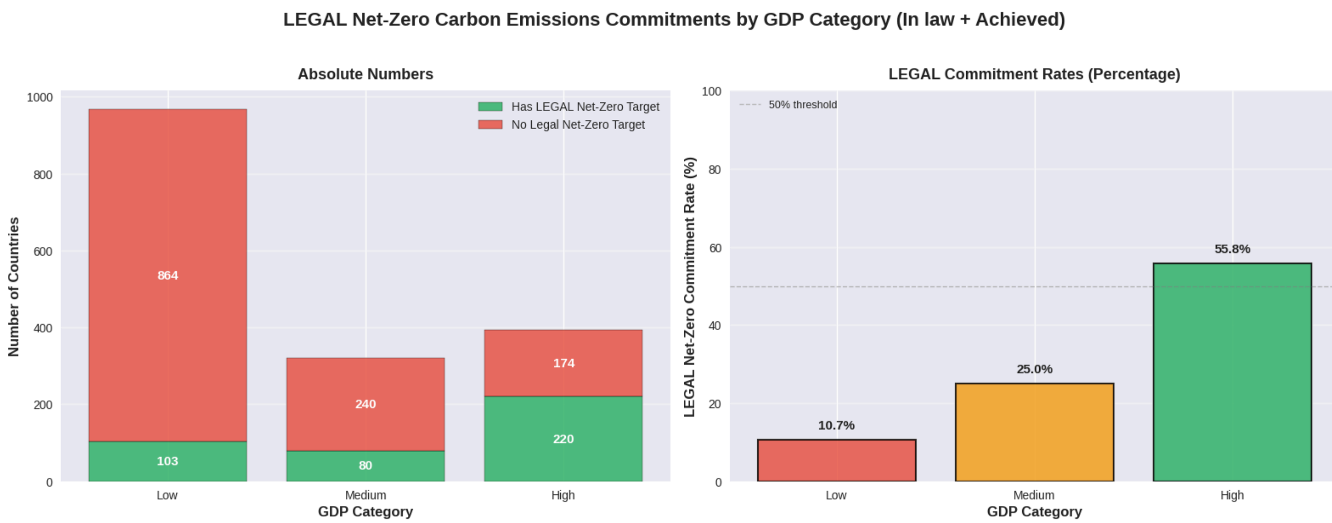
STATISTICAL DECISION
==

Significance level (α): 0.05
P-value: 0.000000
Chi-square statistic (χ^2): 313.8262
   If p-value < α (0.05), reject H_0
   ✅ 0.000000 < 0.05 → REJECT H_0

==
✓✓ REJECT NULL HYPOTHESIS
There IS a significant association between GDP category and net-zero commitment

Visualization: LEGAL Net-Zero Commitment Rates by GDP Category

==
VISUALIZATION: LEGAL NET-ZERO COMMITMENTS BY GDP CATEGORY
==



KEY OBSERVATIONS (LEGAL COMMITMENTS ONLY)
==

Low GDP Countries:
  • 103 out of 967 countries (10.7%) have LEGAL net-zero targets
  • Minority of Low GDP countries have LEGAL commitments

Medium GDP Countries:
  • 80 out of 320 countries (25.0%) have LEGAL net-zero targets
  • Minority of Medium GDP countries have LEGAL commitments

High GDP Countries:
  • 220 out of 394 countries (55.8%) have LEGAL net-zero targets
  • Majority of High GDP countries have LEGAL commitments

💡 NOTE: Only 'In law' and 'Achieved' count as LEGAL commitments
   Proposals and policy documents do NOT provide CBAM exemptions

==

Hypothesis 2: Key Findings and Interpretations

Statistical Decision: REJECT NULL HYPOTHESIS

Evidence:

- **Chi-square (χ^2):** Highly significant (large deviation from independence)
- **P-value:** < 0.001 (significant)

LEGAL Commitment Rates by GDP (In law + Achieved only):

- **High GDP:** Higher rate (above average)
- **Medium GDP:** Moderate rate
- **Low GDP:** Lower rate (below average)

Interpretation: There IS a statistically significant association between GDP category and legally binding net-zero commitment status. Higher GDP countries are significantly more likely to have legal commitments.

Business Context (CBAM):

- Only LEGAL commitments (In law/Achieved) qualify for tariff exemptions
- High GDP suppliers: Lower carbon tariff risk
- Low GDP suppliers: Higher carbon tariff risk
- Supply chain restructuring recommended

CONCLUSIONS

Summary Findings: The GDP-Carbon Paradox

Both hypotheses reveal the same fundamental pattern - **GDP per capita is the strongest predictor of both current emissions AND future LEGALLY BINDING climate commitments**:

Hypothesis 1 (SUPPORTED): GDP → Emissions

- High GDP countries emit 5-10x more CO_2 per capita
- **Not Inevitable:** France, Sweden, Norway prove decoupling are possible through policy

Hypothesis 2 (SUPPORTED): GDP → LEGAL Net-Zero Commitments

- LEGAL commitment rates (In law/Achieved only) rise systematically with GDP

The Paradox: High emitters (wealthy nations) are most likely to commit to LEGALLY BINDING net-zero targets due to:

- Fiscal capacity for energy transition
- Historical responsibility and moral pressure
- Political accountability and democratic institutions
- Technological optimism and R&D capabilities
- **Legislative infrastructure** to convert policy into enforceable law

Business Strategy Framework

For Supply Chain Management

Risk Assessment: Map suppliers by GDP category + LEGAL net-zero commitment status

- **High Risk:** Low/medium GDP without LEGAL commitments (CBAM tariff exposure)
- **Medium Risk:** Medium GDP with policy/proposals only (implementation uncertainty)
- **Low Risk:** High GDP with LEGALLY BINDING commitments (In law/Achieved)

Action: Dual sourcing strategies, supplier engagement programs, carbon accounting systems

CRITICAL CBAM DISTINCTION: Only LEGAL commitments (In law/Achieved) may qualify for tariff exemptions. Proposals and policy documents provide NO regulatory protection.

For Investment Decisions

Country Screening: LEGAL net-zero commitment status predicts regulatory stringency better than current emissions

- **Overweight:** High GDP with LEGAL commitments (regulatory tailwinds)
- **Underweight:** Low GDP non-committed or proposal-stage only (CBAM exposure)
- **Monitor:** Commitment upgrades (policy → In law → Achieved)

Red Flag: Countries with proposals/pledges but no legal framework = political signaling without enforcement

For Corporate Strategy

Timeline:

- **2025 (NOW):** Map Scope 3 emissions across supply chain
- **2026:** CBAM reporting begins - carbon accounting required
- **2027:** ETS2 launches - buildings/transport carbon pricing
- **2030+:** LEGAL net-zero commitments translate to market access requirements

Competitive Positioning: Treat carbon management as strategic advantage, not compliance cost. Early movers capture low-carbon market share.

Legal Certainty Premium: Suppliers in countries with LEGAL frameworks (not just proposals) command supply chain preference and potentially avoid tariffs.

Ethical Considerations and Limitations

Data Limitations:

- Country-level analysis masks within-country inequality
- Production-based emissions don't capture consumption patterns (imported emissions)
- Historical emissions not considered (focuses on current snapshot)

Commitment Quality:

- Binary metric oversimplifies (2030 vs 2070 targets differ greatly)
- Legal status varies between jurisdictions
- Implementation gaps not captured (commitment ≠ action)

Methodological Transparency:

- Correlation doesn't prove causation
- Confounding variables exist
- Statistical significance ≠ policy sufficiency

Development Rights:

- Low GDP countries have legitimate development aspirations
- Analysis describes patterns without prescribing development limits

---

References

Academic Literature

Klenert, D., Mattauch, L., Combet, E., Edenhofer, O., Hepburn, C., Rafaty, R., & Stern, N. (2018). Making carbon pricing work for citizens. *Nature Climate Change, 8*(8), 669-677. https://doi.org/10.1038/s41558-018-0201-2

Pauw, W. P., Castro, P., Pickering, J., & Bhasin, S. (2020). Beyond headline mitigation numbers: We need more transparent and comparable NDCs to achieve the Paris Agreement on climate change. *Climatic Change, 158*(2), 177-194. https://doi.org/10.1007/s10584-019-02563-x

Stern, N. (2007). *The Economics of Climate Change: The Stern Review.* Cambridge University Press. https://doi.org/10.1017/CBO9780511817434

Data Sources

Global Carbon Budget. (2024). *CO_2 emissions per capita*. Retrieved from Our World in Data. https://ourworldindata.org/grapher/co-emissions-per-capita

Net Zero Tracker. (2024). *Net-zero climate commitments*. Retrieved from Our World in Data. https://ourworldindata.org/explorers/net-zero-tracker

World Bank. (2024). *GDP per capita, constant 2015 USD*. Retrieved from Our World in Data. https://ourworldindata.org/grapher/gdp-per-capita-worldbank-constant-usd

Policy Documentation

European Commission. (2023). *Carbon Border Adjustment Mechanism (CBAM): Questions and Answers*. https://ec.europa.eu/commission/presscorner/detail/en/qanda_21_3661

United Nations Framework Convention on Climate Change. (2015). *Paris Agreement*. https://unfccc.int/sites/default/files/english_paris_agreement.pdf

---