Sustainability Trends🌿|
CO_2 Emissions & Deforestation Analysis

By Ritvaj Madotra

Contents

Introduction

Sustainability is a deeply personal subject to me. In today's world, narratives around climate change are often polarizing — with some urging people to avoid even plastic straws, and others dismissing the entire issue as a hoax. As someone who works with data, I’ve come to realize that data is our clearest window into the truth. If data can solve complex business problems, why not apply it to the sustainability crisis?

This project was born from that question. By leveraging publicly available World Bank datasets and as my journey into the Business analysis skills deepens, I aimed to cut through the noise and present a clear, evidence-backed view of sustainability trends — focusing on CO_2 emissions and deforestation.

Objective

To uncover global patterns and disparities in CO_2 emissions and forest loss between 2019 and 2021 by:

- Analyzing emissions and forest trends across countries.
- Comparing the environmental impact of income groups (high to low income nations) and population.
- Creating narrative-rich reports for informed awareness that can lead to better ways of tackling the idea of sustainability in our day-to-day lives.

This project aims to provide easily understandable, insight-rich visuals for both technical and non-technical audiences — encouraging meaningful conversations around climate data.

Tools and Skills Applied

- **Analytical Thinking** – Evaluating the value of information shown on-screen.
- **GPT-4o -** Brainstorming and evaluating ideas and feedback, support with coding errors.
- **ETL Concepts:** End-to-end data pipeline development from choosing and extracting data to loading the final transformed data ready for analysis.
- **Python (Pandas):** CSV file cleaning and export script for one dataset.
- **Excel:** Manual data profiling and cleaning for four additional datasets.
- **SQL (MySQL):** Data modeling, joining, and schema creation.
- **Power BI:** Data import, DAX measures and calculated columns, data modeling, visual storytelling.
- **Accessibility & UX Design:** Creative designing of the report, clean formatting, alt text, high contrast visuals.

Process Overview

The journey began by defining the **scope and narrative** of the project: to tell a story of environmental inequality through reliable data. I chose to focus on **countries** as the unit of analysis, organizing them by population, income level, and development level.

Data was sourced from the **World Bank** across five themes:

- Land Area
- Population
- Net National Income per Capita
- CO_2 Emissions

- Forest Area (Deforestation)

ETL Pipeline in Action:

- **Extract:** CSV data from World Bank
- **Transform:** Cleaned and standardized data using Python(Pandas) & Excel
- **Load:** Imported data into MySQL and built a structured database, then connected to the database using Power BI.
- **Model:** Connected to Power BI, performed final transformations
- **Visualize:** Created dual-page dashboard with focused storytelling

Data Preparation & Cleaning

Out of five datasets, one (Deforestation) was cleaned using Python. The remaining four were handled in Excel:

- **In Python**: I wrote a script using pandas to load the CSV, drop null values, rename columns for clarity, and export a cleaned version.

[20]:
```python
import pandas as pd

# Step 1: Load raw data
df_raw = pd.read_csv("Deforestation.csv")

# Step 2: Drop rows with any null values
df_cleaned = df_raw.dropna().copy()

# Step 3: Renaming columns for clarity
df_cleaned.rename(columns={
    'Country code': 'country_code',
    'year': 'year',
    'name': 'country_name',
    'forest_cover_percent': 'forest_cover_percent'
}, inplace=True)

# Step 4: Save the cleaned version
df_cleaned.to_csv("Deforestation_Cleaned.csv", index=False)

# Display sample output
df_cleaned.head(5)
```

[20]:

	country_code	year	country_name	forest_cover_percent
0	AFG	1990	Afghanistan	1.8528
1	AFG	1991	Afghanistan	1.8528
2	ALB	1990	Albania	28.7883
3	ALB	1991	Albania	28.7172

Jupyter Notebook: Python script for cleaning deforestation csv file

- **In Excel**: I profiled and cleaned data manually — removing empty rows, aligning column names, and deleting irrelevant columns.

Additionally, I merged 4 datasets (Deforestation, Population, Income per capita, & Country land area) into one master table: deforestation and merged 3 datasets (Co2_emissions, Income per capita, and population) into another master table: Co2_emissions.

This allowed me to connect the dimension tables with fact tables consistently.

Additional calculated columns were also created including matching up the average income per capita of each country into their respective economic classes.

Data profiling was conducted before upload to ensure accuracy and completeness.

Data Modeling & Transformation

I created a MySQL database to store and manage the cleaned files. Tables included:

- Co2_emissions (merged Co2_emissions, Population & Income per capita)

- Deforestation (merged Deforestation, Population, Income per capita, & Country land area)

```sql
CREATE DATABASE Sustainability_Analysis2;
USE Sustainability_Analysis2;

CREATE TABLE IF NOT EXISTS Population(
Country_code VARCHAR(5),
COUNTRY_NAME VARCHAR(250),
year int,
Population BIGINT);
ALTER TABLE population
ADD PRIMARY KEY (country_code, year);

INSERT INTO population (country_code, country_name, year, population)
SELECT `Country Code`, `Country Name`, year, Population
FROM population_temporary;
SELECT * FROM population;
```

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content:

Country_code	COUNTRY_NAME	year	Population
ABW	Aruba	2020	108587
AFE	Africa Eastern and Southern	2020	694446100
AFG	Afghanistan	2020	39068979
AFW	Africa Western and Central	2020	474569351
AGO	Angola	2020	33451132

MySQL: Schema and Tables creation using SQL

SQL Joins

Two key joins were written to connect the tables:

```sql
SELECT c.country_name,
  c.country_code,
  c.substance,
  c.emission_value_MtCo2,
  ROUND(c.emission_value_MtCo2 * 1000000 / NULLIF(p.population,0), 4)
    AS emissions_per_capita_tonnes,
  i.income_per_capita_us$ AS income_per_capita,
  CASE  WHEN i.income_per_capita_us$ <= 1135 THEN 'Low income'
        WHEN i.income_per_capita_us$ <= 4465 THEN 'Lower-middle income'
        WHEN i.income_per_capita_us$ <= 13845 THEN 'Upper-middle income'
        WHEN i.income_per_capita_us$ > 13845 THEN 'High income'
        ELSE 'Unknown'
        END AS income_group,
  p.population
FROM co2_emissions c
LEFT JOIN population p
  ON c.country_code = p.country_code
LEFT JOIN income_per_capita i
  ON c.country_code = i.country_code
ORDER BY c.country_name, c.year;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

country_name	country_code	substance	emission_value_MtCo2	emissions_per_capita_tonnes	income_per_capita	income_group	population
Afghanistan	AFG	CO2	1.739541352	0.0445	475.7180842	Low income	39068979
Afghanistan	AFG	CO2	1.737823132	0.0445	475.7180842	Low income	39068979
Afghanistan	AFG	CO2	1.715748742	0.0439	475.7180842	Low income	39068979
Afghanistan	AFG	CO2	1.742813406	0.0446	475.7180842	Low income	39068979
Afghanistan	AFG	CO2	2.198551191	0.0563	475.7180842	Low income	39068979
Afghanistan	AFG	CO2	2.037627184	0.0522	475.7180842	Low income	39068979
Afghanistan	AFG	CO2	1.899895303	0.0486	475.7180842	Low income	39068979

MySQL: Inner Join co2_emissions with population & income and creating calculated columns

```sql
SELECT d.country_code,
       d.year,
       d.country_name,
       d.forest_cover_percent,
       la.land_area_squarekm,
       ROUND(d.forest_cover_percent * la.land_area_squarekm / 100, 2) AS forested_area_squarekm,
       p.population,
       i.income_per_capita_us$,
       CASE WHEN i.income_per_capita_us$ <= 1135 THEN 'Low income'
            WHEN i.income_per_capita_us$ <= 4465 THEN 'Lower-middle income'
            WHEN i.income_per_capita_us$ <= 13845 THEN 'Upper-middle income'
            WHEN i.income_per_capita_us$ > 13845 THEN 'High income'
            ELSE 'Unknown'
            END AS income_group
FROM deforestation d
LEFT JOIN population p
  ON d.country_code = p.country_code
LEFT JOIN income_per_capita i
  ON d.country_code = i.country_code
LEFT JOIN country_land_area la
  ON d.country_code = la.country_code;
```

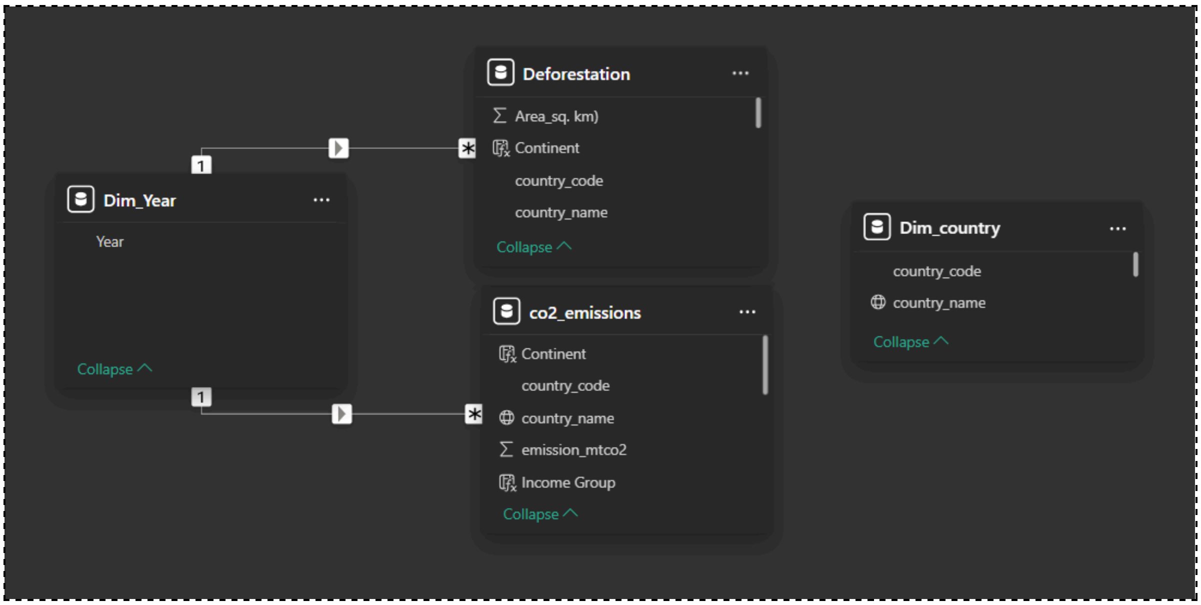
Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

country_code	year	country_name	forest_cover_percent	land_area_squarekm	forested_area_squarekm	population	income_per_capita_us$	income_group
ABW	1990	Aruba	2.3333	180	4.2	108587	20584.79126	High income
ABW	1991	Aruba	2.3333	180	4.2	108587	20584.79126	High income
ABW	1992	Aruba	2.3333	180	4.2	108587	20584.79126	High income
ABW	1993	Aruba	2.3333	180	4.2	108587	20584.79126	High income
ABW	1994	Aruba	2.3333	180	4.2	108587	20584.79126	High income

MySQL: Inner Join deforestation with population, area & income and creating calculated columns

In Power BI

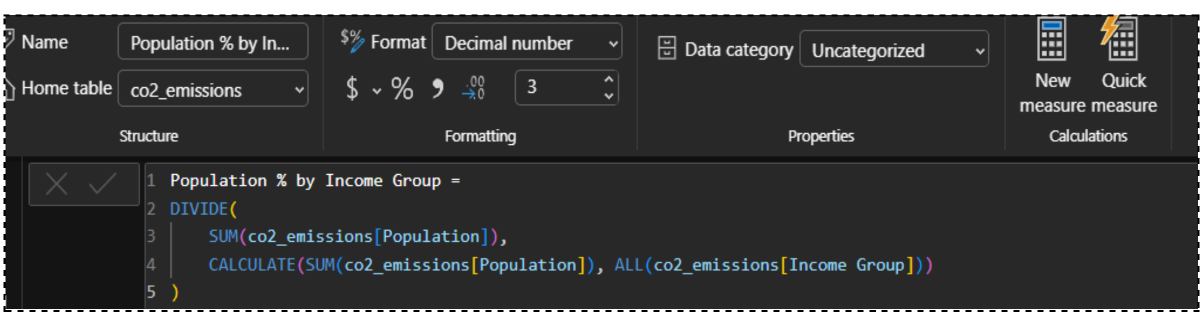
- Connected to MySQL server.
- Used Power Query for final tweaks and data profiling (removing anomaly values).
- Created a **Year** table to enable consistent time filtering.



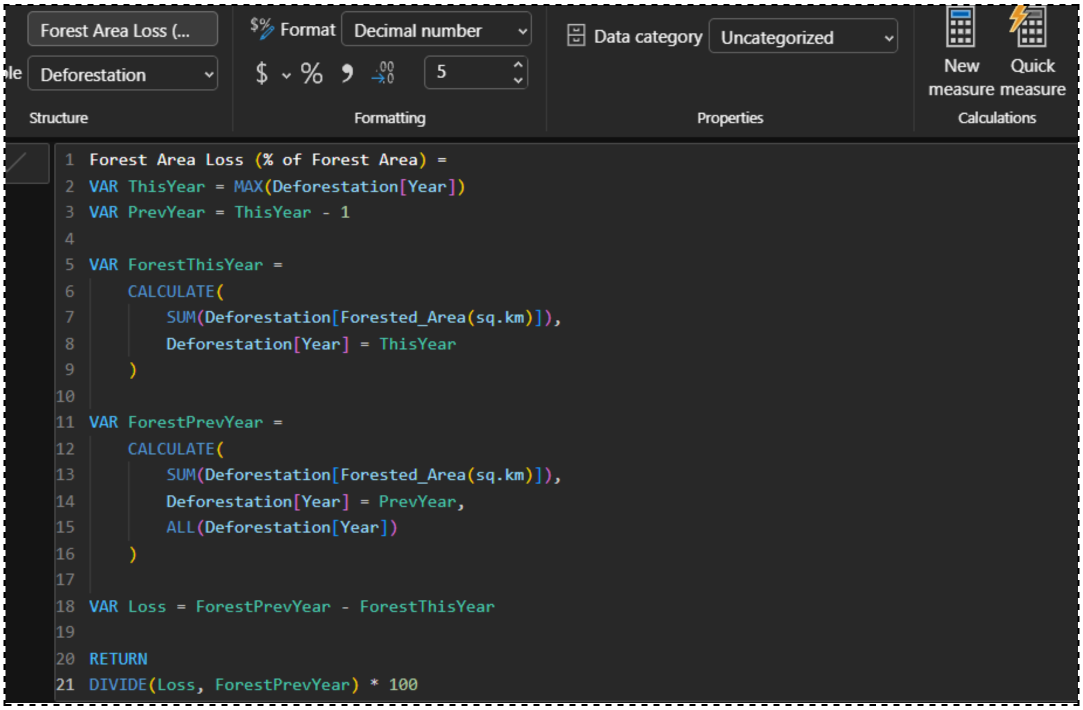
Data Model: **Star Schema** with shared Dim_year linked to both fact tables.

DAX Calculations

- In CO_2 Emissions: Created continent column, and measures like CO2 per Capita, Population % by Income Group.
- In Deforestation: Added continent, and Total Forested Area (2020) measures.



DAX Measure: Population % by Income Group



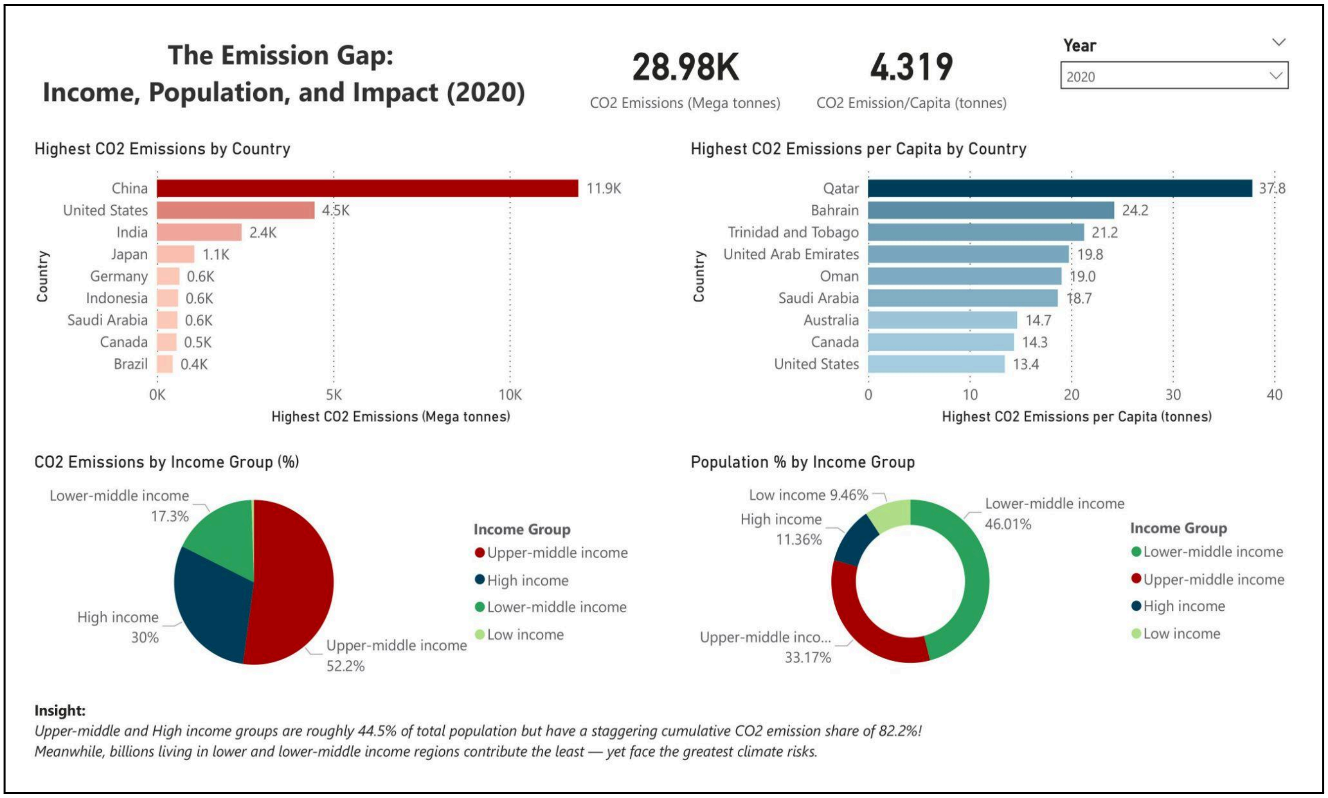

```
Forest Area Loss (% of Forest Area) =
VAR ThisYear = MAX(Deforestation[Year])
VAR PrevYear = ThisYear - 1

VAR ForestThisYear =
    CALCULATE(
        SUM(Deforestation[Forested_Area(sq.km)]),
        Deforestation[Year] = ThisYear
    )

VAR ForestPrevYear =
    CALCULATE(
        SUM(Deforestation[Forested_Area(sq.km)]),
        Deforestation[Year] = PrevYear,
        ALL(Deforestation[Year])
    )

VAR Loss = ForestPrevYear - ForestThisYear

RETURN
DIVIDE(Loss, ForestPrevYear) * 100
```

DAX Measure: Forest Area Loss (% of Forest Area)

Insights Derived from the Data

Page 1: The Emission Gap — Income, Population, and Impact (2020)

- **Top emitters by volume**: China (11.9K Mt), USA, India
- **Top emitters per capita**: Qatar (37.8t), Bahrain, Trinidad
- **Income group comparison**:
    - Upper-middle and high-income nations make only **44.53%** of the total world population and yet contribute **82.2%** of global CO_2!
    - Lower and lower-middle income countries emit far less (**17.8%**) — but suffer disproportionately.

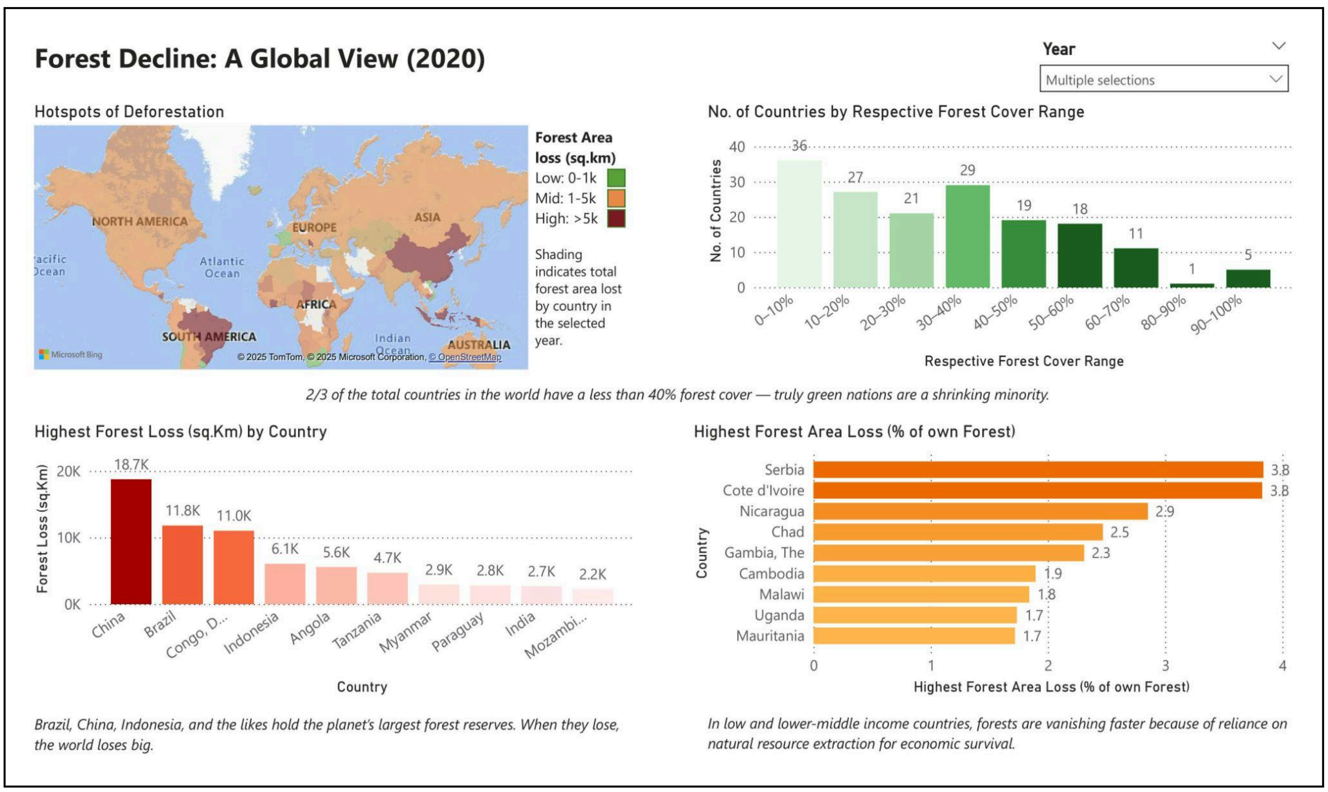


PowerBI Report I: Global CO_2 emissions analysis across income, population and region (2020)

- **Visual highlights**:
    - Card visuals for 2020 total emissions and average emissions by an individual in the year 2020.
    - Bar chart showing top CO_2 emitting countries.
    - Bar chart showing top countries with the highest per capita CO_2 emission.
    - Pie chart showing CO_2 emission distribution by income group.
    - Donut chart showing percentage of population in each income group.

Page 2: Forest Decline — A Global View (2020)

- **Hotspots of forest loss**: Brazil, China, Indonesia, Congo. Countries rapidly developing are deforesting to build infrastructure or using their forests as a resource.
- **Forest loss % of existing forest**: Serbia, Côte d'Ivoire, Nicaragua. Serbia lost **3.8%** of its total forest area in just a year.
- **Global tree cover snapshot**:
  - 2/3 of countries have less than **40% forest cover.**
  - True "Green Nations" are becoming a minority because the central tendency is shifting (reducing) towards 30% of forest cover out of the respective country's total land.
- These insights reflect an urgent imbalance in environmental responsibility vs vulnerability.



PowerBI Report II: Overview of Global forest loss patterns and risk concentration (2020)

- **Visual highlights**:
  - Filled world map of % forest area lost (Red to green gradient, where red depicts deforestation at an alarming rate vs green depicting reforestation efforts).
  - Histogram showing the number of countries with forest cover in ranges of 10% intervals each.
  - Column chart depicting the highest forest land losing countries in a year.
  - Bar chart depicting the highest proportional decrease in the forest percent of own forest.

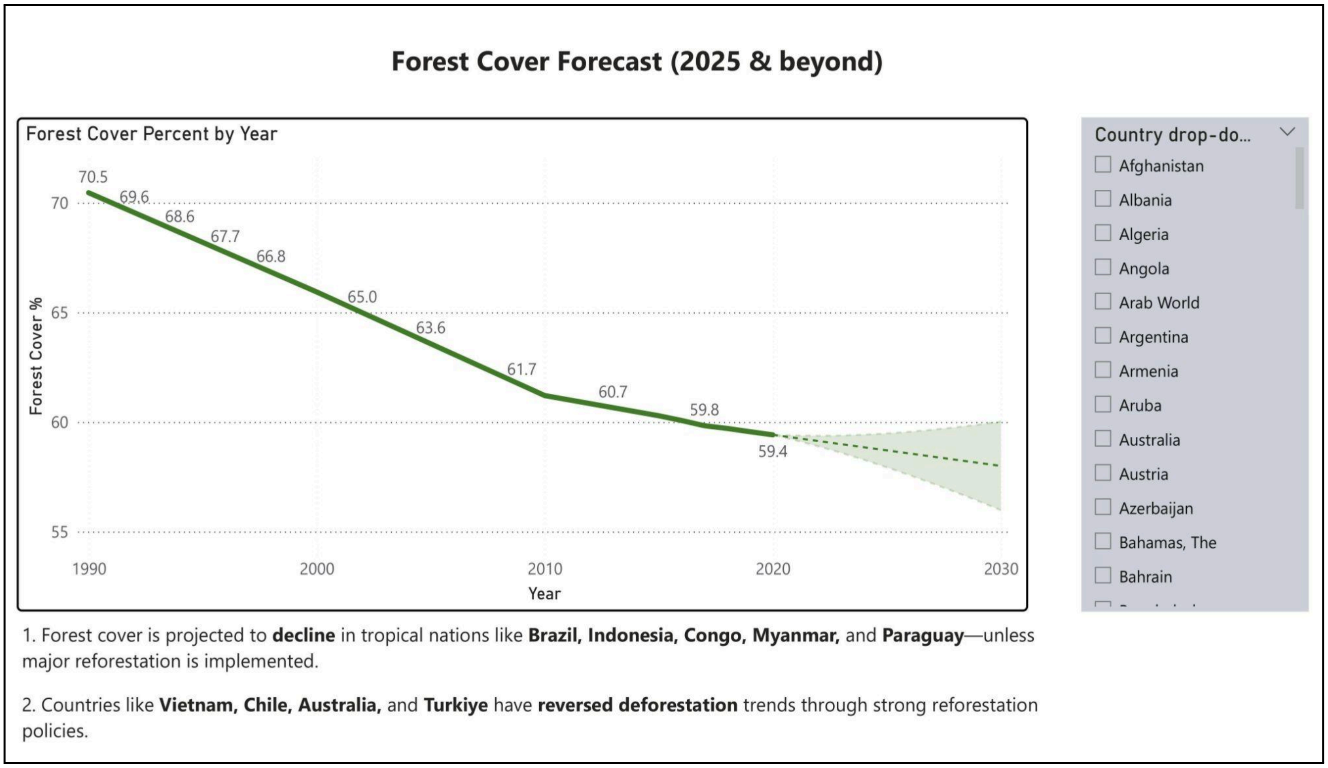
Page 3: Forest Cover Forecast (2025 & Beyond)

To highlight the future impact of current trends, added a forecast based on historical forest cover data. The forecast shows a continued decline in forest cover percentage in tropical nations unless major reforestation efforts are made.

Key Insights:

- **Projected decline** in forest cover for countries like **Brazil, Indonesia, Congo, Myanmar, and Paraguay**, based on trends from 1990–2020.
- **Positive gains** in countries like **Vietnam, Chile, Australia, and Türkiye**, where sustained reforestation policies have reversed deforestation.
- Visualized using **Power BI's forecast analytics**, with a **95% confidence interval** and historical trendline.

🧭This projection turns data into action—highlighting where we're headed unless global sustainability commitments are scaled up.



PowerBI Forecast: Prediction of future forest cover for each country (Current selection: **Brazil**)

Visual Design & Accessibility Features

I focused on making the report easy to digest for any reader. Here are the key elements employed:

- **Two themed pages** with clear headers and sub-sections.
- **Simple, non-cluttered visuals** with intuitive labeling and big font data labels for readability.
- **Accessibility features:**
  - Alt text for every chart
  - Consistent high-contrast color palette
  - Readable fonts

    - Donut and bar charts with categorical legends
    - Patterned shading for visual aid (where applicable)
- Positioning, sizing, and white space were all optimized for comprehension and visual balance.

Reflection

This project has been one of the most meaningful I've worked on. It allowed me to:

- Combine all my core data skills across SQL, Power BI, Python, and Excel and practice them hands-on. Although this was hard, but in the end this has given me immense confidence for my future analyses.
- Use data storytelling to break down sustainability myths. This was enlightening in the sense that there should be more talks about environmental justice, and greenwashing instead of misdirecting to the usual "Small things that you can do to protect the environment", while someone flies solo in their private jet burning tonnes of fuel and having an enormous Carbon Footprint!
- Deliver actionable insights through an elegant and professional report. I tried to deliver on the challenge of making the report easy to understand even for someone who has no relation to sustainability, or the analytics field.

The Python step — although simple — was an exciting way to apply what I'd recently learned. More importantly, the project helped me experience how impactful a clean narrative and good visuals can be in conveying a serious message.

References

- The World Bank Open Data: https://data.worldbank.org
- CO_2 Emissions dataset (World Bank)
- Deforestation dataset (World Bank)
- Population and Income datasets (World Bank)
- Power BI documentation: https://learn.microsoft.com/en-us/power-bi
- pandas Documentation: https://pandas.pydata.org