CO2 Emissions Flow Across Countries over the Decades: An Animated Sankey Story

Project by:
Sayali Pawar, Dept of Computing, DCU
Vaibhav Thalal, Dept of Computing, DCU

Abstract

Our project investigates how global CO2 emissions have changed across continents and countries over the past six decades. Using the World Bank's World Development Indicators dataset we aimed to create a clear and engaging visualization that captures long term shifts in emissions. After thorough data cleaning and exploration we developed an animated Sankey diagram to show how emissions flow from continents to countries across each decade.

The visualization reveals consistent growth in global emissions since 1960 with early dominance by North America and Europe, followed by sharp increases from Asia especially China and India from the 1990s onward. This animated format effectively highlights structural transitions and geographic shifts, illustrating how Asia has become the leading contributor in recent decades.

1. Dataset

The dataset used for this work was obtained from the World Bank's World Development Indicators (WDI) dataset from Kaggle https://www.kaggle.com/datasets/manchunhui/world-development-indicators, a publicly accessible resource that provides annual historical data across a wide range of developmental and environmental metrics. The raw dataset covers more than 150+ countries and regions and spans over sixty years, from 1960 to 2020. It contains 75,78,806 million rows and multiple CO2 related attributes, including emissions measured in kilotons and per-capita measures. The dataset includes six main fields 'CountryName', 'CountryCode', 'IndicatorName', 'IndicatorCode', 'Year' and 'Value' representing a mixture of numeric and categorical data. With its large row count, multi-indicator structure, and longitudinal nature, the dataset demonstrates essential characteristics of big data, particularly volume and variety. The World Bank continuously updates these indicators annually, reflecting a moderate degree of velocity as well. The combined size of the datasets is 800MB meeting the requirement for big-data analysis.

2. Data Exploration, Processing, Cleaning and/or Integration

Our investigation explores the evolution of global CO2 emissions across countries and continents over the past six decades. The dataset spans 263 regions, 1437 distinct indicators and 61 years (1960 to 2020) total of approximately 7.5 million rows. Although the dataset did not contain missing or null values, it was extremely complex because of its scale and the variety dimension of big data. For instance, the number of unique values in key columns illustrates this complexity. This made it essential to perform systematic cleaning, reduction and restructuring before analysis.

Data Cleaning and Processing

Because of the size and diversity of the dataset, our first step was to reduce it to a workable form. We began by extracting only indicators specifically related to CO2 emissions. The 'IndicatorName' column contained 1437 unique entries, many unrelated to emissions (e.g., fertility rates, GDP) We selected only

the indicators relevant to our analysis, including CO2 emissions (kt) and sub-category indicators representing emissions from solid fuel, liquid fuel, and gaseous fuel.

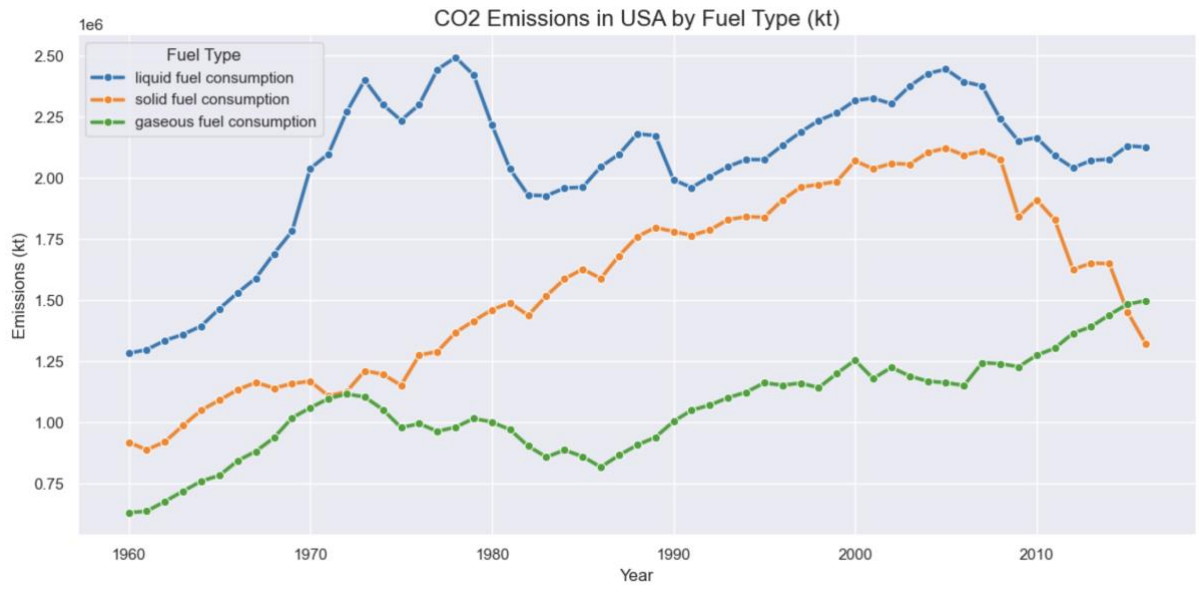
Once extracted, the dataset was further filtered by rows, keeping only years between 1960 and 2020. A decade column was created to support the design of the final Sankey visualisation. Additional processing included converting Year into integers and ensuring numerical consistency in CO2 values. Finally, we mapped each country to its continent using a custom mapping table, which was essential for grouping emissions flows in the Sankey diagram. This layered reduction transformed the original millions of rows into structured subsets tailored to the explanatory plots we intended to build.

The final dataset contained below values.

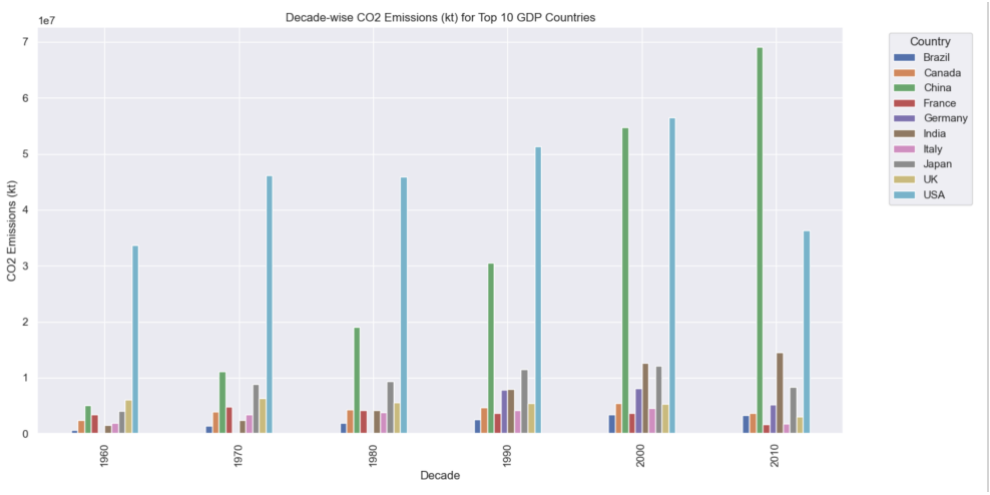
- Top 10 GDP countries.
- Continents where the countries belongs.
- Decade wise totals Source of CO2 emission: Solid, Liquid, Gaseous.
- Total value on CO2 emissions.

Data Exploration

Before designing our final explanatory visualisation, we created two exploratory plots to understand behavioural patterns in the cleaned data. The first exploratory graph focused on the United States world's highest-emitting country over time, plotting its CO2 emissions from solid, liquid and gaseous fuels over the years. This trend chart helped us confirm that the filtered indicators behaved consistently across time and reflected expected historical patterns such as the peak emissions in the 1970s to 1990s.



The second exploratory visualisation examined the top GDP ranking countries and visualised their total CO2 emissions aggregated per decade. This chart revealed how major economies such as the US, China, India, Japan, and Germany have shifted relative to one another in their emission contributions. These explorations guided the selection of attributes and the structure used for the final animated Sankey diagram by demonstrating which temporal patterns and groupings were most meaningful for storytelling.



3. Visualisation

The overall aim of the visualisation phase was to create a single, explanatory figure capable of telling a coherent and compelling story about how CO2 emissions have changed globally over six decades. Although exploratory charts helped us understand patterns within specific subsets of the data, the final visualisation had to integrate time, geography and emission magnitude into one graphic. After several iterations, we selected an animated Sankey diagram as the final representation because it could accurately illustrate the long-term flow of emissions from decades to continents and then to individual countries. This visual structure allowed us to condense the complexity of global emissions into an interpretable and narrative-driven form.

Choice of Chart Type

Selecting the chart type required balancing data complexity with narrative clarity. Our dataset contained multiple layers time (six decades), geography (continents and more than 200 regions), and magnitude of emissions measured in kilotons. Most traditional chart types such as line charts, bar charts, scatter plots or heatmaps would either flatten the hierarchy or relational flow between groups. A Sankey diagram however is specifically designed to present flows between categories while preserving proportional magnitude making it especially useful for multi-level hierarchical data.

In the context of CO2 emissions, the Sankey structure enabled the representation of a flow:
Sankey Flow:

Fuel Type > Continent > Total emission > Countries (timeline 6 decades 1960 – 2010)

The thickness of each connection corresponded directly to the volume of emissions. Unlike a static bar or line chart, which would require separate plots for different continents or countries a Sankey diagram presents everything in a single visual frame. The animation component further impacted the storytelling, as it allows to see how global emission patterns shift across decades rather than viewing each decade as an isolated image.

This chart type was therefore ideal for our dataset because it handles large categories gracefully and highlights proportional differences. The ability to animate transitions made the insights more intuitive, particularly the dramatic increase in Asia's emissions beginning in the 1990s and the relative stabilisation or decline in Europe and North America.

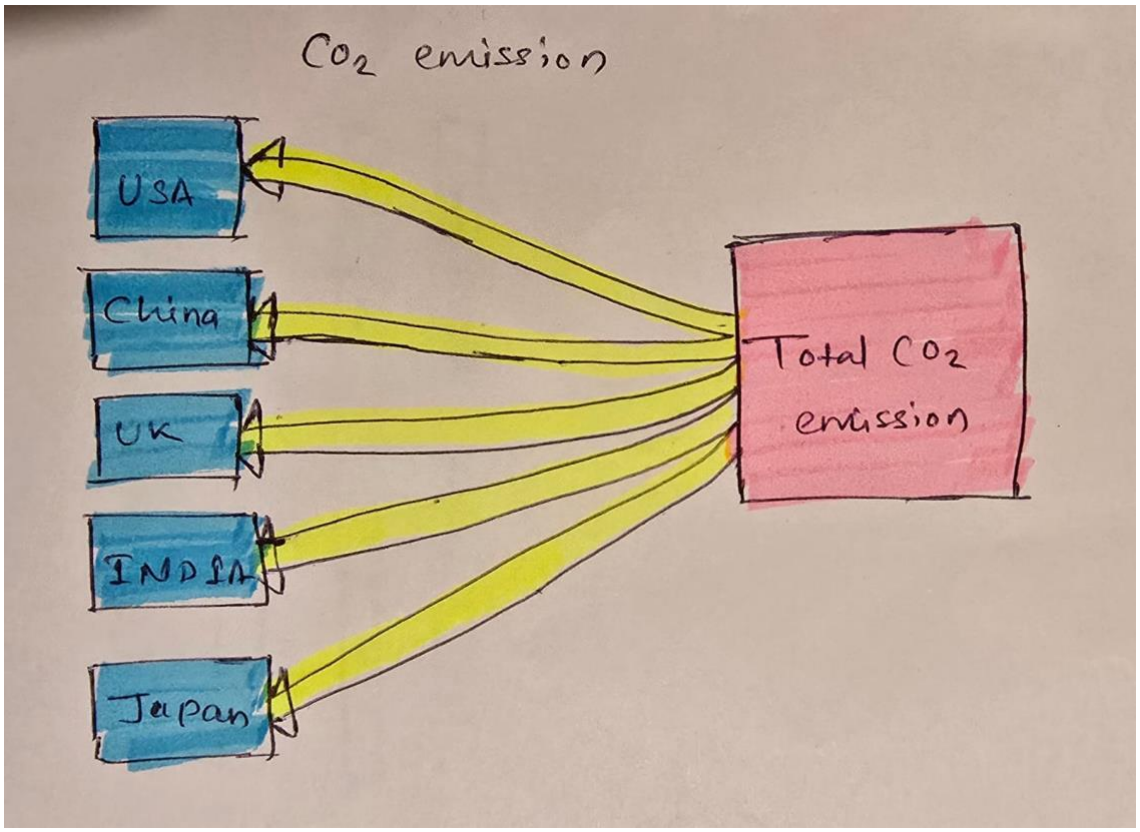
Design Process

The design phase began with constructing a conceptual layout. Initially, we sketched a left-to-right flow with fuel type positioned on the far left, continents in the centre and countries on the right. This arrangement mirrors natural reading patterns and ensures that the direction of flow feels intuitive. Each

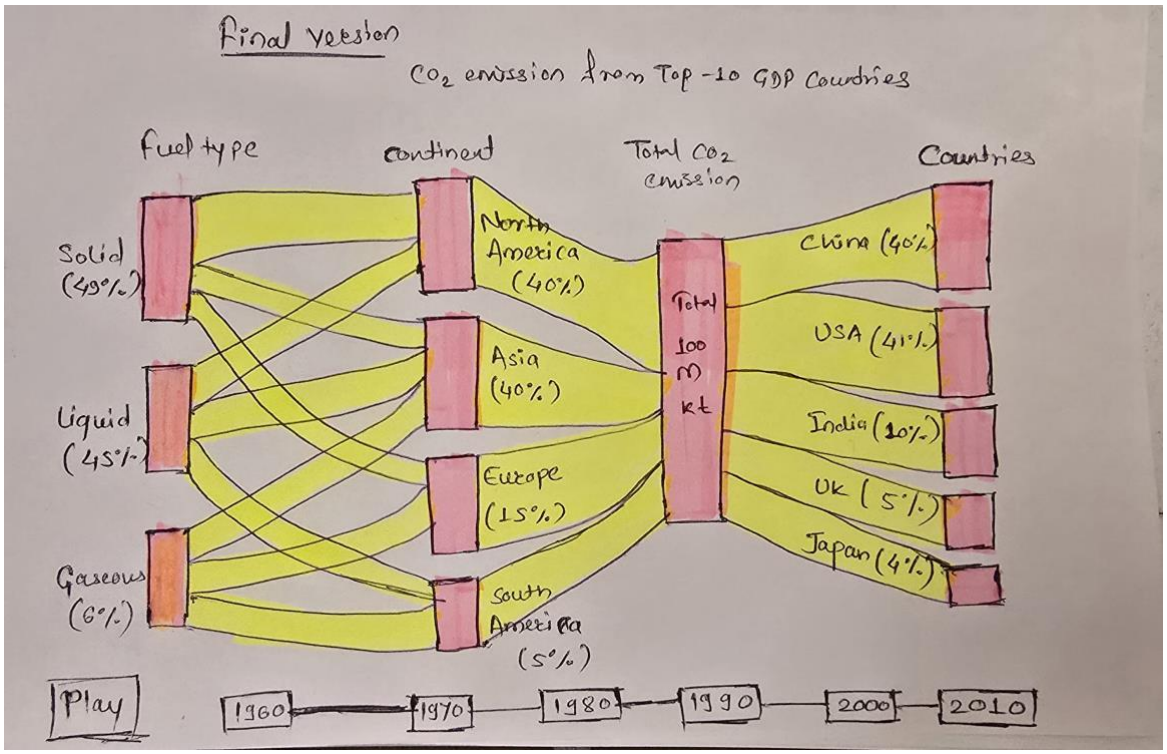
layer had to be spaced evenly to avoid node overlap, especially on the right side where many countries appear.

Sketches

1. **Initial Sketch**

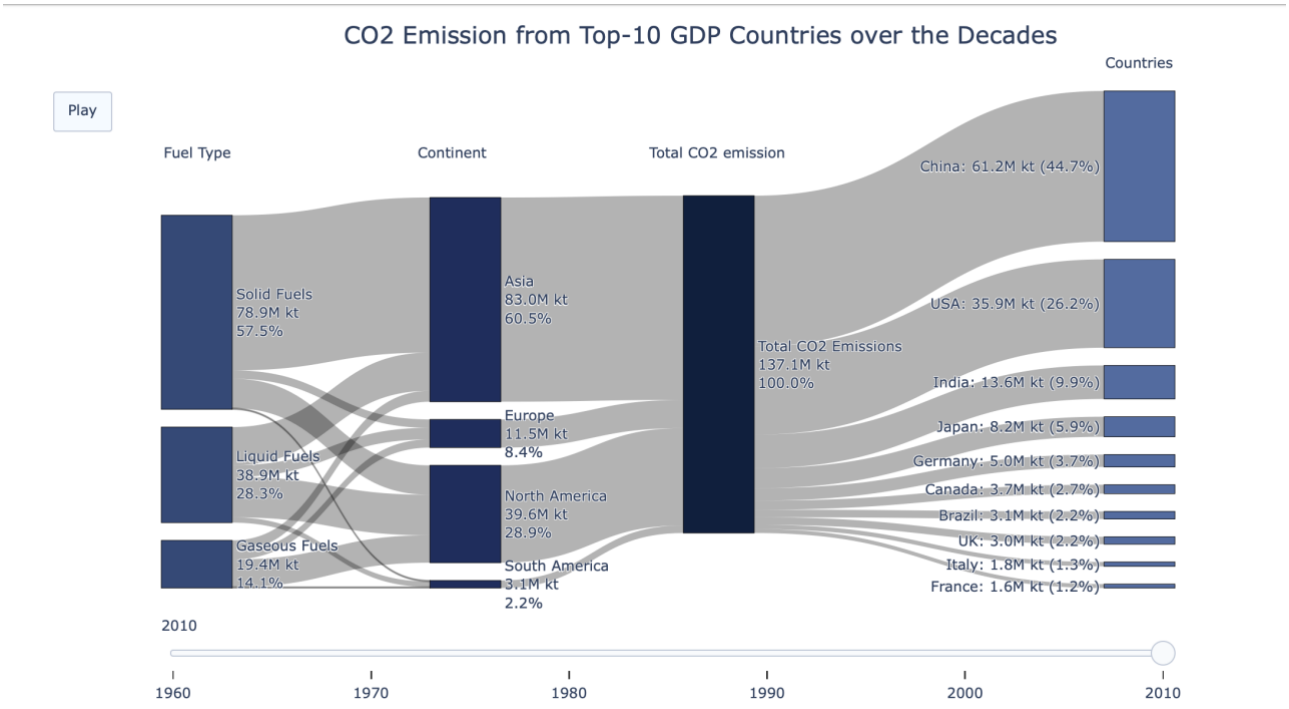


2. **Final Sketch**



After finalising this conceptual design, we implemented it using Plotly, which supports interactive and animated Sankey diagrams. The data was transformed into a hierarchical mapping that Plotly could interpret each node was assigned a unique ID and each flow was defined by source, target and a corresponding emission value. The animation slider was then added to transition smoothly between decades, allowing viewers to click or drag through time. Testing the animation helped refine pacing transitions needed to be slow enough to convey change but fast enough to avoid visual fatigue. Multiple iterations were required to balance clarity, design, spacing and readability.

Our Final Plot:



Design Choices

a. Colour

The colour scheme designed to reflect the relative value of each section using four shades of blue to establish a clear hierarchy. The Total CO2 Emission node representing the sum of all contributions is shown in the darkest shade to emphasize its dominance. The Continent section follows with the second darkest tone highlighting aggregated regional shares. Fuel Type is displayed in a medium shade, while the Country section comprising smaller and individual contributions is rendered in the lightest blue. This gradient approach ensures visual coherence, guiding the viewer's eye from the most significant category to the most granular, in line with design principles from Tamara Munzner and Colin Ware that advocate for perceptual hierarchy and clarity in data visualization.

b. Layout

To enhance clarity and user engagement several thoughtful design choices were made in constructing the Sankey diagram. Typography was kept minimal with concise labels and interactive tooltips enabled by Plotly's hover features used to display exact values without cluttering the layout. Nodes were arranged horizontally to support a linear temporal narrative, with Fuel Type as the root and continents spaced to minimize line crossings, ensuring readability and flow continuity.

c. Animation

Animation played a central role in conveying the progression of global emissions over time. A timeline of decades was added at the bottom of the plot, along with a play button that triggers automatic progression through each decade. As the timeline advances the Sankey diagram dynamically updates changing the values and flows for Fuel Type, Continent, Total Emission and Country sections making structural transitions and emission shifts visually intuitive and easy to follow.

4. Conclusion

The final visualisation successfully communicates how global CO2 emissions have shifted over six decades, revealing both growth in overall emission levels and major geographic transitions in responsibility. The animated Sankey diagram allowed complex and multilayered data to be represented in a single coherent structure. Clearly showing the flow of emissions from decades to continents and then to individual countries. It highlights patterns such as the early dominance of North America and Europe and the significant rise of Asia in more recent decades providing an intuitive understanding of long term global change.

Python, Pandas, Plotly and Jupyter Notebook were the primary tools used to clean, process and visualise the data. While the final visualisation achieves its purpose, there are areas where improvements could be made. The Sankey diagram can become visually dense when many countries are displayed, and a more customisable library such as D3.js could offer greater control over spacing and label placement. Additional contextual layers or supplementary visuals might also enhance the narrative though they were not included due to the requirement of presenting only one explanatory graph.

Overall, the project demonstrates that even highly complex and high volume environmental data can be distilled into a clear visual story when the right design choices and transformations are applied.

Contribution:

This project was completed collaboratively with equal involvement from both team members. Together we selected and cleaned the CO2 dataset. Sayali focused on filtering years, creating the decade column and ensuring numerical consistency, while Vaibhav built the continent mapping grouped data on continent values. Both contributed to exploratory analysis, Sayali developed the USA trend chart and Vaibhav created the comparative chart of top GDP countries. The animated Sankey diagram was co-designed, with Vaibhav structuring the node hierarchy and flows and Sayali refining layout, colours and animation controls.

References

Sankey Diagram in Python by Plotly https://plotly.com/python/sankey-diagram/

Build a Sankey diagram by Sigma https://help.sigmacomputing.com/docs/build-a-sankey-diagram

Amazon Sankey diagram by Sankeyart https://www.sankeyart.com/sankeys/public/60097/

Datacamp - https://www.datacamp.com/tutorial/sankey-diagram

Dataset - https://www.kaggle.com/datasets/manchunhui/world-development-indicators