Analysing Carbon Emissions Using MySQL

Anesu David Pasipanodya

GITHUB

<https://github.com/AnesuorDavid>

Contents

[Analysing Industry Carbon Emissions with SQL 3](#_Toc203531737)

[Introduction 3](#_Toc203531738)

[1. Defining the Project 3](#_Toc203531739)

[Columns 3](#_Toc203531740)

[2. Data Uploading 4](#_Toc203531741)

[3. Data Normalisation 5](#_Toc203531742)

[6 & 7. Data Aggregation + Final Reporting and Documentation 5](#_Toc203531743)

[Question 1. 6](#_Toc203531744)

[Question 2 7](#_Toc203531745)

[Question 3 7](#_Toc203531746)

[Question 4 8](#_Toc203531747)

[Question 5 9](#_Toc203531748)

[Question 6 9](#_Toc203531749)

[Conclusion 10](#_Toc203531750)

# Analysing Industry Carbon Emissions with SQL

## Introduction

This project aims to analyze product-level carbon emissions across various industries using structured data and SQL. Greenhouse gas emissions from manufactured products—including food, clothing, appliances, and consumer goods—account for more than 75% of total global emissions. Understanding the distribution of these emissions is vital for addressing the ecological challenges associated with climate change.

The dataset used in this project is sourced from [The Carbon Catalogue](https://www.nature.com/articles/s41597-022-01178-9), a publicly available research database that compiles product carbon footprint (PCF) data from companies around the world. Each record in the dataset represents a specific product and includes attributes such as the year, product name, company, country, industry group, total carbon emissions, product weight, and the percentage breakdown of emissions by production stage: upstream, operations, and downstream.

The project follows a structured SQL workflow, including data upload, cleaning, normalization, and analysis. Key objectives include identifying the industries and products with the highest emissions, calculating emissions per kilogram of product, and analyzing the contribution of different production stages to the overall carbon footprint.

By utilizing relational database techniques and SQL queries, this analysis provides insights into which sectors and production processes have the largest environmental impact, helping to inform data-driven sustainability decisions in industry and policy.

* To successfully analyse this dataset, I will use my 7 steps to data analysis approach.

**My 7 steps in my Data Analysis Approach**

# Defining the Project

The following project analyses global carbon emissions by industry using SQL and a dataset sourced from The Carbon Catalogue. By examining Product Carbon Footprints (PCFs) from various industries, the goal is to identify which sectors contribute the most to global greenhouse gas emissions and highlight areas for ecological improvement.

**Project Objective**

* Understand how carbon emissions are distributed across industries, companies and various products
* Identify the highest-emitting sectors
* Explore emission intensity relative to product weight
* Break down emissions by lifecycle stage: upstream, operations, downstream

1. **Knowing the Dataset**

*Dataset Overview*

The data is publicly available and includes carbon footprints for products manufactured by various companies. Each row in the dataset corresponds to a product and breaks down its emissions by stage in the supply chain.

***Table: carbon\_emissions***

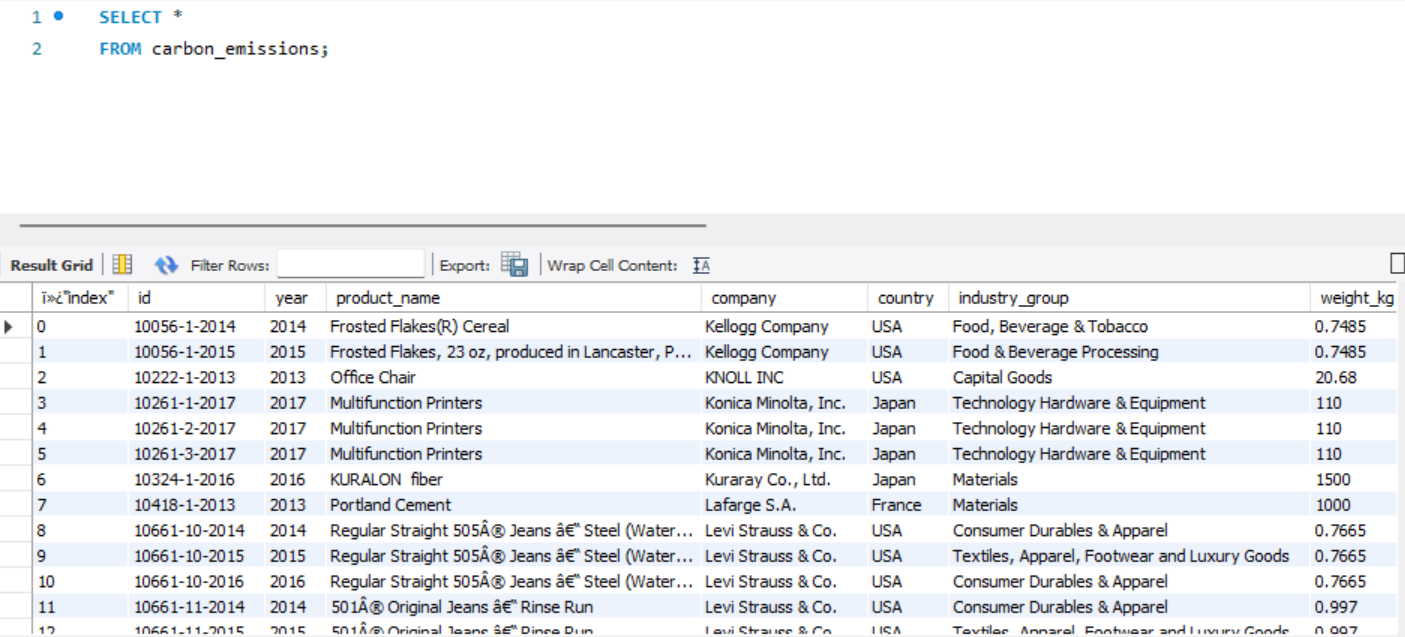
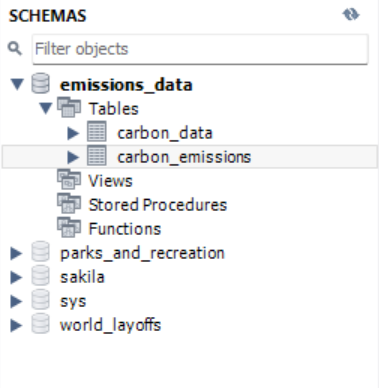
Columns:

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| id | VARCHAR | Unique identifier for each product entry |
| year | INT | The year the emission data was recorded |
| product\_name | VARCHAR | Name of the product |
| company | VARCHAR | Company that produced the product |
| country | VARCHAR | Country where the company is based |
| industry\_group | VARCHAR | Industry category the product belongs to |
| weight\_kg | NUMERIC | Weight of the product in kilograms |
| carbon\_footprint\_pcf | NUMERIC | Total greenhouse gas emissions (in CO₂e) for the product |
| upstream\_percent\_total\_pcf | VARCHAR | Percent of emissions from upstream (raw materials, etc.) |
| operations\_percent\_total\_pcf | VARCHAR | Percent of emissions from operations (manufacturing, processing) |
| downstream\_percent\_total\_pcf | VARCHAR | Percent of emissions from downstream (transportation, use, disposal) |

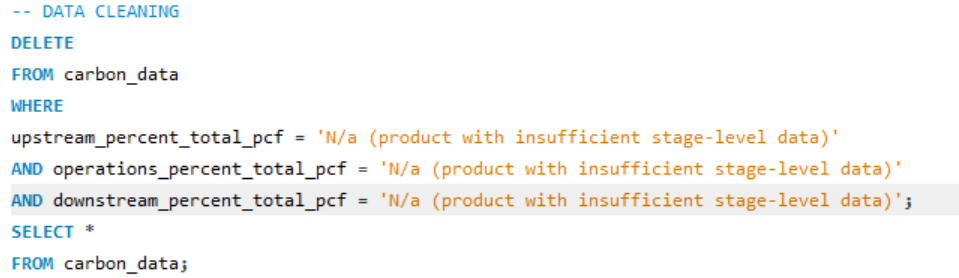
**Key Questions Explored**

1. Which industry groups have the highest total carbon footprint?
2. What is the average carbon footprint per product by industry?
3. Which industries have the highest emissions per kilogram of product?
4. How are emissions distributed across upstream, operations, and downstream stages in each industry?
5. Which countries are associated with the highest emitting industries?
6. What are the top 10 highest-emitting products overall?

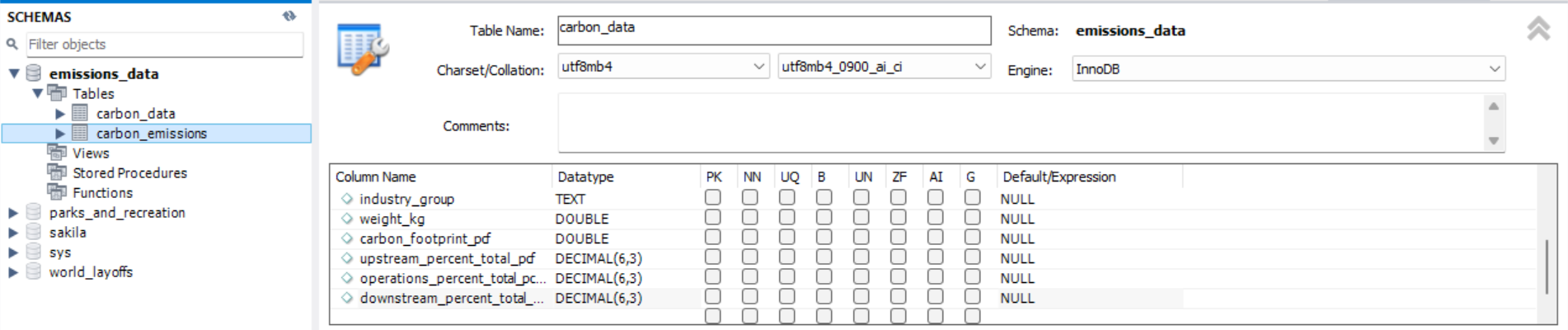
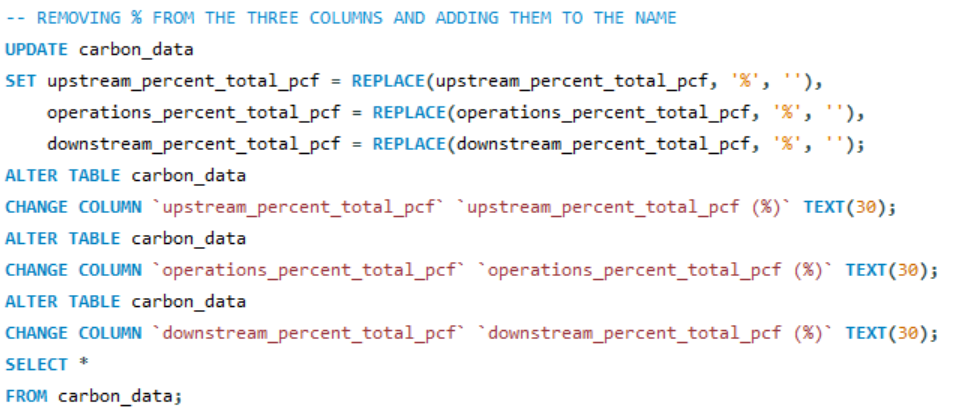
# Data Uploading

I uploaded the carbon\_emissions file as a .csv file. I made a schema named it emissions\_data, defined the table and named it carbon\_emissions.

***Data Cleaning***

First I created a staging table named carbon\_data. Then I got rid of all values from the upstream\_percent\_total\_pcf, operations\_percent\_total\_pcf, downstream\_percent\_total\_pcf columns which contained the following value ***“N/a (product with insufficient stage-level data)”*.** This is because if this is not done it would be impossible to normalise the data later.

## 3.Data Normalisation

****This was done by first removing the aforementioned values and then changing the column names by changing the 3 column names to upstream\_percent\_total\_pcf (%), operations\_percent\_total\_pcf (%), downstream\_percent\_total\_pcf (%). I also changed the data type to double from text. This is so the numbers are not ordered alphabetically but ordered them by value size.

# 6 & 7. Data Aggregation + Final Reporting and Documentation

### Question 1.

***Which industry groups have the highest total carbon footprint?***

**Aggregation**

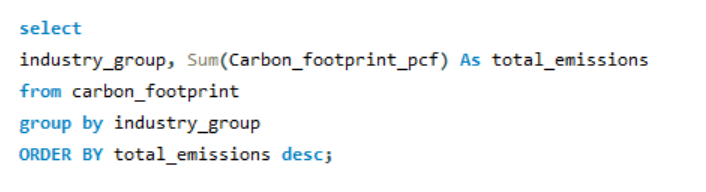
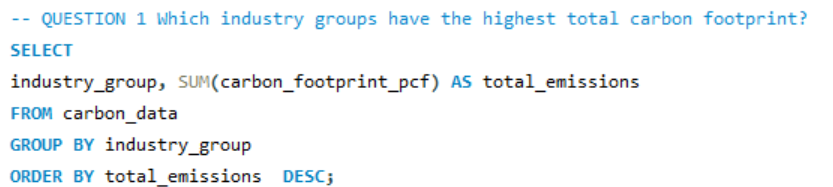
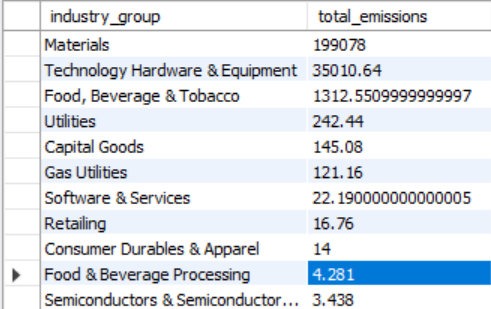
There are 5 columns of data that outline the total output of emissions

|  |
| --- |
| *Fig. Table screenshot from MYSQL.* |

|  |  |
| --- | --- |
| **Column Name** | **Meaning** |
| **weight\_kg** | **Weight of the product** in kilograms. This is used to understand emissions **per kg** or emissions **intensity**. |
| **carbon\_footprint\_pcf** | **Total carbon footprint** of the product, in kilograms of CO₂ equivalent (kgCO₂e). This is the **total emissions** across the entire life cycle of the product. |
| **upstream\_percent\_total\_pcf** | The **percentage of the total carbon footprint** that comes from **upstream activities** — like raw material extraction, processing, and inbound transportation. |
| **operations\_percent\_total\_pcf** | The **percentage of the total carbon footprint** from **operations** — including manufacturing, energy use, and on-site production. |
| **downstream\_percent\_total\_pcf** | The **percentage of the total carbon footprint** from **downstream activities** — like distribution, customer use, and disposal. |

*Fig. AI generated table of definitions based on table values*

So based on these column descriptions, carbon footprint may be the best indicator of the industry with the highest number of emissions or total output in kgCO₂e (carbon\_footprint\_pcf). SO to answer the question I am going to use the SUM of industry groups carbon footprint rounded of to 2sf, and analyse them in order of emissions. I will ignore the year in these prompts.

**QUERY Answer**

Here is the list of Industry groups with highest emissions in order of highest to lowest

|  |
| --- |
| 1. **Materials** |
| 1. **Technology Hardware & Equipment** |
| 1. **Food, Beverage & Tobacco** |
| 1. **Utilities** |
| 1. Capital Goods |
| 1. Gas Utilities |
| 1. Software & Services |
| 1. Retailing |
| 1. Consumer Durables & Apparel |
| 1. Food & Beverage Processing |
| 1. Semiconductors & Semiconductor Equipment |
| 1. Materials |
| 1. Technology Hardware & Equipment |
| 1. Food, Beverage & Tobacco |

**Reporting and Documentation**

Based on the grid above **Materials**, that is, concrete, fibrE, and automative batteries etc., have the highest level of emissions. This is followed by **Technology Hardware & Equipment** then Food, Beverage & Tobacco. In fourth place is Utilities.

Interesting insights in the results

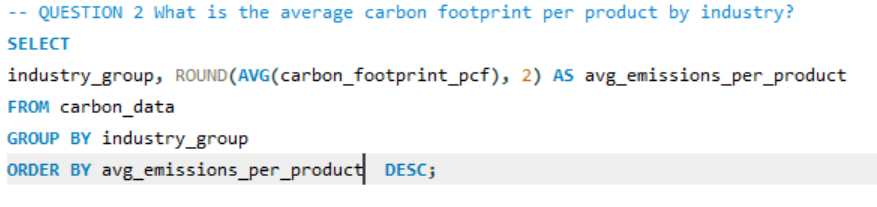
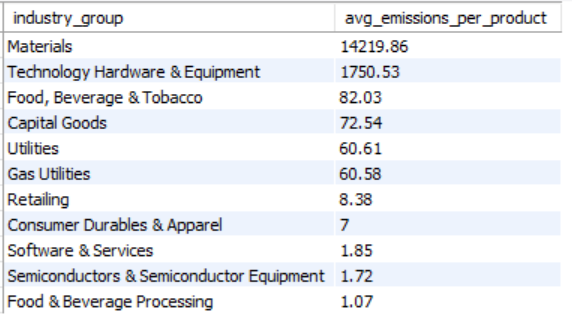
* + The top four contain more emissions by the kilogram than any of the bottom 10.
  + The **Materials** are almost 2 times the total emissions of the **Technology Hardware & Equipment**

### Question 2

***What is the average carbon footprint per product by industry?***

**Aggregation**

Group data by **industry\_group**

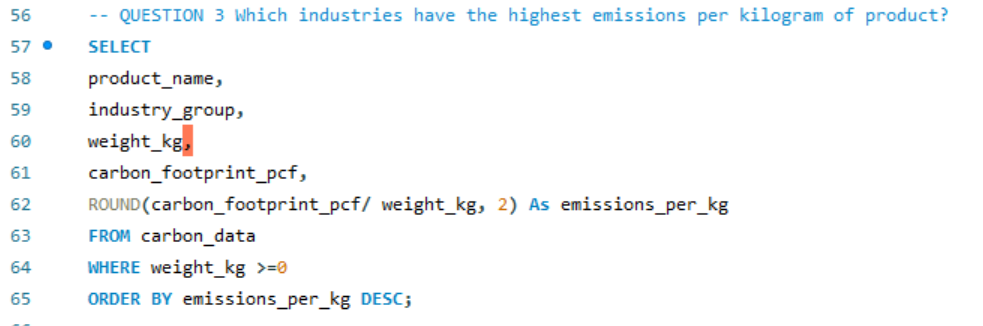
Calculate the **average** of carbon\_footprint\_pcf for all products in each industry. Then round it to 2 decimal places.

**Reporting and Documentation**

Based on the results above Materials has 10 times the average emissions per product than its next counterpart Technology Hardware & Equipment. Like for total emissions Food, Beverage & Tobacco industry groups uses the third most amount but interestingly enough Capital Goods (office chairs) use more average emissions per product than Utilities.

### Question 3

*Which industries have the highest emissions per kilogram of product?*

**Aggregation**

I ordered the data by product name, industry, weight, and the emissions per kg.

**Reporting and Documentation**

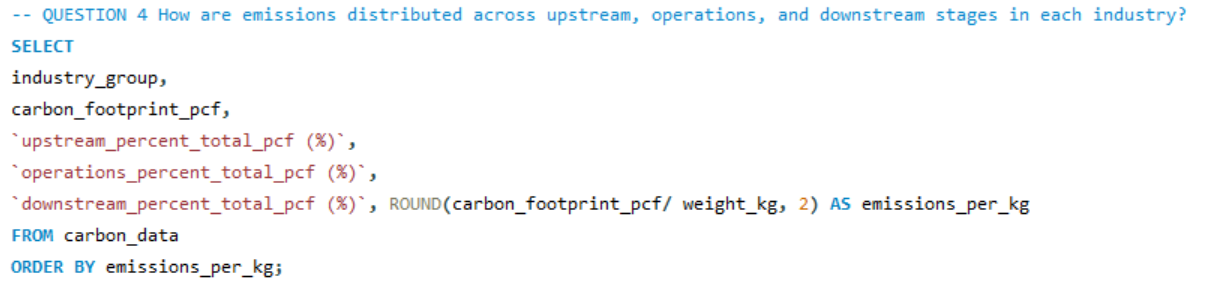
The image above shows that the industry with the highest emissions per kg are Technology Hardware & Equipment,

Semiconductors & Semiconductor Equipment and Software & Services, with TH&E having by far the largest carbon footprint but also the largest weight. This maybe due to them being large printers.

### Question 4

*How are emissions distributed across upstream, operations, and downstream stages in each industry?*

|  |  |
| --- | --- |
| **Column Name** | **Meaning** |
| upstream\_percent\_total\_pcf (%) | The percentage of a product's total carbon footprint (PCF) that comes from **upstream processes**, like raw material extraction or supplier emissions. |
| operations\_percent\_total\_pcf (%) | The percentage of emissions that come from **operations**, including manufacturing, packaging, and factory-level energy use. |
| downstream\_percent\_total\_pcf (%) | The percentage of emissions that come from **downstream processes**, such as transportation, customer use, and end-of-life disposal. |

**Aggregation**

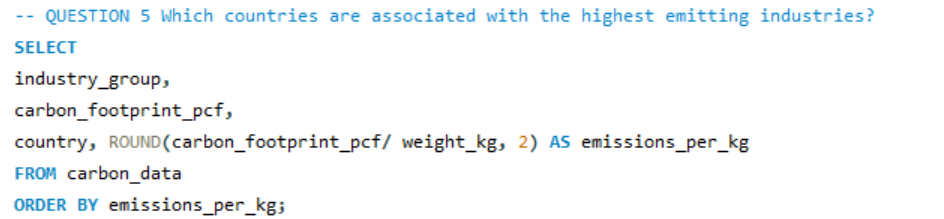
**Reporting and Documentation**

This time the data was grouped by industry, emissions, carbon footprint, upstream, operations, and downstream.

* + Capital Goods have the highest pcf from their upstream processes, because they use various types of materials. Retailing follows second and then Software and services due to chips and other software components.
  + Technology Hardware & Equipment have the highest pcf from operations, probably due to energy consumption, followed by Food, beverage and tobacco then Materials. This makes sense as the operation from all these industries are all energy intensive.
  + Technology Hardware & Equipment also have the highest downstream pcf as well followed by Materials and consumer durables.

### Question 5

*Which countries are associated with the highest emitting industries?*

**Aggregation**

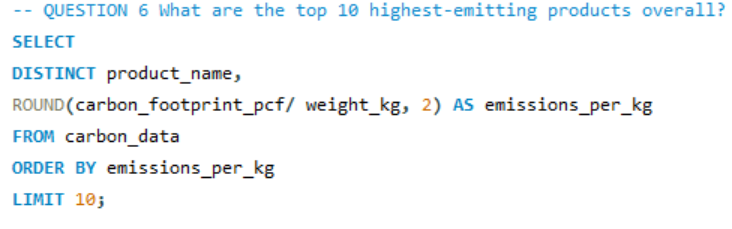
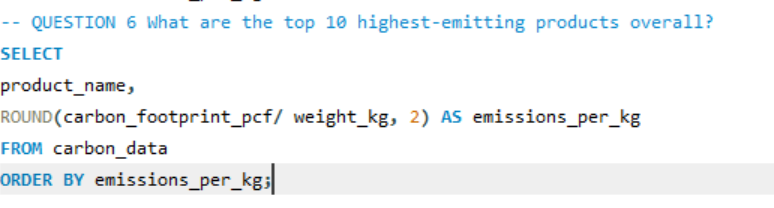
Here we see the industry measured against the carbon footprint, and the country alongside it.

**Reporting and Documentation**

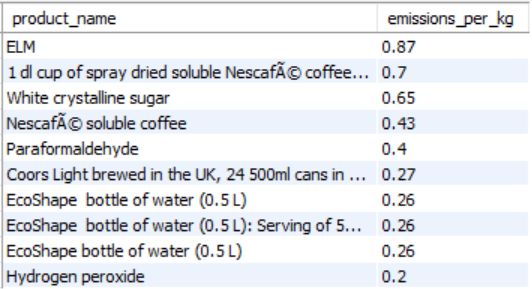
Japan, USA and Finland are high emitters due to their Technology, Hardware and Electric industry, and the Software and Services in their countries. The next closest emitter is South Korea with their Consumer Durables and Applications industries.

### Question 6

*What are the top 10 highest-emitting products overall?*

******Aggregation**

Here I grouped them by product name and emissions per kg.

**Reporting and Documentation**

These are the products that have the highest emissions per kg.

# Conclusion

This project has demonstrated how SQL and structured data can be effectively used to analyze and interpret carbon emissions at the product and industry levels. By querying and aggregating product carbon footprint (PCF) data from The Carbon Catalogue, we gained meaningful insights into how different industries and products contribute to global greenhouse gas emissions.

Our findings highlight several key trends:

* + **Materials** is the leading industry in terms of both total and average carbon emissions, indicating a high environmental impact from heavy manufacturing and construction-related goods.
  + **Technology Hardware & Equipment**, while second in total emissions, exhibits significantly high emissions from **operations and downstream stages**, likely due to energy-intensive production processes and complex logistics.
  + **Food, Beverage & Tobacco** consistently ranks in the top three, underscoring the ecological footprint of agricultural and processed consumables.
  + When emissions are analyzed **per kilogram of product**, industries such as **Semiconductors, Software & Services**, and **Technology Hardware** dominate, reflecting the high emissions intensity of compact yet complex electronic goods.
  + The distribution of emissions across lifecycle stages showed that **Capital Goods** lead in upstream emissions, while **Technology Hardware** dominates operations and downstream processes—suggesting varying sustainability challenges depending on the industry.
  + From a geographic perspective, **Japan, the USA, and Finland** emerged as countries hosting some of the most emission-heavy industries, mostly in the technology and software sectors.
  + Finally, the **top 10 highest-emitting products** include large-scale equipment, industrial supplies, and complex electronics, emphasizing the need for redesign, efficiency improvements, or alternative production methods.

By combining SQL queries with domain understanding, this analysis provides a solid foundation for identifying priority areas for emissions reduction. The methodology used here can support organizations, policymakers, and researchers in making **data-informed decisions to combat climate change** and promote more sustainable production practices across global industries.