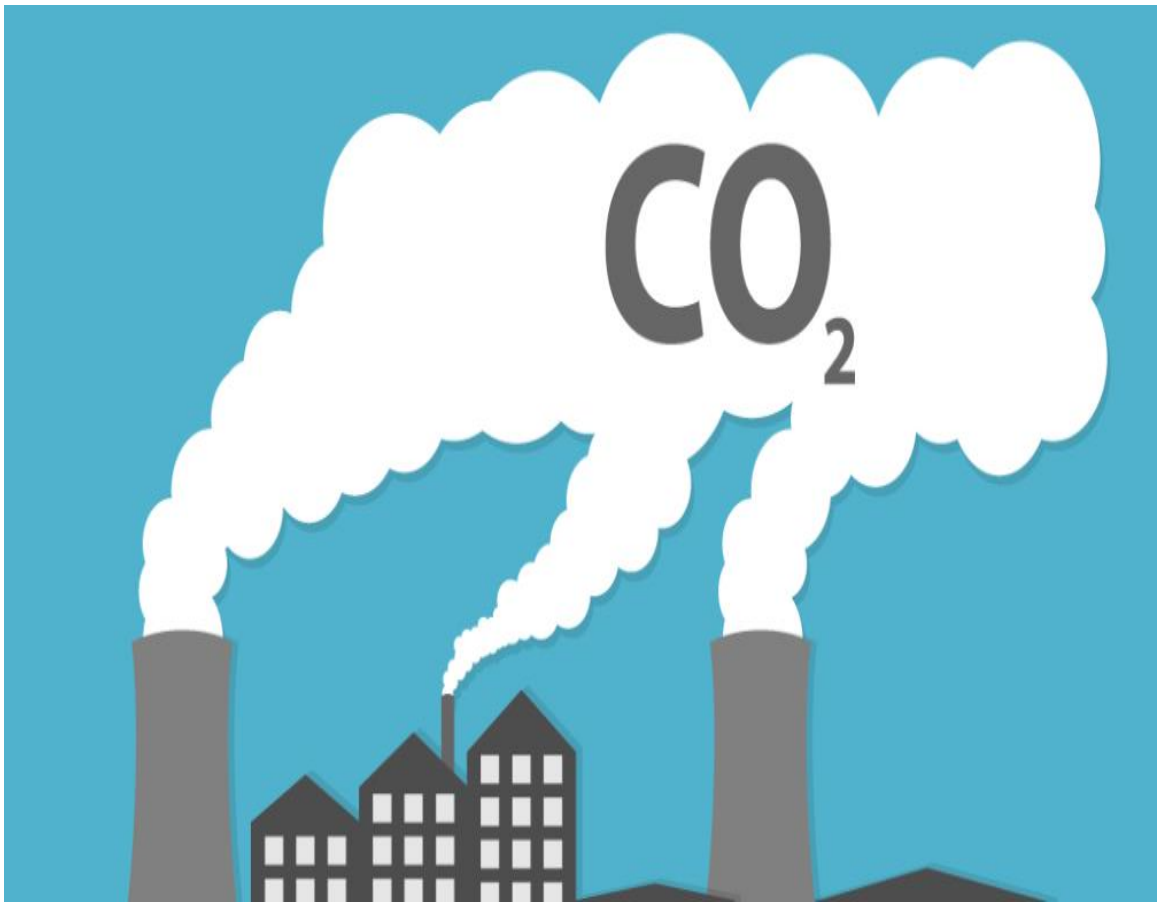

EXPLORATORY DATA ANALYSIS OF CARBON DI OXIDE EMISSION IN UNITED STATES DATASET

BY
OBIORAH CHINENYE JULIET
Data Analyst

Chapter 1 Table of Contents

Chapter 1 Introduction

Objective

This project aims at using SQL to conduct an Exploratory Data Analysis (EDA) on carbon dioxide emissions data across various U.S. states, sectors, and fuel types.
By analyzing these elements, it provides valuable insights into trends in carbon dioxide emissions over time, compare emissions across different states and sectors, and understand the contribution of various fuel types to overall emissions (Alistair, 2024).

Data Source

The data used in this analysis was sourced from Kaggle, a popular platform for data science and machine learning.

Data Overview

This dataset contains carbon dioxide emissions data for various U.S. states from 1970 to 2021. The data is broken down by state, sector (residential, commercial, transportation, electric power, and industrial), and fuel type (coal, petroleum, natural gas, and all fuels combined). The emissions values are measured in million metric tons (MMT) of carbon dioxide. Please see **Appendix A.1** for detailed explanation of the dataset.

Data Format

For this project, I downloaded and worked with a dataset in CSV format named emissions.csv, which I imported into MySQL Workbench for analysis. Using SQL, I performed data cleaning, transformations, and complex queries to extract meaningful insights, then used Excel for data visualization. This approach ensured efficient data organization and streamlined analysis.
Note: This project was designed to demonstrate my SQL skills, as per the instructions, which required using SQL for data cleaning, transformation, and analysis.

Chapter 2 Data Preparation

Data Importation

To import the CSV file into MySQL Workbench I used the following steps:

- Logged into MySQL Workbench and created a database named “emission_db” by applying the command in **Appendix B.1**.
    - Manage connection and access the database using -> **USE emission_db**
    - Create a blank table with the following command by applying the command in **Appendix B.1**:

  Note: Ensure the table structure matches the CSV file's columns.

- **Importing through Command Line Interface:** After creating the database and table, I imported the CSV file using the command line interface, as this method takes less

time to import data into the database. Steps taken for this procedure were mentioned in the A:

- Go to the command line and show directory path of the MySQL bin path by using -> cd C:\Program Files\MySQL\MySQL Server 8.0\bin
- Connect to the MySQL database using -> mysql -u root -p (root is the username and give password)
- If you are successfully logged in, set global variables so that data can be successfully imported from a local computer using -> SET GLOBAL local_infile=1;
- Quit current server connection using (mysql > quit)
- Reconnect to the MySQL server with the local-infile system variable enabled to upload data from a local machine into a file. ->mysql - -local-infile=1 -u root -p (give password)
- Show databases in the MySQL server using -> show databases;
- Connect to the database that was created for the file using -> use emission_db;
- Load the CSV file using the load data statement ->

```
LOAD DATA LOCAL INFILE 'C:\\Users\\HP\\Downloads\\emissions.csv'
INTO TABLE emissions
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\r\n' IGNORE 1 ROWS;
```

(MySQL, Oracle Corporation, 2024).

Data Cleaning and Data Wrangling

- TO avoid syntax errors, change the column name 'year' and column names with special characters '-'. See query in **Appendix B.2**.
    - Result shows that there are no duplicated values in the emissions table.
- Checked for duplicate values in all emissions table columns - see query in **Appendix B.3.**

- Checked for missing/null values in all emissions table columns using the following query in **Appendix B.4.**

- Deleted United States;

  Total emissions from all sectors and all fuel types would confuse my analysis of emissions from individual sectors and fuel types in the sector_name and fuel_name columns. Therefore, I had to remove the rows corresponding to total emissions from all sectors and all fuel types. Similarly, the data for “United States” in the state_name column, which represents a collective aggregation of emissions from individual states, would also interfere with my analysis of emissions at the state level. As a result, I had to remove the rows corresponding to “United States”. This was achieved by using the query in **Appendix B.5**.

- View unique values to check for potential input errors in state_name, sector_name, sand fuel_name columns like unnecessary spaces, mispellings etc (see **Appendix B.6**).

- Result shows that no data was entered in error

- Since "electric power carbon emissions," "industrial power carbon dioxide emissions," "commercial power carbon dioxide emissions," and "residential carbon dioxide emissions" are all sectors of carbon dioxide emissions, it is ideal to rename them to "electric power," "industrial," "commercial," and "residential," respectively, for easier readability in my visualizations. This was done by the following command (see **Appendix B.7**).

Chapter 3 Exploratory Data Analysis

Summary Statistics

Basic Summary Statistics for the overall CO_2 emission (Mean, Median, Min, Max) – See **Appendix B.8**.

SUM	**266976.6**
MEAN	7.65
MEDIAN	1.47
COUNTS	34916
MAX	234.6
MIN	0.000022

Table 1: Basic summary statistics of total CO_2 emission in million metric tons.

Observation:

This statistical summary reveals considerable variability in emissions, with the mean being significantly higher than the median, suggesting the presence of outliers or extreme emission values.

1. To compute basic summary statistics for per sector (see **Appendix B.9**)

Sector Name	**Min**	**Max**	**Mean**	**Total**
Commercial	0	30.06	1.69	12390.88
Residential	0	35.51	2.77	19394.83
Industrial	0	149.02	7.23	55316.14
Transportation	0	234.6	15.55	85546.39

Electric Power	0	160.04	12.66	94328.67

Table 2: Basic summary statistics summary of emission per sector in MMT.

Observation:

- Electric Power and Transportation sectors dominate in terms of both mean and total values, suggesting they are key focus areas for analysis or interventions.
- The Residential sector, while having a modest mean, contributes significantly overall due to the high volume of data points by the total.
- All sectors have a minimum value of **0**, indicating that there are instances with no emissions or activity recorded for some data points.

2. To compute basic summary statistics for per fuel type (see **Appendix B.10**)

Fuel Name	Min	Max	Mean	Total
Natural Gas	0	149.02	4.82	62083.88
Coal	0	160.42	9.8	86130.86
Petroleum	0	234.6	8.7	118762.16

Table 3: Basic summary statistics of emission per fuel type in MMT.

Observation:

- Petroleum stands out with the highest maximum and total values, emphasizing its importance in the dataset.
- Coal has the highest average use per instance, even though its total is less than that of Petroleum CO_2 emission.
- Natural Gas shows lower overall CO_2 emission compared to Coal and Petroleum, both in terms of average and total values.

Insights and Findings

3.1.1 Emission Trend Across Years

The SQL query used to calculate the total amount of carbon di oxide emissions in million metric tons (MMT) and observe the trend is found in the **Appendix B.11**.

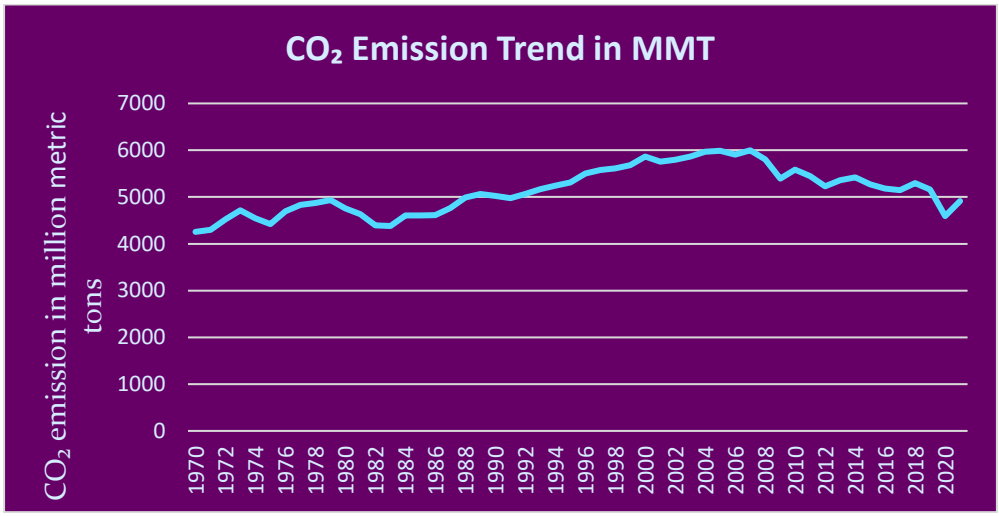

Figure 1: Total emission of carbon dioxide in MMT across years (from 1970 – 2021).

Observation:

CO_2 emissions gradually increased from 4254.94 MMT in 1970 to a peak of about 6000 MMT around 2000. Trend showed limited growth with minor fluctuation in emission in 2000 to 2010, emissions dropped after 2010, clearly in 2020.

3.1.2 How emission of Carbon dioxide emission in million metric tons (MMT) changed over decades see (Appendix B.12).

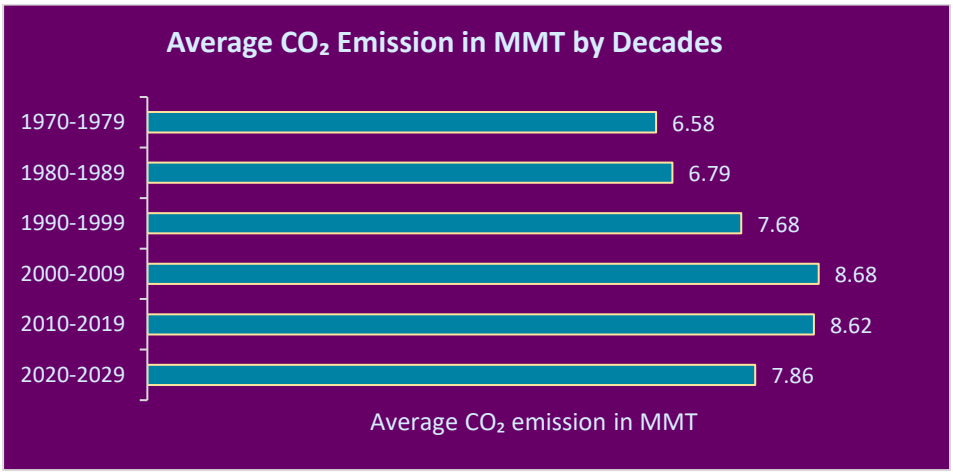

Figure 3: Average CO_2 emission in MMT by Decades

Observation:

- The overall trend shows a rise in CO_2 emissions from the 1970s to 2000s, peaking in the 2000-2009 decade, followed by a slight decline in the 2010-2019 and 2020-2029 decades.
- Comparing to the current (2020 -2029) decade, CO_2 emissions in the current decade are significantly higher than in the 1970s and 1980s showing that emissions have increased slightly by 0.18 MMT while CO_2 emission in the current decade decreased by 0.82 MMT as compared with the emission in 2000s and 2010.

- Since the decade is not yet completed, the analysis for the current decade may change as more data becomes available. If the observed trend continues, the average CO_2 emissions for 2020-2029 could increase, stabilize or further decline.

3.1.3 To calculate the percentage contribution of carbon dioxide in million metric tons (MMT) emission by each sector in all decades (see Appendix B.13).

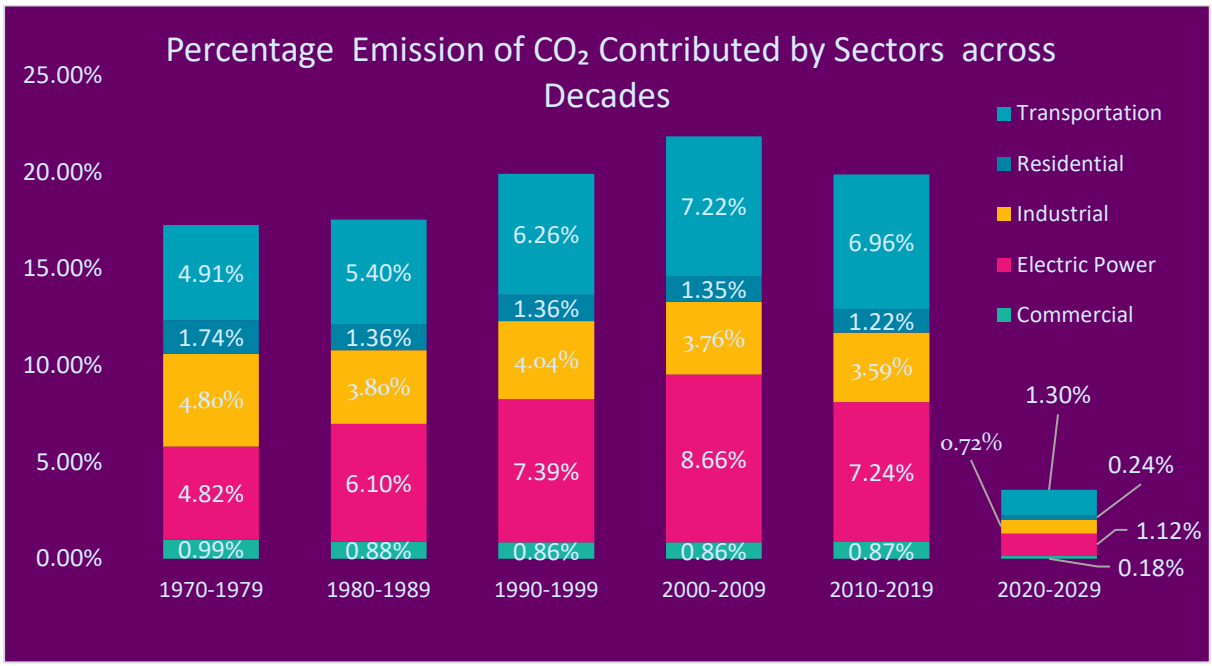

Figure 3: Percentage contribution of CO_2 emissions by sectors across different decades.

Observation:

- Electric Power has been the largest contributor to CO_2 emissions across decades, reaching a peak of 8.66% in 2000-2009 before showing a decline in recent years.
- Transportation emissions steadily increased from 4.91% in the 1970s to a peak of 7.22% in the 2000s but saw a slight reduction to 6.96% in 2010-2019.
- The Residential and Commercial sectors have consistently made relatively small contributions, remaining below 2% across all decades.
- Industrial emissions peaked at 4.80% in the 1970s and gradually declined, reaching 1.12% so far in the current decade.
- For the ongoing decade (2020-2029), there has been a notable drop in emissions across all sectors, with Transportation ranking the highest contributor at 1.30%, followed by Industrial at 1.12%, Electric Power at 0.72%, Commercial at 0.24% and Residential as the lowest contributor at 0.18%.
- These reductions may reflect incomplete data, as the decade is not yet complete, and the final trends could change as more information becomes available.

3.1.4 Comparing the above result with the percentage contribution of carbon dioxide in million metric tons emission between 2020 and 2021.

see **Appendix B.14** for the SQL commands.

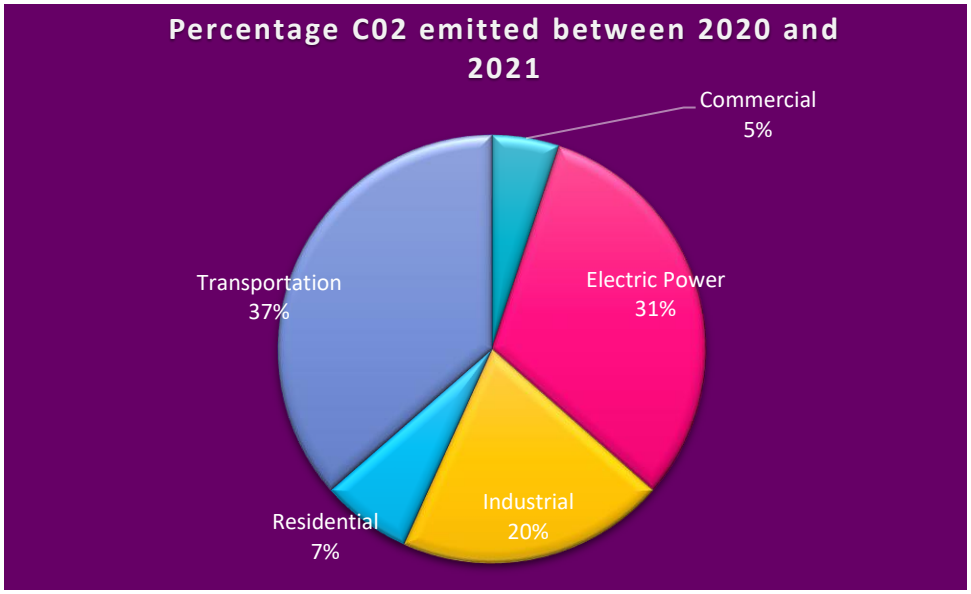

Figure 4: Percentage Contribution of CO_2 emissions by sectors in 2020-2021.

Observation:

- The comparison between CO_2 emissions in 2020-2021 and the current decade (2020-2029) reveals significant differences, likely due to incomplete data for the decade.
- In 2020-2021, Transportation contributed 37%, a much higher figure than the 1.30% previously reported for the current decade, reinforcing its role as a major emitter.
- Similarly, Electric Power accounted for 31% in 2020-2021, while the earlier decade figure was only 0.72%, highlighting a large discrepancy.
- The Industrial sector contributed 20% in the 2020-2021 period compared to 1.12% in the decade's earlier data, again reflecting incomplete reporting.
- Residential and Commercial emissions also show notable differences, with 7% and 5% in 2020-2021 compared to 0.18% and 0.24% respectively for the current decade.
- Overall, the 2020-2021 data confirms that Transportation and Electric Power remain the largest contributors to CO_2 emissions, with significant shares from the Industrial, Residential, and Commercial sectors.
- The much lower values reported so far for the 2020-2029 decade suggest incomplete data, and it is expected that the full decade's trends will align more closely with the proportions observed in the 2020-2021 period as additional data becomes available.

3.1.5 To compute the percentage contribution of carbon di oxide in million metric tons emission by fuel type in all decades.

See **Appendix B.15** for the SQL commands.

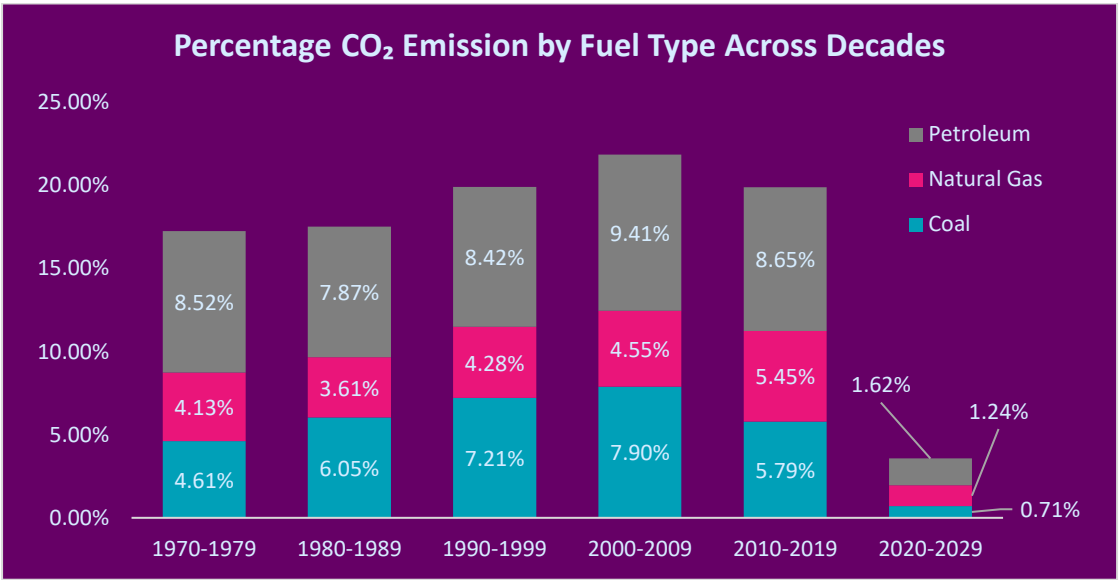

Figure 5: Percentage contribution of CO_2 emissions by fuel type across different decades.

Observation:

- The analysis of CO_2 emissions by fuel type across decades highlights notable trends in energy transitions and emissions reduction.
- Petroleum remained the largest contributor over the decades, reaching a peak of 9.41% in 2000-2009 but saw a sharp decline to 1.62% in the current decade.
- Coal emissions, which started at 4.61% in the 1970s and peaked at 7.90% in 2000-2009, declined sharply to just 0.71% in the current decade.
- Similarly, natural gas emissions gradually increased from 4.13% to a peak of 5.45% in 2010-2019 before dropping to 0.24% in 2020-2029.
- While there are drastic reductions observed for all fuel types in 2020-2029 these values cannot yet be fully ascertained due to incomplete data availability for the current decade. As the decade is still ongoing, the trends may shift, and the final emissions data could more accurately reflect the contributions of each fuel type.

3.1.6 Comparing the above result with the percentage contribution of carbon di oxide in million metric tons emission by fuel type between 2020 and 2021.

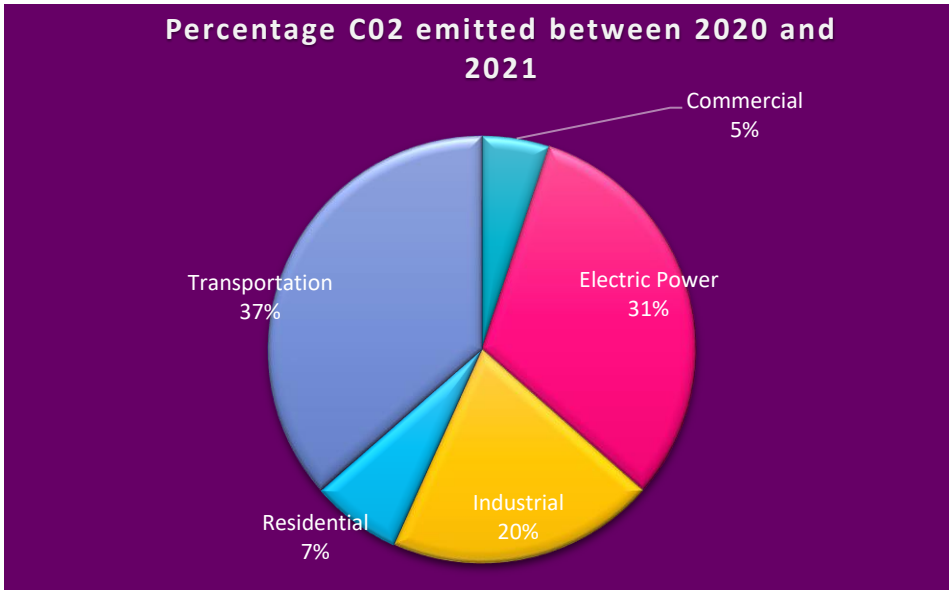

Figure 6: Percentage contribution of CO_2 emissions by fuel type from 2020 to 2021.

Observation:

- In the earlier charts for 2020-2029, emissions for all major fuel types coal, petroleum, and natural gas showed substantial declines. Coal dropped to 0.71%, petroleum to 1.62%, and natural gas to 0.24%.
- However, the 2020-2021 pie chart paints a different picture, petroleum still remains the dominant contributor accounting for 45%.
- Natural gas contributed to 35% of CO_2 emissions and coal, while reduced, still contributes 20% of CO_2 emissions.

3.1.7 Contribution of CO_2 emission in each sector by fuel types.

See **Appendix B.17** for the SQL command.

Sectors	Coal	Natural Gas	Petroleum
Commercial	0.19%	3.06%	1.39%
Electric Power	27.61%	5.68%	2.04%
Industrial	4.40%	8.75%	7.57%
Residential	0.06%	5.04%	2.17%
Transportation	0.00%	0.72%	31.32%

Table 4: percentage distribution of CO_2 emission across various sectors by fuel type.

Observation:

- The Transportation sector dominates the entire CO_2 emissions from petroleum, contributing a significant 31.32%.
- The Electric Power sector is the second largest emitter of CO_2 from coal, contributing 27.61% of total emissions
- Other sectors have minimal contributions from coal, with the Industrial sector following at 4.40% and the Commercial and Residential sectors contributing less than 0.20% each.
- Transportation does not contribute to CO_2 emissions from coal.

- Natural gas emissions are highest in the Industrial sector accounting for 8.75% of total emission, with the Residential and Electric Power sectors also making notable contributions.
- The Commercial and Transportation sectors account for smaller shares of emissions from this fuel type.
- The Industrial sector is the second-highest contributor at 7.57%, while other sectors contribute

Chapter 4: Conclusion

Limitations

Recent Discrepancies and Incomplete Data:
The sharp decline observed in emissions for the 2020-2029 period cannot yet be conclusively attributed to energy transition, as the data for the current decade remains incomplete. These values cannot yet be fully ascertained due to incomplete data availability for the current decade as the decade is still ongoing, the trends may shift. However, the specific data for 2020-2021 shows significantly higher emissions, suggesting that reliance on fossil fuels has persisted in the short term.

Key Highlights

- CO_2 emissions in the current decade are significantly higher than in the 1970s and 1980s
- District of Columbia recorded the most substantial emission reduction at 82% while New York ranked second largest emission reduction by 45% and Massachusetts with 44% reduction as the third.
- North Dakota experienced the highest increase in CO_2 emissions from 1970 to 2021, with a staggering 283% rise. Nevada and Alaska followed closely, with emissions increasing by 266% and 242%, respectively.
- Transportation and Electric Power remain the dominant sources of CO_2 emissions.
- The Electric Power sector is heavily reliant on coal, which is a key area to target for emission reduction strategies.
- Transportation primarily contributes through petroleum emissions.
- Natural gas has a more balanced distribution, with significant emissions across Industrial, Residential, and Electric Power sectors.

Recommendation

One of the major objectives of analyzing the emission of CO_2 is to understand and manage their impact on the environment. By identifying trends, sources, and sectors contributing to emissions, we can:

- Generate data-driven insights help encourage the adoption of clean energy, efficient transportation, and sustainable industrial practices
- Use data analytics to continuously monitor CO_2 emissions across sectors and decades.

- Discover industries or regions with the highest emissions helps focus efforts on areas with the most significant environmental impact.
- Reducing CO_2 emissions is critical to slowing global warming.

In 2010-2019 average annual global greenhouse gas emissions were at their highest levels in human history, but the rate of growth has slowed. Without immediate and deep emissions reductions across all sectors, limiting global warming to 1.5°C is beyond reach. Limiting global warming will require major transitions in the energy sector (Intergovernmental Panel on Climate Change (IPCC), 2022). They involved the following.

- Agriculture, forestry, and other land use can provide large-scale emissions reductions and also remove and store carbon dioxide at scale (Intergovernmental Panel on Climate Change (IPCC), 2022).

- A substantial reduction in fossil fuel use, widespread electrification, improved energy efficiency, and use of alternative fuels such as hydrogen (Intergovernmental Panel on Climate Change (IPCC), 2022).

- Having the right policies, infrastructure and technology in place to enable changes to our lifestyles and behaviour can result in a 40-70% reduction in greenhouse gas emissions by 2050 (Intergovernmental Panel on Climate Change (IPCC), 2022).

- Reducing emissions in industry will involve using materials more efficiently, reusing and recycling products and minimizing waste (Intergovernmental Panel on Climate Change (IPCC), 2022).

- Cities and other urban areas also offer significant opportunities for emissions reductions. These can be achieved through lower energy consumption (such as by creating compact, walkable cities), electrification of transport in combination with low-emission energy sources, and enhanced carbon uptake and storage using nature (Intergovernmental Panel on Climate Change (IPCC), 2022).

References

Alistair, King. (2024, April 20). *CO_2 Emissions (in U.S.)*. Retrieved 12 13, 2024, from Kaggle: www.kaggle.com/datasets/alistairking/u-s-CO_2 -emissions

Intergovernmental Panel on Climate Change (IPCC). (2022, April 04). *IPCC PRESS RELEASE - The evidence is clear: the time for action is now. We can halve emissions by 2030*. Retrieved 12 17, 2024, from IPCC: https://www.ipcc.ch/site/assets/uploads/2022/04/IPCC_AR6_WGIII_PressRelease_English.pdf

MySQL, Oracle Corporation. (2024, December 06). *Character Set and Collation*, 8.0. Retrieved 12 15, 2024, from MySQL 8.0 Reference Manual: https://dev.mysql.com/doc/refman/8.0/en/charset.html

MySQL, Oracle Corporation. (2024, December 06). *LOAD DATA Statement*. (Oracle Corporation) Retrieved from MySQL 8.0 Reference Manual: https://dev.mysql.com/doc/refman/8.0/en/load-data.html

Appendices

Appendix A: General Dataset overview

Appendix A.1: Dataset Overview

Columns

1. year: The year for which the emissions data is provided (e.g., 1970).
2. state-name: The name of the U.S. state (e.g., Alabama, Alaska, Arizona).
3. sector-name: The sector for which the emissions data is provided. The sectors include:
    - Residential carbon dioxide emissions
    - Commercial carbon dioxide emissions
    - Transportation carbon dioxide emissions
    - Electric Power carbon dioxide emissions
    - Industrial carbon dioxide emissions
    - Total carbon dioxide emissions from all sectors

4. fuel-name: The type of fuel contributing to the carbon dioxide emissions. The fuel types include:
    - Coal
    - Petroleum
    - Natural Gas
    - All Fuels (representing the total emissions from all fuel types combined)

5. value: The carbon dioxide emissions value in million metric tons for the specified year, state, sector, and fuel type (Alistair, 2024).

Appendix A.2: Dataset Importance

This dataset can be used in cases where:

- Researchers and policymakers can use this data to study the effectiveness of emissions reduction policies and identify areas for improvement (Alistair, 2024).
- Environmentalists and advocacy groups can leverage this information to raise awareness about the carbon footprint of different states and sectors, and push for cleaner energy solutions (Alistair, 2024).
- Businesses and investors can utilize this data to assess the environmental sustainability of various industries and make informed decisions about investments and partnerships (Alistair, 2024).
- Educators and students can explore this dataset to learn about the factors contributing to carbon dioxide emissions and develop projects or research papers on related topics (Alistair, 2024).

Appendix B: SQL Queries and Definition.

Appendix B.1: Table and Database Creation

```
-- [1] -- Database Creation
DROP DATABASE IF EXISTS emission_db;
CREATE DATABASE IF NOT EXISTS emission_db
CHARACTER SET utf8mb4
COLLATE utf8mb4_general_ci;
```

Explanation:

- The following command drops command deletes the database named emission_db if it exists. Using IF EXISTS prevents an error from occurring if the database does not exist. Creates a new database named emission_db if it does not already exist.
- The statement CHARACTER SET utf8mb4 COLLATE utf8mb4_general_ci; is used in MySQL to define the character set and collation for a database or table. UTF8MB4 allows for the storage of a wide range of characters, including emojis and other symbols, as it fully supports Unicode. The utf8mb4_general_ci collation specifies a case-insensitive comparison of strings (MySQL, Oracle Corporation, 2024).

```
 --[2]—Table Creation
DROP TABLE IF EXISTS emissions;
CREATE TABLE emissions (
`year` INT,
`state-name` VARCHAR (20),
`sector-name` VARCHAR (60),
`fuel-name` VARCHAR (15),

`value` DECIMAL (15,8)
);
```

Explanation:

- This command deletes the table named **emissions** if it exists. It ensures that there are no errors if the table is not present Iand creates a new table called **emissions** with the following column types:
    - VARCHAR: A variable-length string (number of characters) representing the state, sector and fuel names.
    - INT: Integer representing the year
    - Decimal: Used when creating decimal values in SQL. Syntax for defining a decimal column -> DECIMAL (precision, scale) where precision is the total number of digits that can be stored, both to the left and right of the decimal point (inclusive) while the scale is the number of digits that will be stored to the right of the decimal point.

Why Choose Backticks for column names?
Since year is a reserved keyword in SQL (used for date-related functions), and column names that contain special characters (‘- ‘) enclosing it in backticks avoids syntax errors. Without backticks, the database might misinterpret year as a keyword or function rather than a column name or table name.

Appendix B.2 - Changing Column name:

```
ALTER TABLE emissions
CHANGE COLUMN `year` `emission_year` INT,
CHANGE COLUMN `state-name` `state_name` TEXT,
CHANGE COLUMN `sector-name` `sector_name` TEXT,
CHANGE COLUMN `fuel-name` `fuel_name` TEXT,
CHANGE COLUMN `value` `emission_value` DOUBLE;
```

Appendix B.3 - Check for duplicate using the following commands below:

```
SELECT emission_year,
 state_name,
 sector_name,
fuel_name,
 emission_value,
COUNT (*)
FROM emissions
GROUP BY emission_year, state_name, sector_name, fuel_name, emission_value
HAVING COUNT (*) > 1;
```

Explanation:

- The SQL query retrieves data on emissions, grouping by year, state, sector, fuel type, and the emission value while counting occurrences. It filters results to include only those groups that appear more than once, indicating duplicate records in the dataset.
- Result shows that there are no duplicated values in the emissions table.

Appendix B.4 - Checking for null/missing values

```
SELECT
        *
FROM emissions
WHERE emission_year IS NULL OR state_name IS NULL OR sector_name IS NULL OR fuel_name
IS NULL OR emission_value IS NULL;
```

Explanation:

- This SQL query you've written retrieves all records from the emissions table where any of the specified fields is null.

Appendix B.5 Deleting unwanted rows:

```
DELETE FROM emissions
WHERE sector_name LIKE ‘%all sectors%’ OR state_name LIKE ‘%united%’ OR fuel_name LIKE
‘%ALL%';
```

Explanation:

DELETE FROM emissions: This deletes rows from the table named emissions.

WHERE sector_name LIKE '%all sectors%': Deletes rows where the sector_name contains the substring "all sectors".
OR state_name LIKE '%united%': Deletes rows where the state_name contains the substring "united".
OR fuel_name LIKE '%All%': Deletes rows where the fuel_name contains the substring "All".

Appendix B.6 - Checking for unique values to ensure that no data was entered in error:

- View unique values for state name

```
SELECT
DISTINCT state_name
FROM emissions;
```

- View unique values for sector_name

```
SELECT
DISTINCT sector_name
FROM emissions;
```

- View unique values for fuel name

```
SELECT
DISTINCT fuel_name
FROM emissions;
```

Appendix B.7 —Renaming rows in the sector_name field:

```
UPDATE emissions
SET sector_name = CASE
WHEN sector_name = 'Electric Power carbon dioxide emissions' THEN 'Electric Power'
        WHEN sector_name = 'Residential carbon dioxide emissions' THEN 'Residential'
    WHEN sector_name = 'Transportation carbon dioxide emissions' THEN 'Transportation'
        WHEN sector_name = 'Industrial carbon dioxide emissions' THEN 'Industrial'
    WHEN sector_name = 'Commercial carbon dioxide emissions' THEN 'Commercial'
    ELSE sector_name
     END;
```

Explanation:

- The “CASE” statement allows for conditional logic within the UPDATE. Each condition checks the sector_name, and if it matches one of the specified phrases, it changes it to a more simplified name while the “ELSE” clause ensures that any sector_name not matching the specified conditions will remain unchanged.

Appendix B.8 - Basic summary statistics for overall CO_2 emission:

```
SELECT
        ROUND (SUM (emission_value),2) AS `SUM`,
         ROUND (AVG (emission_value),2) AS "MEAN",
         COUNT (emission_year) AS `COUNTS`,
          ROUND (MAX (emission_value),2) AS `MAX`,
```

```
        MIN (emission_value) AS `MIN`
FROM emissions;
```

- This query offers statistics snapshot of the emissions data, allowing for quick analysis on total emissions, average emissions, and extremes (min/max). The use of ROUND for sums and averages ensures that the output is presented neatly.

Most SQL databases do not provide a built-in MEDIAN function. However, the median was calculated using a combination of SQL window functions or aggregation. Here's how it was achieved:

```
    WITH CTE AS (
            SELECT
                emission_value,
                ROW_NUMBER() OVER (ORDER BY emission_value) AS
row_num,
                COUNT(*) OVER () AS total_rows
             FROM emissions
)
SELECT
        ROUND (CASE
        WHEN total_rows % 2 = 1 THEN
                    (SELECT emission_value
                     FROM CTE
                    WHERE row_num = FLOOR (total_rows / 2) + 1)
        ELSE
            (SELECT AVG(emission_value)
             FROM CTE
            WHERE row_num IN (total_rows / 2, (total_rows / 2) + 1))
       END,2) AS `Median`
FROM CTE
LIMIT 1;
```

The explanation of the code snippet above is as follows:

- CTE Definition:
  - "The WITH CTE AS ()" clause creates a CTE named CTE, which includes:
  - emission_value: The actual emission value from the emissions table.
  - ROW_NUMBER () OVER (ORDER BY emission_value) AS row_num: This calculates the row number for each emission value which is ordered by emission_value.
  - COUNT (*) OVER () AS total_rows: This counts the total number of rows in the CTE, giving the total number of emission values.

- Calculating the Median:

The main SELECT statement:

- Uses a CASE statement to determine how the median should be calculated based on whether total_rows is odd or even:

Odd Total Rows: If total_rows is odd, it selects the middle value:

- SELECT emission_value FROM CTE WHERE row_num = FLOOR (total_rows / 2) + 1

Even Total Rows: If total_rows is even, it calculates the average of the two middle values:

- SELECT AVG (emission_value) FROM CTE WHERE row_num IN (total_rows / 2, (total_rows / 2) + 1)

The ROUND (..., 2) function is applied to round the result to two decimal places.
Limiting the Result:

- LIMIT 1: Ensures that only one result is returned, which is the calculated median.

- Final Output:

The result of this query is a single value representing the median of the emission_value column in the emissions table, rounded to two decimal places.

Appendix B.9 - To compute basic summary statistics (min, max, mean, and total) per sector.

```
SELECT
   sector_name,                    -- Group results by sector_name
   ROUND(MIN(emission_value),2) AS “min”,   -- Minimum emission value for each sector
   ROUND(MAX(emission_value),2) AS “max”,  -- Maximum emission value for each sector
   ROUND(AVG(emission_value),2) AS “mean,  --Average emission value for each sector
   ROUND(SUM(emission_value),2) AS total     -- Total emission value for each sector
FROM emissions
GROUP BY sector_name            -- Group data by the sector_name column
ORDER BY total_value;        -- Sort results by total_value in ascending order
```

- **ROUND e**nsures the results are rounded to 2 decimal places for better readability.

Appendix B.10 - To compute basic summary statistics (min, max, mean, and total) per fuel type.

```
SELECT
   fuel_name,                    -- Group results by fuel_name
   ROUND(MIN(emission_value),2) AS “min”, -- Minimum emission value for each fuel
   ROUND(MAX(emission_value),2) AS “max”,  -- Maximum emission value for each fuel
   ROUND(AVG(emission_value),2) AS “mean”,  --Average emission value for each fuel
   ROUND(SUM(emission_value),2) AS “total”     -- Total emission value for each fuel
FROM emissions
GROUP BY fuel_name                -- Group data by the fuel_name column
```

```
ORDER BY total_value;        -- Sort results by total_value in ascending order
```

- **ROUND e**nsures the results are rounded to 2 decimal places for better readability.

Appendix B.11 - Emission Trend Across Years

```
SELECT emission_year,
                ROUND (SUM (emission_value),2) AS "total_value"
FROM emissions
GROUP BY emission_year;
```

Explanation:

- SELECT emission_year:
  - Selects the emission_year column from the emissions table. This is the year of emissions, and the results will be grouped by year.
- ROUND(SUM(emission_value), 2) AS "total_value":
  - SUM(emission_value): Adds up the emission_value for each emission_year (aggregates emissions within each year).
  - ROUND(..., 2): Rounds the summed emission values to two decimal places.
  - The result is labeled as "total_value".
- FROM emissions:
  - Specifies the emissions table as the data source.
- GROUP BY emission_year:
  - Groups the data by emission_year, so the total emission value is calculated for each distinct year.

Appendix B.12 - How emission of Carbon dioxide emission in million metric tons (MMT) changed over decades.

```
SELECT
CONCAT(FLOOR(emission_year/10) * 10, "-", FLOOR (emission_year/10) * 10 + 9) AS "decade",
        ROUND (AVG (emission_value),2) AS "Average_emission"
FROM emissions
GROUP BY decade;
```

Explanation:

- CONCAT(FLOOR(emission_year/10) * 10, "-", FLOOR(emission_year/10) * 10 + 9) AS "decade":
  - FLOOR(emission_year/10) * 10: This operation divides the emission_year by 10, rounds it down (using FLOOR), and then multiplies by 10 to get the starting year of the decade.
  - FLOOR(emission_year/10) * 10 + 9: This adds 9 to the start year to get the ending year of the decade.

- CONCAT(...): Combines the start and end year of the decade into a string (e.g., "1990-1999").
- The result is labeled as "decade".

- ROUND(AVG(emission_value), 2) AS "Average_emission":
  - AVG(emission_value): Calculates the average of emission_value for each decade.
  - ROUND(..., 2): Rounds the average emission value to 2 decimal places.
  - The result is labeled as "Average_emission".
- FROM emissions:
  - Specifies the emissions table as the data source.
- GROUP BY decade:
  - Groups the data by the calculated decade, so that the average emission value is computed for each decade**.**

Appendix B.13 - To calculate the percentage contribution of carbon dioxide in million metric tons (MMT) emission by each sector in all decades.

```
SELECT
    CONCAT (FLOOR (emission_year/10) * 10, "-", FLOOR (emission_year/10) * 10 + 9) AS
    "decade",
    sector_name,
    CONCAT (ROUND (ROUND (SUM (emission_value)/ (SELECT SUM (emission_value)
    FROM emissions), 4) *100,2),'%') AS "percentage emission"
FROM emissions GROUP BY decade, sector_name;
```

Explanation:

- CONCAT (FLOOR (emission_year/10) * 10, "-", FLOOR (emission_year/10) * 10 + 9) AS "decade":
  - It groups data into decades by rounding the emission_year to the nearest decade (e.g., "1990-1999").see Appendix B.12 for detailed explanation.

- CONCAT (ROUND (ROUND (SUM (emission_value) / (SELECT SUM (emission_value) FROM emissions), 4) * 100, 2), '%') AS "percentage emission":
  - SUM(emission_value): Calculates the total emissions for each sector in a given decade.
  - SELECT SUM(emission_value) FROM emissions: This subquery calculates the total emissions across all sectors and years.
  - ROUND(..., 4): Divides the sector's total emission by the overall total emissions to get the fraction.
  - * 100: Converts the fraction into a percentage.
  - ROUND(..., 2): Rounds the percentage to 2 decimal places.
  - CONCAT(..., '%'): Appends the '%' sign to the calculated percentage.
  - The result is aliased as "percentage emission".
- FROM emissions:
  - Specifies the emissions table as the data source.

- GROUP BY decade, sector_name:
    - Groups the data by decade and sector_name, so the query calculates emissions and their percentages for each sector within each decade.

Appendix B.14 - Comparing the above result with the percentage contribution of carbon dioxide in million metric tons emission between 2020 and 2021.

```
SELECT
        sector_name,
        CONCAT (ROUND (ROUND (SUM (emission_value)/ (SELECT SUM (emission_value)
        FROM emissions WHERE emission_year BETWEEN 2020 AND 2021),4) *100,2),'%') AS
        "percentage emission"
FROM emissions
WHERE emission_year BETWEEN 2020 AND 2021
GROUP BY sector_name;
```

Explanation:

- sector_name:
    - Selects the sector_name column to group emissions by sector.
- CONCAT(ROUND(ROUND(SUM(emission_value) / (SELECT SUM(emission_value) FROM emissions WHERE emission_year BETWEEN 2020 AND 2021), 4) * 100, 2), '%') AS "percentage emission":
    - SUM(emission_value): Sums the emission_value for each sector in the years 2020 and 2021.
    - SELECT SUM(emission_value) FROM emissions WHERE emission_year BETWEEN 2020 AND 2021: This subquery calculates the total emissions in the years 2020 and 2021 for all sectors.
    - ROUND(..., 4): Divides the sector's total emissions by the total emissions from the subquery to get the fraction.
    - * 100: Converts the fraction into a percentage.
    - ROUND(..., 2): Rounds the percentage to two decimal places.
    - CONCAT(..., '%'): Adds a % symbol to the rounded percentage.
    - The result is aliased as "percentage emission".
- FROM emissions:
    - Specifies the emissions table as the data source.
- WHERE emission_year BETWEEN 2020 AND 2021:
    - Filters the data to include only the years 2020 and 2021.
- GROUP BY sector_name:
    - Groups the data by sector_name to calculate the emissions and percentage for each sector separately.

Appendix B.15 - To compute the percentage contribution of carbon di oxide in million metric tons emission by fuel type in all decades

```
SELECT
        CONCAT (FLOOR (emission_year/10) * 10, "-", FLOOR (emission_year/10) * 10 + 9) AS
        "decade",
```

```
    fuel_name,
    CONCAT (ROUND (ROUND (SUM (emission_value)/ (SELECT SUM (emission_value)
    FROM emissions),4) *100,2),'%') AS "percentage emission"
FROM emissions
GROUP BY decade, fuel_name;
```

- This SQL command has already been explained in **Appendix B.13** of this document. The only difference here is that instead of grouping by **sector_name**, the query groups by **fuel_name**.

Appendix B.16 – Comparing the above result with the percentage contribution of carbon di oxide in million metric tons emission by fuel type between 2020 and 2021

```
SELECT
    fuel_name,
    CONCAT (ROUND (ROUND (SUM (emission_value)/(SELECT SUM(emission_value) FROM
    emissions WHERE emission_year BETWEEN 2020 AND 2021),4)*100,2),'%') AS
    "percentage emission"
FROM emissions
WHERE emission_year BETWEEN 2020 AND 2021
GROUP BY fuel_name;
```

- This SQL command has already been explained in **Appendix B.14** of this document. The only difference here is that instead of grouping by sector_name, the query groups by fuel_name**.**

Appendix B.17 – Contribution of CO_2 emission in each sector by fuel types.

```
SELECT
    state_name,
    ROUND(SUM(emission_value),2) AS "total_value"
FROM emissions
GROUP BY state_name
ORDER BY total_value desc;
```