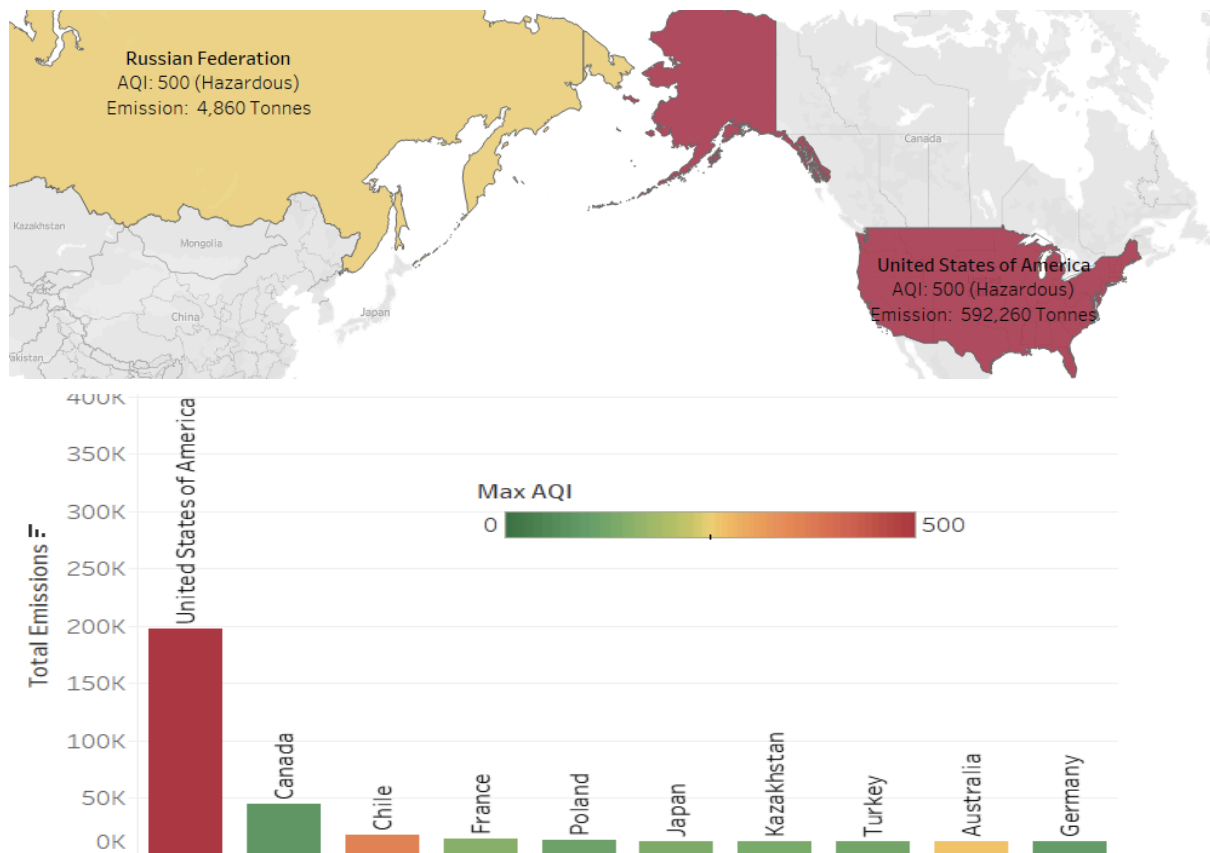


# Project Summary

This project focuses on the analysis of air quality and pollutant emissions to understand their impact on environmental sustainability and public health by utilising two datasets detailing the Air Quality Index (AQI) by city and emissions of air pollutants by country. The primary objectives of this project were to identify the most harmful types of pollutants and to investigate the relationship between emissions and air quality across various countries. Both datasets were thoroughly inspected for missing values and inconsistencies. Necessary adjustments, such as adding a foreign key, extracting data snippets, and reformatting data, were performed. The datasets were then joined in an SQLite database, leveraging the common attribute of country codes. Specific SQL queries were then designed to answer the project's objectives.

The key findings from this project are that Carbon Monoxide (CO), Particulate Matter (PM10), and Non-Methane Volatile Organic Compounds (NMVOC) were among the most harmful pollutants, with CO being the most harmful in its 66% contribution to worsening the overall air quality. As for the relationship between emission and air quality, there were some expected correlations between countries that were high emitters such as the USA and their reflective hazardous AQI. However, there were also outliers such as Russia which emitted less but had bad air quality, or Canada which was a high emitter but did not have a bad air quality.



However, there were limitations within the data sources that may have led to inaccurate findings. Among these include missing data, data inconsistencies, and differing scopes.

# Wrangling Details

## DATA SOURCE 1

The first data source used in this report is the **World Air Quality Index by City and Coordinates** dataset. It is compiled on [Kaggle by Aditya Ramachandran](#) for the purposes of creating a comprehensive dataset on cities, latitude, longitude, and pollution levels. The dataset was downloaded in CSV format.

Upon inspection, the dataset contains 16695 rows in the following columns:

Column Name	Description
Country	Name of the country in which the values were measured
City	Name of the city in which the values were measured
AQI Value	The measured air quality index
AQI Category	Quantitative categories of the AQI values, including; Good = 0-50 Moderate = 51-99 Unhealthy for Sensitive Groups = 100-149 Unhealthy = 150-200 Very Unhealthy = 201-300 Hazardous = 301-500
CO AQI Value	The measured AQI in regards to carbon monoxide levels
CO AQI Category	Quantitative categories of the carbon monoxide AQI values. The range of AQI values for each category was unclear
Ozone AQI Value	The measured AQI in regards to ozone levels
Ozone AQI Category	Quantitative categories of the Ozone AQI values. The range of AQI values for each category was unclear
NO2 AQI Value	The measured AQI in regards to nitrogen dioxide levels
NO2 AQI Category	Quantitative categories of the nitrogen dioxide AQI values. The range of AQI values for each category was unclear
PM2.5 AQI Value	The measured AQI in regards to PM2.5 levels
PM2.5 AQI Category	Quantitative categories of the PM2.5 AQI values, including; Good = 0-50 Moderate = 51-99 Unhealthy for Sensitive Groups = 100-149 Unhealthy = 150-200 Very Unhealthy = 201-300 Hazardous = 301-500
lat	The latitude of the measured location
lng	The longitude of the measured location

Upon further inspection, some “Country” values were missing. There was an attempt to address this issue by matching the latitude and longitude values of rows with missing “Country” values to a latitude and longitude master list using VLOOKUP. However, this method did not prove to be accurate, and using an API to automatically fill in the missing values would be too time-consuming. An executive decision was made to exclude these values.

The original CSV file was modified to become an XLSX file for easier manipulation, as there was a need to add a foreign key column for future joining purposes. The foreign key choice was the alpha-3 country code (which was used in the second data source). This was added by obtaining the country codes from [IBAN](#) and then using VLOOKUP. Below is the Excel function used for the new column titled “Country Code”. TRUE was used instead of FALSE due to the fact that some of the country names in the data source did not exactly match the external table from IBAN (e.g., “United States of America (the)” verses “United States of America”). The rows that were missing “Country” values displayed #N/A as expected.

```
=VLOOKUP(A2, Table2[#All], 2,TRUE)
```

For the purposes of consistency, the decision was made to only use the overall AQI values when analysing the data. This was due to the fact that the range of CO, Ozone, and NO2 AQI values for each category were unclear, even though the PM2.5 categories were clear.

Once the foreign key was inserted, the XLSX was saved as a CSV called “Source\_1” and imported into SQLite via the DB Browser.



## DATA SOURCE 2

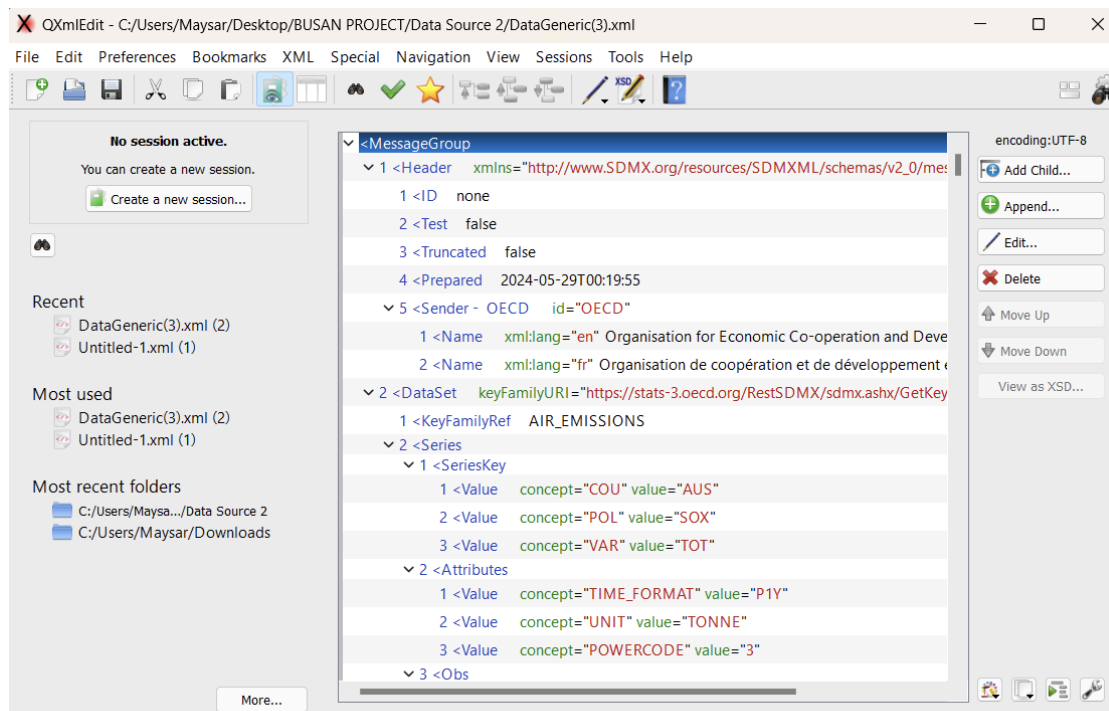
The second data source used in this report is the **Emissions of Air Pollutants** dataset. It is compiled by the OECD on [their website](#) for the purpose of providing selected information on national emissions of traditional air pollutants. The dataset was downloaded in XML format.

The original dataset was difficult to deal with, as there were some syntax issues and VSCode was having issues formatting the file. The data was also structured unconventionally. Rather than using nodes to identify column names (e.g., <COU>AUS</COU>), the dataset stored these in <Value> and <Obs> nodes, as shown below. This made it difficult to see where one entry started or ended, and it was assumed that the extraction process would be a lot harder than originally anticipated.

```

XML DOCUMENT
  </> MessageGroup (children: 2)
    </> Header (attributes: 1, children: 5)
    </> DataSet (attributes: 1, children: 5442)
      @ keyFamilyURI = "https://stats-3.oecd.o...
      </> KeyFamilyRef
      </> Series (children: 34)
        </> SeriesKey (children: 3)
          </> Value (attributes: 2)
            @ concept = "COU"
            @ value = "AUS"
          </> Value (attributes: 2)
          </> Value (attributes: 2)
        </> Attributes (children: 3)
          </> Value (attributes: 2)
            @ concept = "TIME_FORMAT"
            @ value = "P1Y"
          </> Value (attributes: 2)
          </> Value (attributes: 2)
        </> Obs (children: 2)
          </> Time
          </> ObsValue (attributes: 1)
            @ value = "1585.176"
```

A decision was then made to inspect the data in QXmlEdit, which automatically formats any XML data. This made it possible to understand the data structure and identify the best pathway to extracting the data.



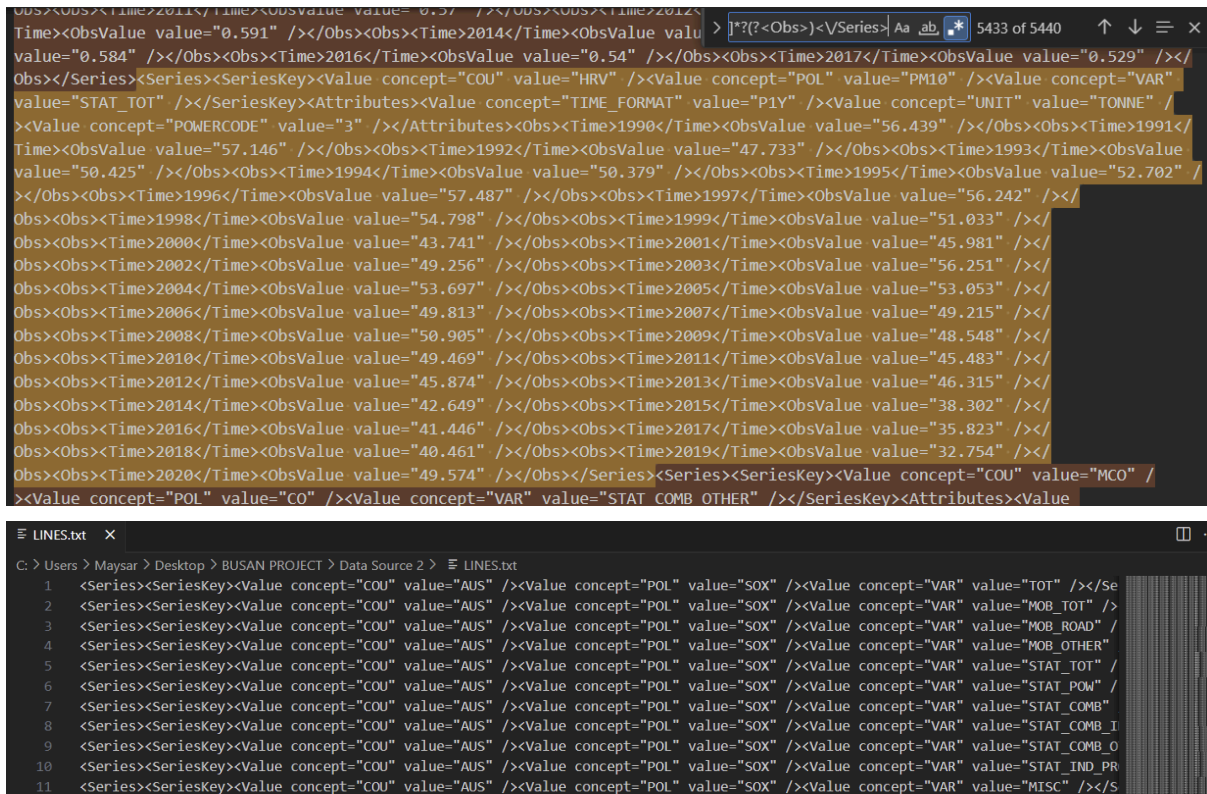
The data contains 5440 entries and the observed columns for this dataset are as follows:

Column Name	Description
COU	Country name, using the alpha-3 code
POL	The name of the pollutant
VAR	The variable of the pollutant
TIME_FORMAT	The format of the recorded time. All entries were P1Y, so this column was ignored in the extraction
UNIT	The units used to measure the pollutants are as follows: IDX, KG_1000USD, KG_HAB, and TONNE. Some rows were also missing data for this column
POWERCODE	The power code for each pollutant
REFERENCEPERIOD	The period of time in which the data was recorded. This column was dismissed in the extraction process because it was deemed redundant since there were already columns with specific years
1990	The amount of pollutants emitted in the year 1990
...	...
2021	The amount of pollutants emitted in the year 2021

A RegEx query was then used in VSCode to select and copy the relevant parts of the dataset into a new text file, as shown below. This query would extract each data entry (as divided by <Series></Series>) into new rows, making the data easier to work with.

```
<Series>[\s\S]*?(?<Obs>)</Series>
```

The process of extracting the data into a new text file is shown in the screenshots below.



The first screenshot shows a text editor window with XML data. The data is structured as a series of observations over time, with attributes like COU, POL, VAR, and POWERCODE. The second screenshot shows a file explorer window with the file 'LINES.txt' selected, displaying the same XML data in a text format.

The text file was saved as Lines.txt and imported into Excel for data extraction. These are examples of the functions used to extract the data into new columns. The decision was made not to include some attributes that were missing in most rows. In hindsight, the columns VAR and POWERCODE were not necessary for future analysis either.

```
//ATTRIBUTE VALUES

=IFERROR(MID(A2, FIND("COU", A2) + 12, FIND("COU", A2) - FIND("COU", A2) - 12), "")

=IFERROR(MID(A2, FIND("POL", A2) + 12, FIND("POL", A2) - FIND("POL", A2) - 12), "")

=IFERROR(MID(A2, FIND("VAR", A2) + 12, FIND("VAR", A2) - FIND("VAR", A2) - 12), "")

=IFERROR(MID(A2, FIND("UNIT", A2) + 13, FIND("UNIT", A2) - FIND("UNIT", A2) - 13), "")

=IFERROR(MID(A2, FIND("POWERCODE", A2) + 18, FIND("POWERCODE", A2) - FIND("POWERCODE", A2) - 18), "")

//YEAR VALUES: 1990 - 2021

=IFERROR(MID(A2, FIND("1990 =", A2) + 7, FIND("1990 =", A2) - FIND("1990 =", A2) - 7), "")

(etc.)
```

AutoSaveOff

Lines\_Pol\_clean

Search

FileHomeInsertDrawPage LayoutFormulasDataReviewViewAutomateDeveloperHelpPower PivotTable DesignQuery

Get Data

From Text/CSV

From Web

From Table/Range

From Picture

Recent Sources

Existing Connections

Get & Transform Data

Refresh All

Queries & Connections

Properties

Workbook Links

Queries & Connections

Stocks

Currencies

Data Types

Sort

Filter

Clear

Reapply

Advanced

Sort & Filter

Text to Columns

What-If Analysis

Data Tools

G2

:

</

The newly extracted data were then loaded into Power Query to replace NULL values in the year columns with 0, using the M query:

```
= Table.ReplaceValue("#"Removed Columns",null,0,Replacer.ReplaceValue,{"1990", "1991", "1992", "1993", "1994", "1995", "1996", "1997", "1998", "1999", "2000", "2001", "2002", "2003", "2004", "2005", "2006", "2007", "2008", "2009", "2010", "2011", "2012", "2013", "2014", "2015", "2016", "2017", "2018", "2019", "2020", "2021"})
```

Query Settings

PROPERTIES

Name

LINES\_POL (2)

All Properties

APPLIED STEPS

Source

Changed Type

Removed Columns

Replaced Value

= Table.ReplaceValue("#"Removed Columns",null,0,Replacer.ReplaceValue,{"1990", "1991", "1992", "1993", "1994", "1995", "1996", "1997", "1998", "1999", "2000", "2001", "2002", "2003", "2004", "2005", "2006", "2007", "2008", "2009", "2010", "2011", "2012", "2013", "2014", "2015", "2016", "2017", "2018", "2019", "2020", "2021"})

COU	POL	VAR	UNIT	POWERCODE	1990
AUS	SOX	TOT	TONNE	3	
AUS	SOX	MOB_TOT	TONNE	3	
AUS	SOX	MOB_ROAD	TONNE	3	
AUS	SOX	MOB_OTHER	TONNE	3	
AUS	SOX	STAT_TOT	TONNE	3	
AUS	SOX	STAT_POW	TONNE	3	
AUS	SOX	STAT_COMB	TONNE	3	
AUS	SOX	STAT_COMB_IND	TONNE	3	
AUS	SOX	STAT_COMB_OTHER	TONNE	3	
AUS	SOX	STAT_IND_PROC	TONNE	3	
AUS	SOX	MISC	TONNE	3	
AUS	SOX	MISC_AGR	TONNE	3	
AUS	SOX	MISC_WASTE	TONNE	3	
AUS	SOX	INDEX_1990	IDX	0	
AUS	SOX	TOT_CAP	KG_HAB	0	
AUS	NOX	TOT	TONNE	3	

The connection was then saved into a new CSV file called "Source\_2" and imported into SQLite via the DB Browser in the same way that data source 1 was imported.



# THE DATABASE

Once the data was imported correctly, the database was saved as "[matk663\\_DB.db](#)". The following SQL query was run to ensure that the foreign keys were working correctly:

```
SELECT *
FROM "Source_1" AS s1
JOIN "Source_2" AS s2
ON s1.CountryCode = s2.COU;
```

The join was successful and ready for further queries.

SQL 1

```
1 SELECT *
2 FROM "Source_1" AS s1
3 JOIN "Source_2" AS s2
4 ON s1.CountryCode = s2.COU;
```

	IAQIValue	PM2.5AQICategory	lat	lng	CountryCode	COU	POL	VAR	UNIT	POWERCODE	1990	1991	199
1	51	Moderate	44.7444	44.2031	ROU	ROU	CO	INDEX_1990	IDX	0	100.0	79.077	67
2	51	Moderate	44.7444	44.2031	ROU	ROU	CO	INDEX_2000	NULL	0	114.064	90.199	77
3	51	Moderate	44.7444	44.2031	ROU	ROU	CO	MISC	TONNE	3	156.168	120.812	101
4	51	Moderate	44.7444	44.2031	ROU	ROU	CO	MISC_AGR	TONNE	3	0.309	0.298	0
5	51	Moderate	44.7444	44.2031	ROU	ROU	CO	MISC_WASTE	TONNE	3	7.965	8.447	8
6	51	Moderate	44.7444	44.2031	ROU	ROU	CO	MOB_OTHER	TONNE	3	2.65	2.725	85
7	51	Moderate	44.7444	44.2031	ROU	ROU	CO	MOB_ROAD	TONNE	3	577.759	477.525	324
8	51	Moderate	44.7444	44.2031	ROU	ROU	CO	MOB_TOT	TONNE	3	580.41	480.25	405
9	51	Moderate	44.7444	44.2031	ROU	ROU	CO	STAT_COMB	TONNE	3	389.289	282.106	235
10	51	Moderate	44.7444	44.2031	ROU	ROU	CO	STAT_COMB_IND	TONNE	3	126.469	97.071	
11	51	Moderate	44.7444	44.2031	ROU	ROU	CO	STAT_COMB_OTHER	TONNE	3	262.871	185.034	165

Execution finished without errors.  
Result: 1114736 rows returned in 769ms  
At line 1:  
SELECT \*  
FROM "Source\_1" AS s1  
JOIN "Source\_2" AS s2  
ON s1.CountryCode = s2.COU;

A disclaimer about the database is that the scope of the data may not be fully reflective of real-world data. Data source 1 does not include some countries that are significantly represented in data source 2, such as India and Pakistan. As aforementioned, there were also missing data in source 1, so the data for cities without a valid country code were omitted. Furthermore, only data with units in tonnes were used for data source 1 was utilised in the analysis for the sake of consistency.



# Questions and Answers

A topic of concern upon analysing the data is the timeframe for the data; given that source 2 contains data only as recently as 2021, while source 1 was updated as recently as 2023. A decision was made to limit the scope of the question to only the year 2021 for the most optimal correlation between the two datasets.

The three questions to achieve the objective of this project are outlined below, along with the rationale behind the queries used to answer these questions.

## Question 1: Which were the top three most harmful pollutant types?

This question was asked to best narrow down the scope of future questions and focus on the pollutants with the most impact on air quality. To do this, the query was structured to divide the average value of the pollutant for the year 2021 by the average AQIValue from source 1. This is done using conditional aggregation with the AVG and CASE statements, which will result in percentages. The GROUP BY clause is then used to display the result set by the pollutant (POL), so the calculations are performed for each pollutant separately.

```
SELECT
  POL,
  (AVG(CASE WHEN POL = 'SOX' THEN "2021" ELSE NULL END) * 1.0 / AVG(s1.AQIValue)) AS
  SOXImpact,
  (AVG(CASE WHEN POL = 'NOX' THEN "2021" ELSE NULL END) * 1.0 / AVG(s1.AQIValue)) AS
  NOXImpact,
  (AVG(CASE WHEN POL = 'PM10' THEN "2021" ELSE NULL END) * 1.0 / AVG(s1.AQIValue))
  AS PM10Impact,
  (AVG(CASE WHEN POL = 'PM2-5' THEN "2021" ELSE NULL END) * 1.0 /
  AVG(s1.AQIValue)) AS PM2_5Impact,
  (AVG(CASE WHEN POL = 'CO' THEN "2021" ELSE NULL END) * 1.0 / AVG(s1.AQIValue)) AS
  COImpact,
  (AVG(CASE WHEN POL = 'NMVOC' THEN "2021" ELSE NULL END) * 1.0 / AVG(s1.AQIValue))
  AS VOCImpact
FROM
  "Source_1" AS s1
JOIN
  "Source_2" AS s2
ON
  s1.CountryCode = s2.COU
WHERE
  s2.UNIT = 'TONNE'
GROUP BY
  POL;
```

The results showed that CO contributed the most to the average AQI by 66%, followed by PM10 at 22% and NMVOC at 18% - identifying these as the most harmful pollutants.

	POL	SOXImpact	NOXImpact	PM10Impact	PM2_5Impact	COImpact	VOCImpact
1	CO	NULL	NULL	NULL	NULL	65.575537265817	NULL
2	NMVOC	NULL	NULL	NULL	NULL	NULL	18.0685671860074
3	NOX	NULL	12.5278620614293	NULL	NULL	NULL	NULL
4	PM10	NULL	NULL	22.3543536717001	NULL	NULL	NULL
5	PM2-5	NULL	NULL	NULL	6.65832883683894	NULL	NULL
6	SOX	3.293168687349	NULL	NULL	NULL	NULL	NULL

**Question 2: What countries have a hazardous AQI and how much of the three most harmful pollutants did they emit?**

This question was asked to get a snapshot of how much a country's emission in the recent year impacted its current hazardous AQI values (i.e. if AQIValue is 300 or more). This was done by filtering data using WHERE and then using SUM to calculate the total emissions for the three most harmful pollutants that were identified in the previous question.

```
SELECT
    s1.Country,
    SUM(CASE WHEN s2.POL = 'CO' AND s2.UNIT = 'TONNE' THEN s2."2021" ELSE 0
END) AS TotalCO,
    SUM(CASE WHEN s2.POL = 'PM10' AND s2.UNIT = 'TONNE' THEN s2."2021" ELSE 0
END) AS TotalPM10,
    SUM(CASE WHEN s2.POL = 'NMVOC' AND s2.UNIT = 'TONNE' THEN s2."2021" ELSE 0
END) AS TotalNMVOC,
    AVG(AQIValue)
FROM
    "Source_1" AS s1
JOIN
    "Source_2" AS s2
ON
    s1.CountryCode = s2.COU
WHERE
    s1.AQIValue >= 300
GROUP BY
    s1.Country
HAVING
    SUM(CASE WHEN s2.POL = 'CO' AND s2.UNIT = 'TONNE' THEN s2."2021" ELSE 0
END) > 0 OR
    SUM(CASE WHEN s2.POL = 'PM10' AND s2.UNIT = 'TONNE' THEN s2."2021" ELSE 0
END) > 0 OR
    SUM(CASE WHEN s2.POL = 'NMVOC' AND s2.UNIT = 'TONNE' THEN s2."2021" ELSE 0
END) > 0
ORDER BY AVG(s1.AQIValue) DESC;
```

Of the three countries in the database that fit this qualification, Chile and the USA's data were reasonable, as their high emission levels seemed to correlate with the hazardous AQI value. However, Russia's emission was surprisingly low considering the AQI value maxing out at 500. This could be due to lasting effects from Russia's high emissions in previous decades or that they were high emitters of other pollutants that were not this project's identified top three. Note that the average AQI value displayed below represents values for areas where the measured AQI was over 300, not the country's overall average.

	Country	TotalCO	TotalPM10	TotalNMVOC	AVG(AQIValue)
1	United States of America	369314.415	123300.795	99644.532	500.0
2	Russian Federation	3490.556	577.024	792.034	500.0
3	Chile	23534.942	1115.41	12200.316	358.0

**Question 3: What were the top 10 countries that emitted these three gasses and what was their maximum AQI value?**

As identified in the previous question, there were possibilities of outliers like Russia which did not have high emission levels for the top three pollutants despite a hazardous AQI. This question was asked to identify whether there were similar outliers of countries with high emission levels but a low AQI value. This was achieved by using SUM to calculate the total emissions of CO, PM10, and NMVOC, as well as the total emissions of these three pollutants combined. ORDER BY and LIMIT were then used to find the top ten emitters. MAX was used instead of AVG for the AQI values to highlight the worst-case scenario of air quality in each country.

```
WITH TotalEmissions AS (  
    SELECT  
        s2.COU AS CountryCode,  
        SUM(CASE WHEN s2.POL = 'CO' AND s2.UNIT = 'TONNE' THEN s2."2021" ELSE 0  
END) AS TotalCO,  
        SUM(CASE WHEN s2.POL = 'PM10' AND s2.UNIT = 'TONNE' THEN s2."2021" ELSE 0  
END) AS TotalPM10,  
        SUM(CASE WHEN s2.POL = 'NMVOC' AND s2.UNIT = 'TONNE' THEN s2."2021" ELSE 0  
END) AS TotalNMVOC,  
        SUM(CASE WHEN s2.POL IN ('CO', 'PM10', 'NMVOC') AND s2.UNIT = 'TONNE' THEN  
s2."2021" ELSE 0 END) AS TotalEmissions  
    FROM  
        "Source_2" AS s2  
    GROUP BY  
        s2.COU  
)  
SELECT  
    s1.Country AS CountryName,  
    te.TotalCO,  
    te.TotalPM10,  
    te.TotalNMVOC,  
    te.TotalEmissions,  
    MAX(s1.AQIValue) AS MaxAQI  
FROM  
    TotalEmissions AS te  
JOIN  
    "Source_1" AS s1  
ON  
    s1.CountryCode = te.CountryCode  
GROUP BY  
    s1.Country, te.TotalCO, te.TotalPM10, te.TotalNMVOC, te.TotalEmissions  
ORDER BY  
    te.TotalEmissions DESC  
LIMIT 10;
```

The results were staggering. Six out of the ten top emitters had maximum AQI values under 150, which is not considered unhealthy for the general population. Canada, the second most emitter, appeared to have a “moderate” rating for their worst polluted area, which is a stark contrast when compared to its neighbour the USA. This may be due to several factors. The first possibility is that the dispersion of those pollutants in the atmosphere might be more spread out compared to countries with higher population densities because Canada is a vast country with a relatively small population density in many areas. The second possibility is that Canada may have stricter environmental regulations and policies in place compared to other countries on the list, leading to better control of emissions and improved air quality

despite significant emissions. The third possibility is that the data is skewed due to the aforementioned missing data and the limited scope of the database.

	CountryName	TotalCO	TotalPM10	TotalNMVOC	TotalEmissions	MaxAQI
1	United States of America	123104.805	41100.265	33214.844	197419.914	500
2	Canada	15535.251	24827.041	4388.123	44750.415	83
3	Chile	11767.471	557.705	6100.158	18425.334	358
4	France	9684.583	943.158	3681.39	14309.131	151
5	Poland	9632.021	1448.512	2347.536	13428.069	113
6	Japan	10148.342	0	2523.167	12671.509	133
7	Kazakhstan	8417.776	1398.684	2728.132	12544.592	130
8	Turkey	6748.135	1998.283	3667.64	12414.058	121
9	Australia	8662.133	0	3669.714	12331.847	264
10	Germany	8324.13	576.891	3182.552	12083.573	94

*AQI Category: Green = Moderate, Yellow = Unhealthy for sensitive groups*

## REFERENCES

IBAN. (n.d.). *List of country codes by alpha-2, alpha-3 code (ISO3166)* [Data set]. Retrieved from <https://www.iban.com/country-codes>

OECD. (n.d.). *Emissions of air pollutants* [Data set]. Retrieved from [https://stats.oecd.org/Index.aspx?DataSetCode=AIR\\_EMISSIONS](https://stats.oecd.org/Index.aspx?DataSetCode=AIR_EMISSIONS)

Ramachandran, A. (n.d.). *World Air Quality Index by City and Coordinates* [Data set]. Kaggle. Retrieved from <https://www.kaggle.com/datasets/adityaramachandran27/world-air-quality-index-by-city-and-coordinates>