CIP Project Report

Group Number : 03

Student A: Aishwarya Patil

Student C: Simao Garcia

Date 22.04.2024

Table of Contents

1. Introduction

The global energy sustainability metrics and trends consist of an extensive analysis of how renewable energy production, greenhouse gas (GHG) emissions, oil production interconnect and impact on climate change and economic viability of the world. As the world faces an urgent need to shift away from fossil fuels to more sustainable energy sources, understanding these correlations becomes critical. Meanwhile, oil production remains a significant component of the global energy landscape, affecting both GHG emissions . By analysing trends in renewable energy adoption, changes in GHG emission levels, fluctuations in oil production, we can evaluate the progress being made towards a more sustainable and resilient energy future.

- **Motivation**

The 'Climate Index' project utilizes Python's web scraping capabilities to collect data on various countries' environmental performance metrics. It gathers information from diverse sources, processes it into a suitable format, and conducts analyses. This project compares factors such as renewable energy production, greenhouse gas emissions and oil production, to gauge each country's environmental sustainability.

- **Goals**

By examining these factors, the project aims to offer insights into which countries are making strides in combating climate change. The compiled data serves as a valuable resource for researchers and policymakers, enabling them to make informed decisions and comprehend the global environmental landscape. Python's web scraping tools streamline the process of data collection and analysis, making it accessible to individuals interested in understanding global climate efforts.

The primary objective of the project is to enhance our comprehension of environmental conditions by analyzing data in a straightforward and efficient manner. We concentrate on gathering information such as renewable energy generation, GHGs emissions and oil production. By scrutinizing these metrics, we can better assess a country's environmental resilience, aiding in the formulation of more effective climate strategies and policies.

- **Questions**

We have several questions we want to understand:

1. The correlation between renewable energy generation and GHGs emissions.

2. Determine countries with highest percentage of renewable energy production.

3. The correlation between oil production and GHGs emissions.

2. ETL process plan

Our project involves an ETL (Extract, Transform, Load) process for climate related web pages. In simpler terms, we gather information from these web pages, organize and reformat it, and then store it in a way that is easier to analyze and use.

During the extraction phase, we collect data from various energy, oil and emissions related by scraping and downloading from two different web pages. This can include details about countries, renewable electricity generation, oil production, greenhouse gases (GHGs) and other relevant information.

Next, we transform the data with pandas to make it consistent and structured. We organize the information, convert it into a standard format, and ensure it is ready for analysis.

Finally, we load the transformed data into a MariaDB database facilitating easy access and utilization. This empowers researchers, policymakers, and stakeholders to conveniently access the information and make informed decisions regarding climate-related issues. By

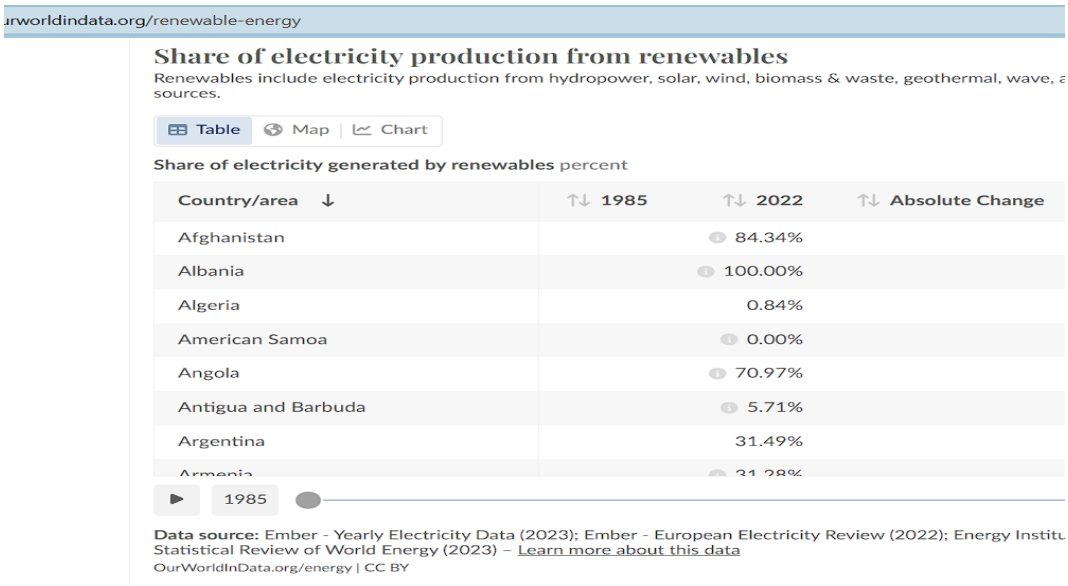
executing this ETL process on climate-related web pages, we streamline data collection, enhance accessibility, and enable comprehensive analysis for more effective decision-making in climate initiatives and policies.

2.1 Context diagram

3. Data Source .

3.1 Student C

Student A scraped the share of electricity production from renewables data from the website Renewable Energy - Our World in Data as shown in fig. 1



.

Fig.3.1 Dynamic table from the website.

Scrapped total 11 columns from 10 different tables. The names of the columns are shown below in Table 2.

Merged DataFrame:

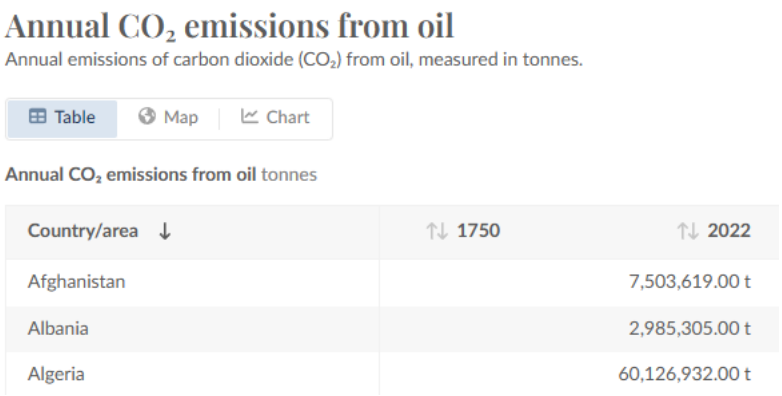
	Country/area	2022_Renewable(%)	2022_Hydro_generation(TWH)	2022_Hydro_generation(%)	2022_Hydro_consumption(%)	2022_Wind_gene
0	Algeria		0.40 TWh	5.26%	4.8%	
1	Argentina		1.23 TWh	45.59%	1.1%	
2	Australia		7.69 TWh	11.27%	5.3%	
3	Austria		16.08 TWh	70.36%	25.1%	
4	Azerbaijan			6.34%		
...	...	...	...	...	...	
95	South America	77.56%	39.58 TWh	76.47%	11.0%	
96	South and Central America (EI)	70.85%	41.30 TWh	69.37%	8.9%	
97	Upper-middle-income countries	25.53%	152.64 TWh	25.29%	4.4%	
98	Western Africa (EI)	45.30%	0.48 TWh	45.07%	2.5%	
99	World	20.82%	923.20 TWh	20.02%	6.3%	

100 rows × 11 columns

Table 3.1. Merged dataframe of all the tables

3.2 Student C

Student C downloaded the CO2 emission from oil data and the total GHG emission data from the website [CO₂ and Greenhouse Gas Emissions - Our World in Data] in fig. 3.2 and 3.3.



Annual CO_2 emissions from oil

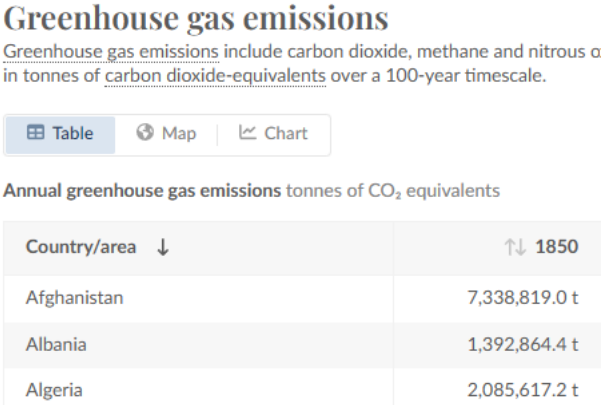
Annual emissions of carbon dioxide (CO_2) from oil, measured in tonnes.

Table | Map | Chart

Annual CO_2 emissions from oil tonnes

Country/area ↓	↑↓ 1750	↑↓ 2022
Afghanistan		7,503,619.00 t
Albania		2,985,305.00 t
Algeria		60,126,932.00 t

Fig. 3.2. Annual CO2 emission from oil (dynamic table from the website)



Greenhouse gas emissions

Greenhouse gas emissions include carbon dioxide, methane and nitrous o
in tonnes of carbon dioxide-equivalents over a 100-year timescale.

Table | Map | Chart

Annual greenhouse gas emissions tonnes of CO_2 equivalents

Country/area ↓	↑↓ 1850
Afghanistan	7,338,819.0 t
Albania	1,392,864.4 t
Algeria	2,085,617.2 t

Fig. 3.3. Greenhouse gas emissions (dynamic table from the website)

In addition, the oil production by country data was downloaded from the website [Oil Production by Country 2024 (worldpopulationreview.com).] The data is illustrated in the fig 3.4.

Country	Oil Production 2022 (Million Tons)	Daily Production 2022 (1K Barrels Per Day)	Production 2021 (Million Tons)	Daily Production 2021 (1K Barrels Per Day)	Production 2020 (Million Tons)	Daily Production 2020 (1K Barrels Per Day)
Search... (47)	Min Max	Min Max	Min Max	Min Max	Min Max	Min Max
United States	759.50	17770.00	715.20	16,679	713.30	16,492
Saudi Arabia	573.10	12136.00	515.00	10,954	519.60	11,039
Russia	548.50	11202.00	538.80	11,000	524.40	10,666
Canada	274.00	5576.00	266.60	5,414	252	5,130
Iraq	221.30	4520.00	200.80	4,102	202	4,114
China	204.70	4111.00	198.90	3,994	194.80	3,901
United Arab Emirates	181.10	4020.00	163.40	3,640	165.90	3,679
Iran	176.50	3822.00	168.80	3,653	144.40	3,120
Brazil	163.10	3107.00	156.90	2,990	159.30	3,030
Kuwait	145.70	3028.00	129.90	2,704	131.20	2,721
Mexico	97.70	1944.00	96.50	1,928	95.10	1,912
Norway	89.00	1901.00	93.90	2,028	92.10	2,006
Kazakhstan	84.10	1769.00	85.90	1,805	85.70	1,796
Qatar	74.10	1768.00	72.80	1,736	71.70	1,703

Fig 3.4. Oil production by country

4.Tools used

4.1.Selenium

Selenium was utilized in this project to automate web browser interactions, facilitating the retrieval of data from web pages and enabling efficient web scraping tasks. By simulating user interactions, such as button clicks and form submissions, Selenium helped streamline data extraction processes from dynamic websites.

4.2. MariaDB

For this project, MariaDB serves as the database management system, providing a robust and scalable solution for storing and managing structured data. It enables efficient data storage, retrieval, and manipulation, supporting various SQL operations required for the project's database interactions. MariaDB's compatibility with MySQL ensures seamless integration with existing systems and libraries, facilitating smooth development and deployment processes. Additionally, MariaDB's features such as high availability, replication, and security enhancements contribute to the project's reliability and data integrity.

5. Solution steps

After loading each individual dataset into MariaDB, both datasets went underwent a transformation processes, to get ready for further analysis.

The First step was to upload both that required to be merged. The Figures 5.1 & 5.2 show the information about the data column, data type and the number of non – null from both datasets. It is clear that the Student A dataset has columns with nulls value. For this reason, data cleaning has to be performed on this data.

#	Column	Non-Null Count	Dtype
0	Country_area	235 non-null	object
1	Code	235 non-null	object
2	Year	235 non-null	int64
3	Greenhouse_gas_emissions_in_CO_2_equivalents(t)	235 non-null	float64
4	CO_2_emissions_from_oil(t)	235 non-null	float64
5	Oil_Production_(Kbbl/Day)	235 non-null	float64

Fig 5.1 shows the student_C dataset attributes . The full data has 235 rows and 6 columns.

#	Column	Non-Null Count	Dtype
0	Country/area	100 non-null	object
1	2022_Renewable(%)	57 non-null	float64
2	2022_Hydro_generation(TWH)	67 non-null	float64
3	2022_Hydro_generation(%)	95 non-null	float64
4	2022_Hydro_consumption(%)	58 non-null	float64
5	2022_Wind_generation(TWH)	30 non-null	float64
6	2022_Wind_generation(%)	66 non-null	float64
7	2022_Wind_consumption(%)	22 non-null	float64
8	2022_Solar_generation(TWH)	28 non-null	float64
9	2022_Solar_generation(%)	59 non-null	float64
10	2022_Solar_consumption(TWH)	18 non-null	float64
11	OtherRenewables_2022(%)	57 non-null	float64
12	Total_EnergyProd_2022(TWH)	41 non-null	float64
13	Norm_RenEnergy_produced_2022	57 non-null	float64

Fig. 5.2 shows the student_A dataset attributes. The full data has 100 rows and 14 columns.

The second step is to rename the Student_A column names, so that the dataset can have a key variable to link to the student B data. After renaming the columns names, the datasets are merged.

The Figure 5.3 shows that the merged dataset has null values in most of the columns. This means that some cleaning has to be performed to remove the null values.

```
Data columns (total 19 columns):
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
 0   Country_area                                    235 non-null    object
 1   Code                                            235 non-null    object
 2   Year                                            235 non-null    int64
 3   Greenhouse_gas_emissions_in_CO₂_equivalents(t)  235 non-null    float64
 4   CO₂_emissions_from_oil(t)                       235 non-null    float64
 5   Oil_Production_(Kbbl/Day)                       235 non-null    float64
 6   Renewable(%)                                    44 non-null     float64
 7   Hydro_generation(TWH)                           56 non-null     float64
 8   Hydro_generation(%)                             77 non-null     float64
 9   Hydro_consumption(%)                            43 non-null     float64
 10  Wind_generation(TWH)                            24 non-null     float64
 11  Wind_generation(%)                              53 non-null     float64
 12  Wind_consumption(%)                             16 non-null     float64
 13  Solar_generation(TWH)                           23 non-null     float64
 14  Solar_generation(%)                             46 non-null     float64
 15  Solar_consumption(TWH)                          13 non-null     float64
 16  Other_Renewables(%)                             44 non-null     float64
 17  Total_EnergyProd(TWH)                           33 non-null     float64
 18  Norm_RenEnergy_produced                         44 non-null     float64
```

Fig. 5.3 shows the merged datasets. It also shows the variable atributes, its corresponding data type and the number of non-null values in each column. The full data has 235 rows and 19 columns.

The third step consisted of cleaning the nulls shown on Figure. The cleaning method used here is the .fillna(), where all null values are filled by 0 (zero) . The Figure xx shows the outcome of the cleaning process.

```
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
 0   Country_area                                    235 non-null    object
 1   Code                                            235 non-null    object
 2   Year                                            235 non-null    int64
 3   Greenhouse_gas_emissions_in_CO₂_equivalents(t)  235 non-null    float64
 4   CO₂_emissions_from_oil(t)                       235 non-null    float64
 5   Oil_Production_(Kbbl/Day)                       235 non-null    float64
 6   Renewable(%)                                    235 non-null    float64
 7   Hydro_generation(TWH)                           235 non-null    float64
 8   Hydro_generation(%)                             235 non-null    float64
 9   Hydro_consumption(%)                            235 non-null    float64
 10  Wind_generation(TWH)                            235 non-null    float64
 11  Wind_generation(%)                              235 non-null    float64
 12  Wind_consumption(%)                             235 non-null    float64
 13  Solar_generation(TWH)                           235 non-null    float64
 14  Solar_generation(%)                             235 non-null    float64
 15  Solar_consumption(TWH)                          235 non-null    float64
 16  Other_Renewables(%)                             235 non-null    float64
 17  Total_EnergyProd(TWH)                           235 non-null    float64
 18  Norm_RenEnergy_produced                         235 non-null    float64
```

Fig. 5.4. Shows the attributes of the final merged dataset. It has 235 rows and 19 columns

Finaly, the final merged data is uploaded into MariaDB thought python.

6. Data analysis

The study carried on this project consists of answering three questions:

- Are countries with higher Oil production generally associated with higher GHGs emissions?
- Are countries with higher renewable energy production generally associated with lower GHGs emissions?
- Which countries have the highest percentage of renewable energy production?

1. **Relationship between CO_2 emissions from oil and greenhouse gas emissions:**

As a starting point, it will be interesting to compare the CO_2 emissions that comes out of the oil production and the greenhouse gas emissions. It is expected that countries with higher oil production tend have higher greenhouse gas emissions. Let's see if this is indeed the case.

A scatter plot is made for visualizing and comparing the CO_2 emissions that comes out of the oil with the greenhouse gas emissions for each country. The Figure 1 illustrates that there is a linear correlation between both parameters. However, it is important to determine the correlation factor between both parameters.

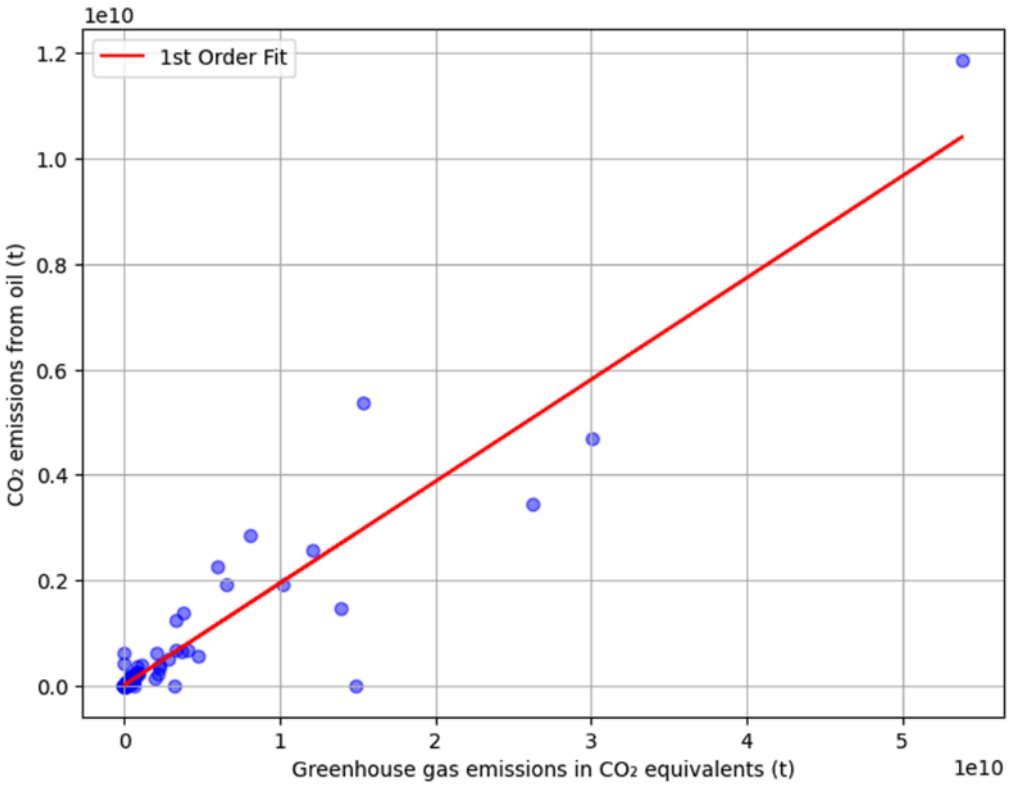


Fig. 6.1. Linear regression between GHG and CO_2 emission from oil

The table 6.1 is made for calculating the correlation factor between GHG and CO_2 emission from oil. It shows that there is a correlation of 0.96. This suggests that countries with higher oil production tend to have higher greenhouse gas emissions. This relationship is likely due to the fact that increased oil production often involves both direct emissions from extraction and refining processes and indirect emissions from consumption of the produced oil.

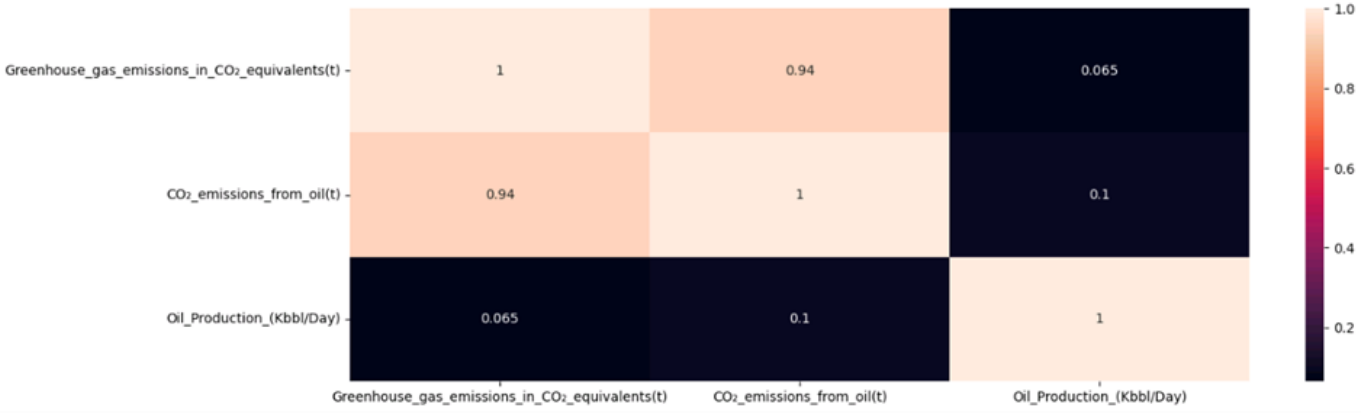


Table 6.1. Correlation between GHG, CO_2 emission from oil and oil production

2. **<u>Relationship between greenhouse gas emissions and renewable energy production:</u>**

As the next step, an attempt is made to see if the countries having a large share of their energy produced from renewable energy also have lower greenhouse gas emissions. Again, a scatter plot was made to answer this question.

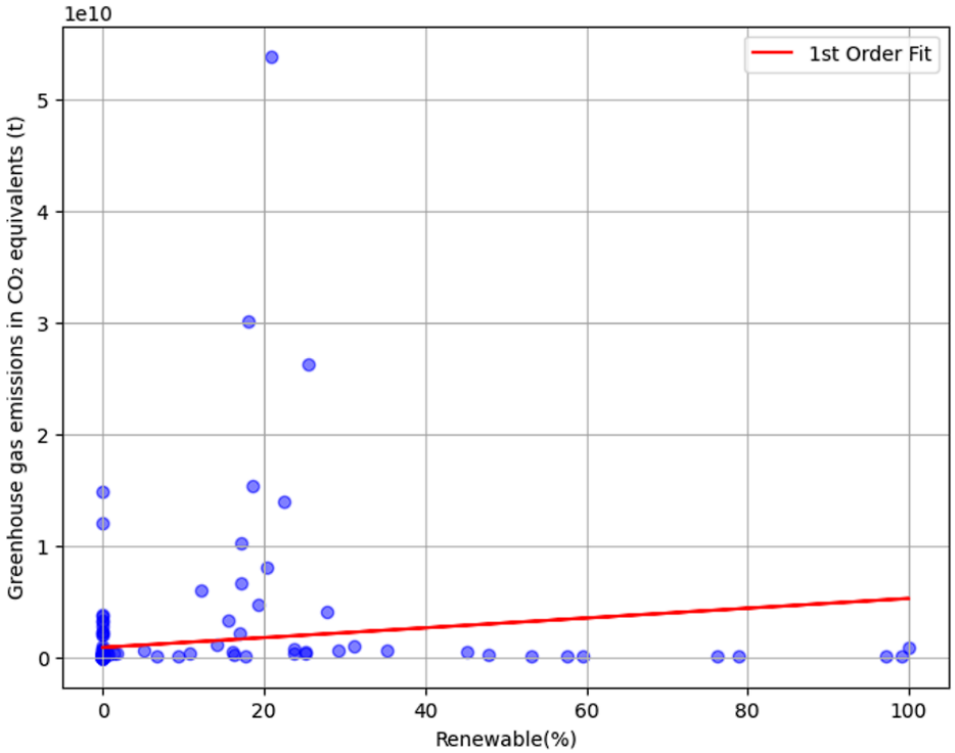


Fig. 6.2. Linear regression between GHG and renewable energy

It is observed from Figure 6.2 that there is no direct relation between the aforementioned parameters. It maybe that even though the percentage share of renewable energy production is high, the absolute value is much lower. Apart from that, other factors independent of renewable energy such as automobile pollution, etc can be the contributor to the high amounts of greenhouse gas emissions in several countries.

3. **Top 5 countries having maximun % of electricity generated from Renewable resources is shown in the fig 6.3**

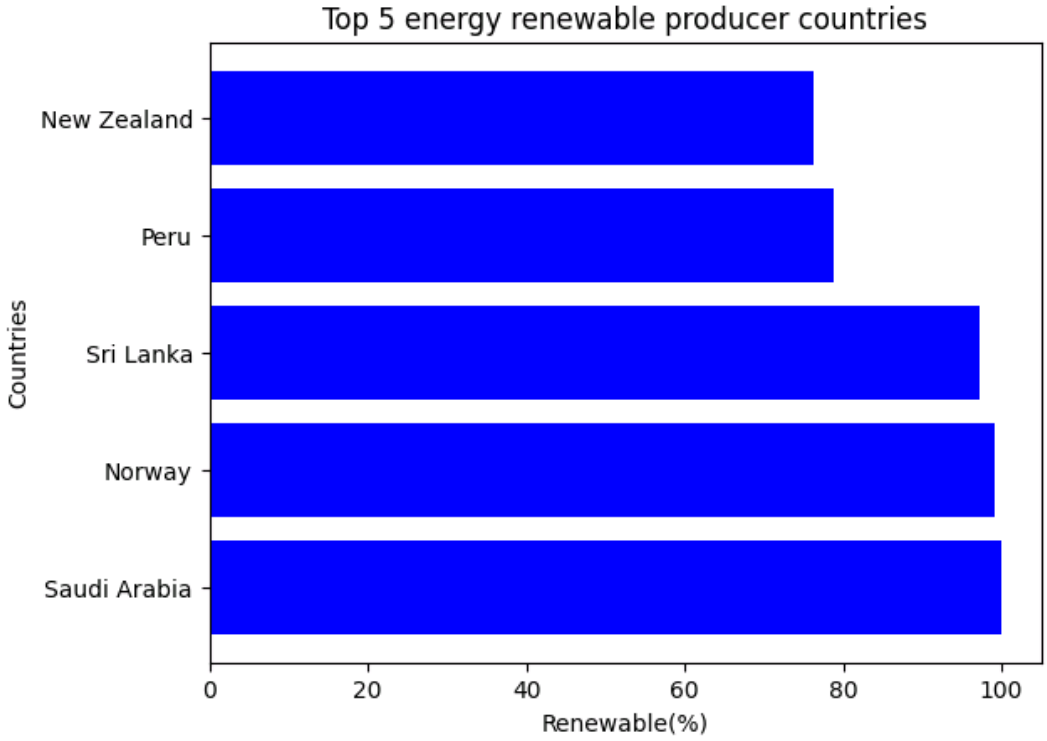


Fig. 6.3. Top 5 renewable energy producers countries in percentage

7. Lesson learned / self-reflection

1. Choosing the right project topic was important for its success. It should be flexible and easy to expand as needed. We struggled to find an appropriate topic which is both interesting and we can find irrelevant data from web pages.

2. Teamwork and communication are crucial for a project's quality. We have a strong collaboration between us. Even though we experienced a lack of Student B, we managed to deliver our project documentation with answers.

3. A context diagram was to be simple, easy to understand, and cover all the necessary information. It should have given the whole process of the project that someone can

understand the project just by looking at this diagram. We had some difficulties before we started the project. We were able to draw our diagram while we were doing a project.

5. We struggled in the analysis part because one of the team members (Role Student B) who stopped studies. We tried to answer questions, but it was not as good as we expected before starting the project. However, we have answers for all the questions. This resulted in delayed submission of the group work.

6. GPT played a crucial role in providing guidance, explanations, and assistance throughout the project by generating explanations, and answering questions related to data processing, database management, and coding practices. This enhanced our overall development process and learning experience.