Analysis of the Impact of Renewable Energy Consumption on CO_2 Emissions Across U.S. Sectors (1973–2024) for Sustainable Policy Insights

Abstract

This report details the development of an automated data pipeline for analyzing the impact of renewable energy adoption on U.S. carbon emissions. The project involved acquiring and processing data from two Kaggle datasets: renewable energy consumption and CO_2 emissions. The pipeline performed data merging, filtering, and validation to ensure consistency, and stored the final data in an SQLite database for further analysis. Key challenges included handling inconsistent columns and ensuring data integrity, which were resolved through targeted coding solutions. Future improvements could involve real-time data integration, adaptive normalization, and geospatial analytics for enhanced analysis.

---

1. **Question** How does renewable energy consumption influence carbon dioxide emissions across different sectors in the United States, and what insights can be drawn to guide sustainable policy formulation

---

2. Data Sources

2.1 Description of Data Sources

- **Dataset 1: Renewable Energy Consumption in the U.S.**
    - **Source**: Kaggle - Alistair King
    - **Data URL**: Renewable Energy Consumption in the U.S.
    - **Data Type**: CSV
    - **Description**: This dataset provides insights into renewable energy consumption in the U.S., categorized by year, state, sector, and fuel type. It is essential for understanding renewable energy adoption trends.
- **Dataset 2: CO2 Emissions in the U.S.**
    - **Source**: Kaggle - Abdelrahman16
    - **Data URL**: CO2 Emissions in the U.S.
    - **Data Type**: CSV
    - **Description**: This dataset includes detailed records of CO2 emissions by year, allowing analysis of emission trends and comparison with renewable energy data.

2.2 Data Structure and Quality

- **Dataset 1 (Renewable Energy Consumption)**: This dataset is structured in a tabular format with columns representing year, state-name, sector-name, fuel-name, and value (the energy consumption value). The data spans from 1973 to 2024, offering comprehensive insights into renewable energy trends. Missing values were addressed during data processing.

Additional Data (CO2 Emissions) preview:

	Year	Month	Sector	Hydroelectric Power	Geothermal Energy	...	Total Renewable Energy	Renewable Diesel Fuel	Other Biofuels	Conventional Hydroelectric Power	Biodiesel
0	1973	1	Commerical	0.00	0.00	...	0.570	0.0	0.0	0.000	0.0
1	1973	1	Electric Power	0.00	0.49	...	89.223	0.0	0.0	88.522	0.0
2	1973	1	Industrial	1.04	0.00	...	99.973	0.0	0.0	0.000	0.0
3	1973	1	Residential	0.00	0.00	...	30.074	0.0	0.0	0.000	0.0
4	1973	1	Transportation	0.00	0.00	...	0.000	0.0	0.0	0.000	0.0

- **Dataset 2 (CO2 Emissions)**: The dataset is structured similarly, with columns like Year, Country, and various CO2 emission types (cement_co2, coal_co2, gas_co2, oil_co2). It provides a global

Data (Renewable Energy Consumption) preview:

	year	state-name	sector-name	fuel-name	value
0	1970	Alabama	Industrial carbon dioxide emissions	Coal	26.721507
1	1970	Alabama	Industrial carbon dioxide emissions	Petroleum	3.577779
2	1970	Alabama	Industrial carbon dioxide emissions	Natural Gas	8.944097
3	1970	Alabama	Industrial carbon dioxide emissions	All Fuels	39.243383
4	1970	Alabama	Total carbon dioxide emissions from all sectors	All Fuels	102.646851

overview of CO2 emissions from 1970 to 2021, with a specific focus on U.S. emissions post-1950.

2.3 Licenses

- **Renewable Energy Dataset**: The dataset is under Kaggle's open data license, allowing free access for educational purposes.
- **CO2 Emissions Dataset**: This dataset is also open access, following Kaggle's data usage policies for educational and research purposes.

Both datasets are used in compliance with their licenses, focusing solely on non-commercial, educational analysis.
I will use these datasets just for educational projects.

3. Data Pipeline

The data pipeline was implemented in Python using libraries such as pandas, os, and sqlite3. The pipeline automates the download, preprocessing, merging, and storage of the datasets into a SQLite database.

3.1 Overview of the Pipeline Process

1. **Downloading Datasets**: Data is fetched from Kaggle using the os library to execute Kaggle CLI commands.
2. **Loading and Previewing Data**: The CSV datasets are read into pandas DataFrames.
3. **Data Preprocessing**: Columns are selected and renamed to ensure consistency between the datasets. The Year column is processed to ensure data alignment.
4. **Merging Datasets**: DataFrames are merged on the year column to create a unified dataset.
5. **Data Filtering and Saving**: Unnecessary columns are dropped, and relevant data is stored in an SQLite database for further analysis.

3.2 Challenges and Solutions

- **Challenge 1: Missing Column Values**

- **Solution**: Used a conditional check to ensure the presence of the Year column before processing. Implemented exception handling to catch any missing data issues.

- **Challenge 2: Data Alignment Issues**
  - **Solution**: Renamed columns and performed group-by operations to align data formats and ensure consistency.

- **Challenge 3: Ensuring Data Integrity**
  - **Solution**: Used data preview checks and validation steps to confirm that only the necessary columns were retained.

---

4. Results and Limitations

4.1 Output Data Structure

The processed data is stored in an SQLite database named renewable_energy.sqlite3. This database includes a table named renewable_energy, containing columns for year, state-name, sector-name, fuel-name, and value.

4.2 Data Quality

- **Accuracy**: The datasets were sourced from reputable platforms (Kaggle), ensuring high-quality data.
- **Completeness**: The data was filtered to include records from 1973 to 2021, minimizing gaps.
- **Consistency**: The data underwent normalization steps such as renaming columns and grouping, resulting in consistent formatting.
- **Timeliness**: The data includes the most recent records, up to 2021.

4.3 Data Format

The final output format chosen was SQLite due to its portability, ease of use, and compatibility with data analysis tools.

4.4 Potential Issues and Limitations

- **Data Gaps**: While the data was filtered for completeness, some periods may still have missing values, impacting analysis.
- **Data Correctness**: Even with reputable sources, there is always the possibility of inaccuracies due to errors during data collection.
- **Scalability**: As the analysis grows, performance may be impacted by the database's size and complexity.

```
Data from SQLite:
   year state-name                                  sector-name    fuel-name       value
0  1973    Alabama            Industrial carbon dioxide emissions         Coal   23.552431
1  1973    Alabama            Industrial carbon dioxide emissions    Petroleum    5.541595
2  1973    Alabama            Industrial carbon dioxide emissions  Natural Gas    8.300523
3  1973    Alabama            Industrial carbon dioxide emissions    All Fuels   37.394549
4  1973    Alabama  Total carbon dioxide emissions from all sectors    All Fuels  109.563135
```