ANALYZING THE IMPACT OF DEVELOPMENT INDICATORS ON CO_2 EMISSIONS: A DATA SCIENCE AND LOW-CODE APPROACH

Bushra Jabeen
Capstone Project 1
Data Talent Program – Cohort 17

INDUSTRY

Environmental and Sustainable Development industry, focusing on the **impact of development indicators on CO_2 emissions**.

OBJECTIVE:

To analyze how various development indicators influence CO_2 emissions across countries and identify patterns or clusters of countries based on their development and environmental impact, using both traditional data science methods and low-code tools.

KEY QUESTIONS

- Which development indicators have the strongest relationship with CO_2 emissions?
- How do countries cluster based on their development metrics and emissions levels?
- Can we predict CO_2 emissions using selected development indicators?
- Can low-code tools like KNIME be effectively used for building predictive models using development and environmental data?

DATA COLLECTION:

World Bank Data: You pulled development indicators including CO_2 emissions per capita, CO_2 intensity (kg per PPP $ of GDP), total CO_2 emissions, etc., from the **World Bank's World Development Indicators (WDI)** via the World Bank Open Data / DataBank portal https://databank.worldbank.org/source/world-development-indicators

Kaggle Dataset: You downloaded the **CO_2 Emissions by Sectors** dataset (covering multiple countries and years) from Kaggle, specifically the "Co2_Emissions_by_Sectors" https://www.kaggle.com/datasets/avinashsingh004/co2-emissions-by-sectors?utm_source=chatgpt.com

DATA PREPARATION:

- Two separate datasets:
  - World Bank development indicators
  - CO_2 emissions dataset from OWID
- Filtered data for the year **2022** only
- Aligned datasets by **country name**
- Selected relevant indicators and CO_2 columns
- Merged both datasets into a single dataframe for analysis

DATA CLEANING

•**Dropped Duplicate and Null Columns**:
Removed columns with many null values and duplicates:

- •**Current health expenditure (% of GDP)**
- •**Literacy rate, adult total (% of people ages 15 and above)**
- •**Country Name_y**

•**Filled Missing Values with Median**:
Some columns had missing values, and I filled them using the **median** to prevent bias and maintain consistency.
Columns filled with median:

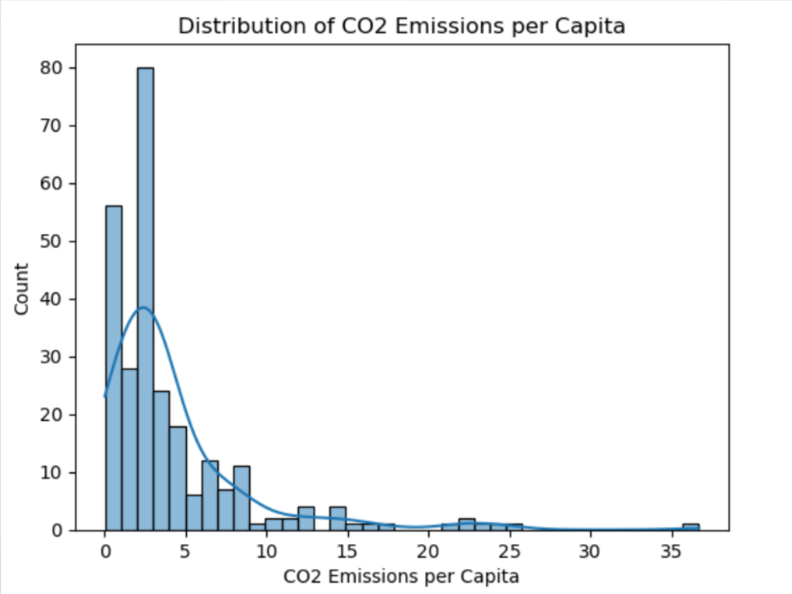
- •**Access to electricity (% of population)**
- •**GDP per capita (current US$)**
- •**Individuals using the Internet (% of population)**
- •**School enrollment, secondary (% gross)**
- •**Unemployment, total (% of total labor force) (modeled ILO estimate)**
- •**CO_2 Emissions (million tonnes)**
- •**CO_2 Emissions per Capita**
- •**Population**
- •**GDP (CO_2 dataset)**

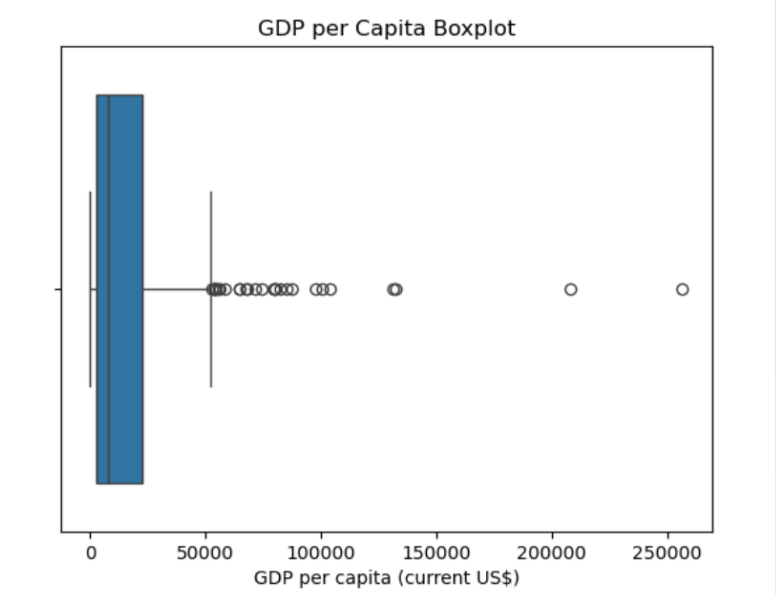
```
[8]: df_final.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 265 entries, 0 to 264
Data columns (total 16 columns):
 #   Column                                                          Non-Null Count  Dtype
---  ------                                                          --------------  -----
 0   Country Name                                                    265 non-null    object
 1   Country Code                                                    265 non-null    object
 2   Access to electricity (% of population)                         263 non-null    float64
 3   Current health expenditure (% of GDP)                           21 non-null     float64
 4   GDP per capita (current US$)                                    249 non-null    float64
 5   Individuals using the Internet (% of population)                186 non-null    float64
 6   Life expectancy at birth, total (years)                         265 non-null    float64
 7   Literacy rate, adult total (% of people ages 15 and above)      44 non-null     float64
 8   Population growth (annual %)                                    265 non-null    float64
 9   School enrollment, secondary (% gross)                          120 non-null    float64
 10  Unemployment, total (% of total labor force) (modeled ILO estimate)  232 non-null    float64
 11  Country Name_y                                                  206 non-null    object
 12  CO2 Emissions (million tonnes)                                  204 non-null    float64
 13  CO2 Emissions per Capita                                        204 non-null    float64
 14  population                                                      206 non-null    float64
 15  GDP (CO2 dataset)                                               163 non-null    float64
dtypes: float64(13), object(3)
memory usage: 33.3+ KB
```

EXPLORATORY DATA ANALYSIS:

Univariate Analysis:

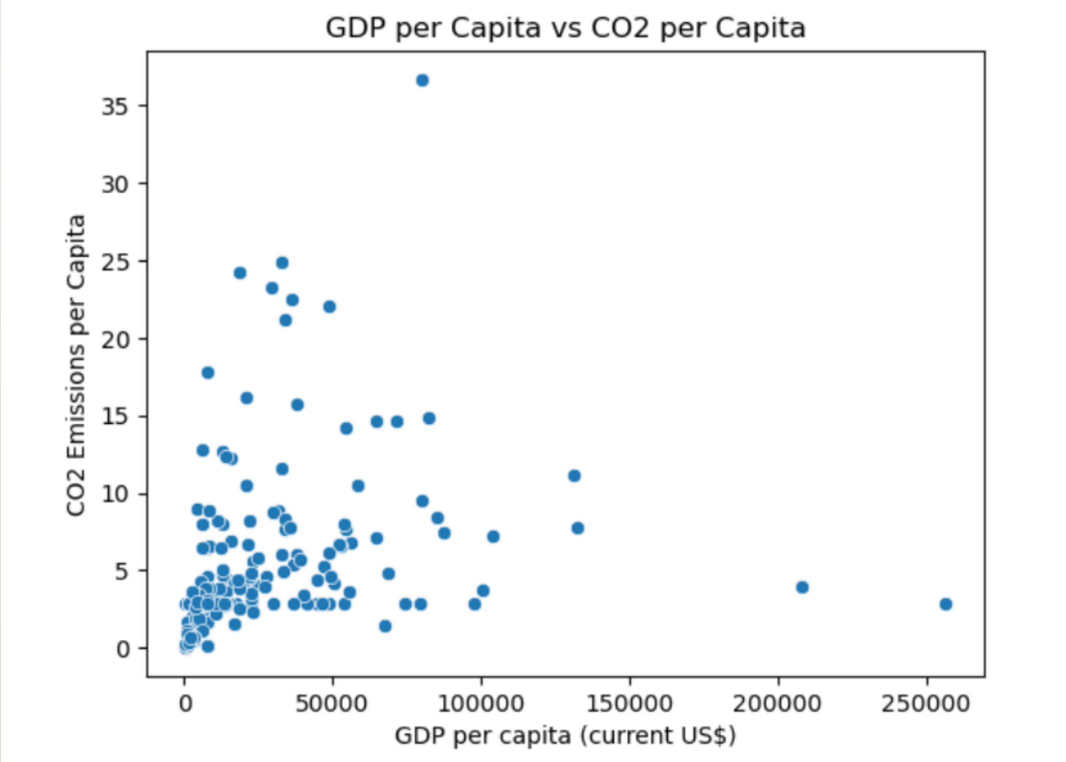
The distributions reveal strong right-skewness in GDP per capita and CO_2 emissions per capita, indicating a few countries dominate the upper end. Most countries have high electricity access and life expectancy, suggesting global progress in basic infrastructure.





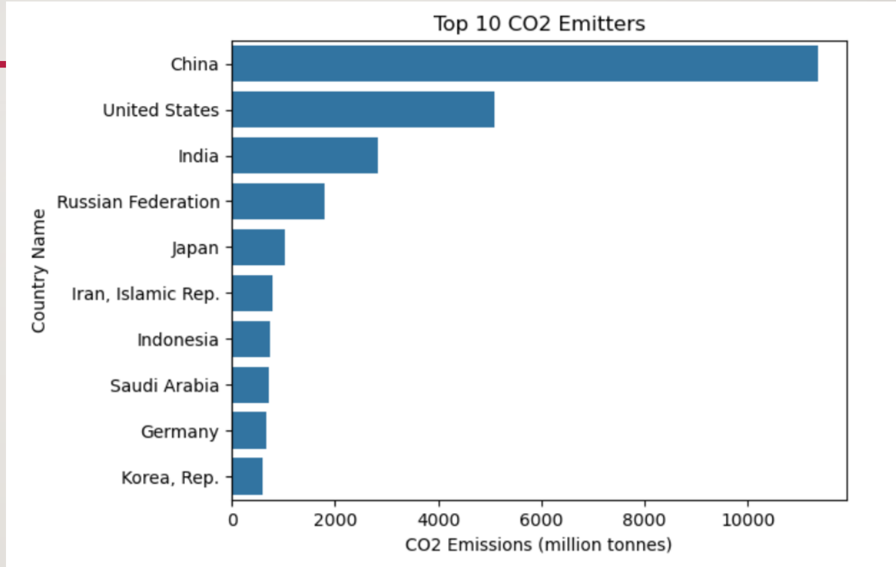
BIVARIATE ANALYSIS:

GDP per capita shows a positive correlation with CO_2 emissions per capita, suggesting wealthier countries tend to pollute more. Similarly, countries with high electricity access also report higher internet usage, reflecting digital inclusion.

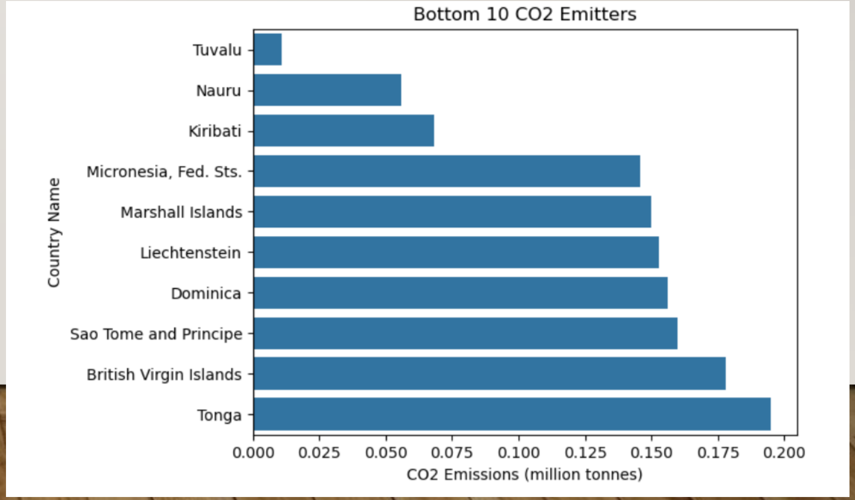


COUNTRY-LEVEL COMPARISON:

- The top CO_2 emitters in absolute terms are large economies like China, USA, and India. However, on a per capita basis, smaller nations like Qatar and Kuwait rank high due to high fossil fuel dependence despite smaller populations.

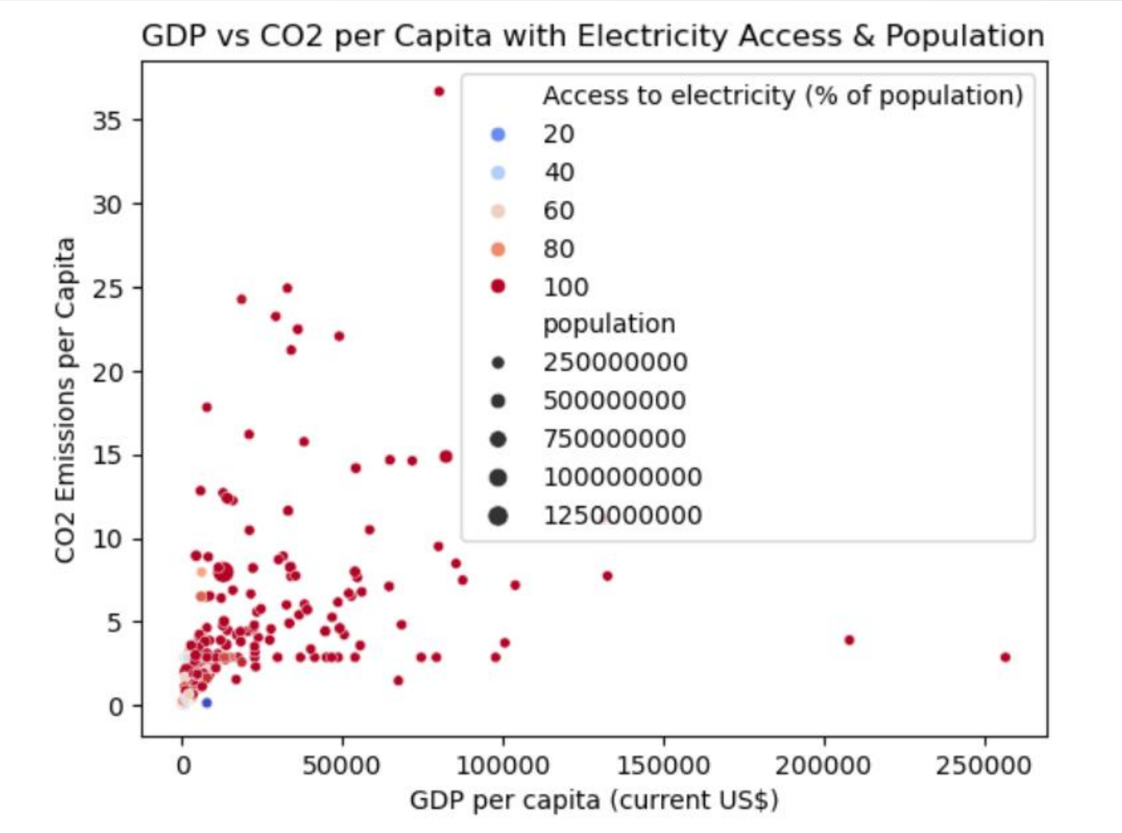


- The bottom 10 countries by GDP per capita and emissions are primarily low-income nations, often from Sub-Saharan Africa. These countries also tend to lag in internet usage, electricity access, and education, indicating development gaps.



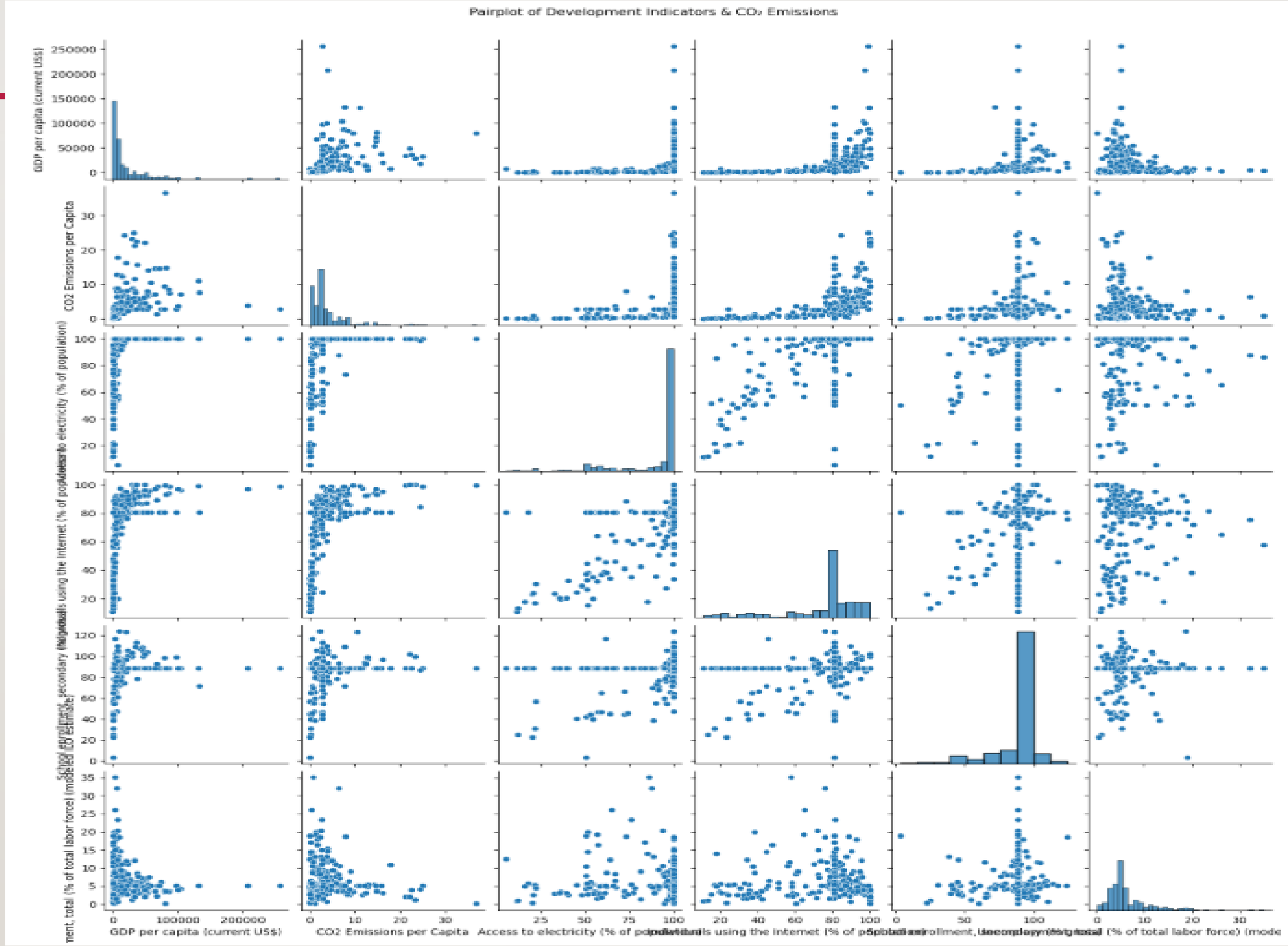
MULTIVARIATE ANALYSIS:

Multivariate plots highlight clear clusters of countries: those with high GDP, high CO_2, and strong infrastructure, versus countries with lower development and minimal emissions. Population size further differentiates these groups.



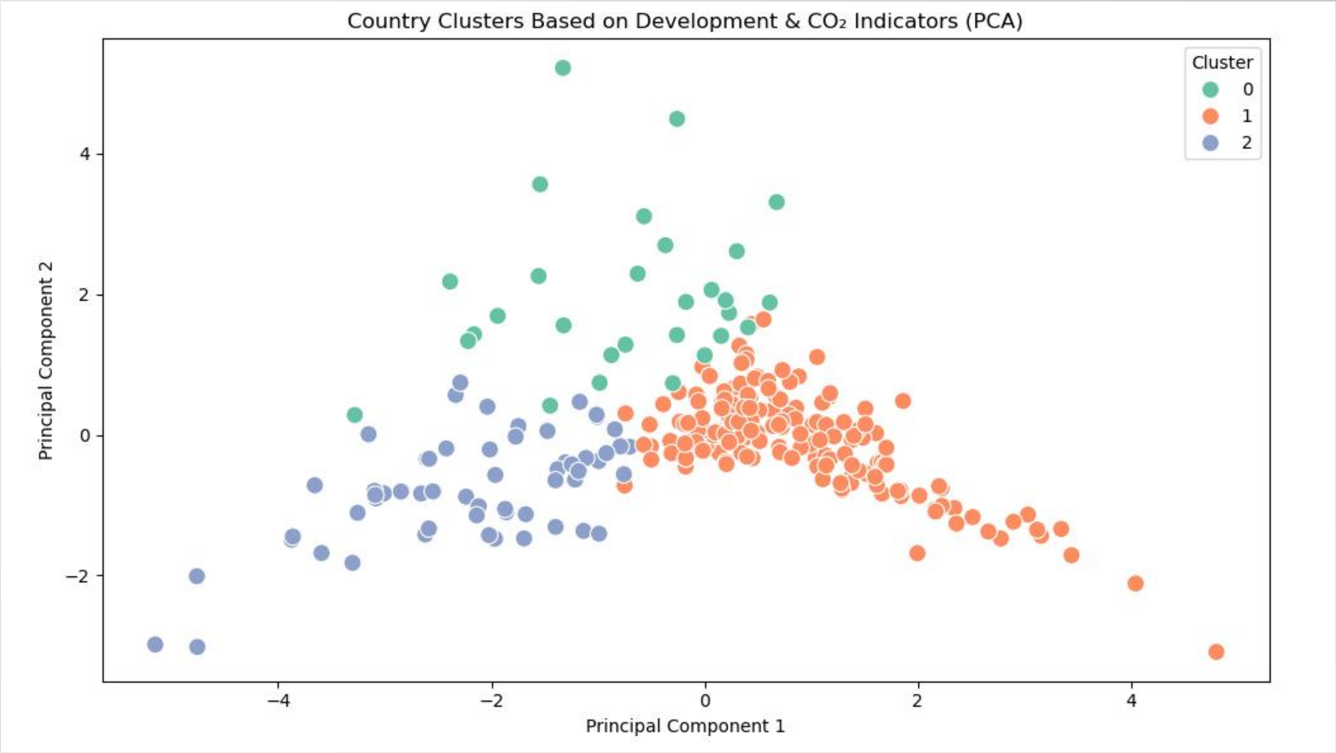
ADVANCED EXPLORATORY ANALYSIS:

The pairplot reveals strong linear relationships between GDP per capita, electricity access, and internet usage — indicators often associated with higher development. There's also a noticeable cluster of high-GDP, high-emission countries, indicating co-movement of economic and environmental metrics.



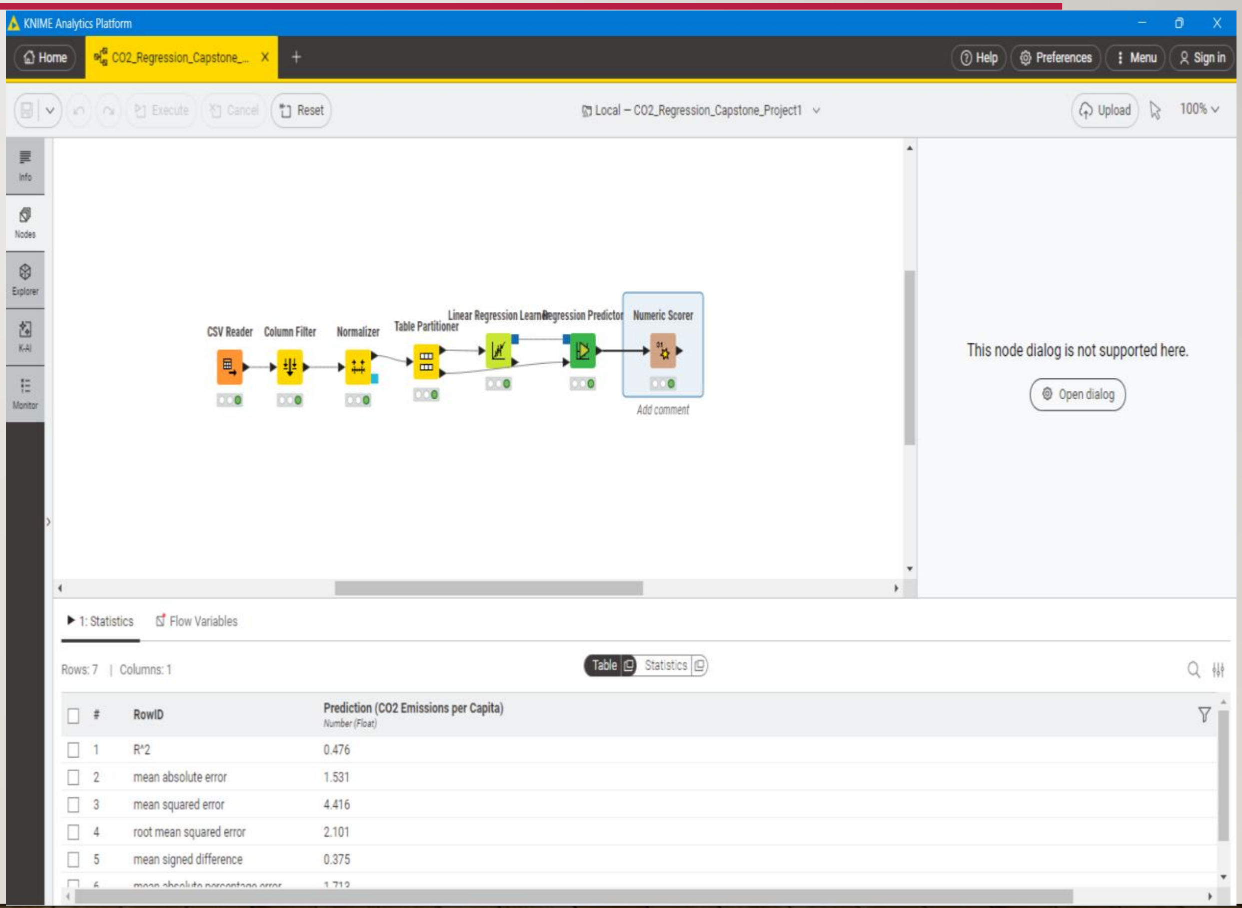
PRINCIPAL COMPONENT ANALYSIS FOLLOWED BY K-MEANS CLUSTERING

PCA reduced multiple indicators to two principal components, capturing most variance. K-Means clustering revealed three distinct groups of countries based on development and emissions profiles, enabling targeted regional comparisons.



KNIME (LOW-CODE/NO-CODE APPROACH):

- Used **KNIME Analytics Platform** for modeling
- Selected **Linear Regression** as the algorithm
- Target variable: **CO_2 Emissions per Capita**
- Input features:
  - Access to electricity (% of population)
  - GDP per capita (current US$)
  - Individuals using the Internet (% of population)
  - School enrollment, secondary (% gross)
  - Unemployment (% of total labor force)
  - Population
  - GDP (from CO_2 dataset)



RESULTS & INTERPRETATION:

- The model achieved an **R^2 score of 0.476**
- This means around **47.6% of the variance** in CO_2 emissions per capita is explained by the selected indicators
- Indicates a **moderate linear relationship** between development indicators and CO_2 emissions
- Suggests that while these indicators are significant, **other external factors** may also influence emissions
- Demonstrates that **low-code tools** like KNIME can produce valuable insights without manual coding

WHY R^2 IS MODERATE (0.476):

- **CO_2 emissions per capita** are influenced by many complex and country-specific factors not captured in the dataset
- Possible missing variables:
  - Industrial activity levels
  - Energy sources (renewables vs fossil fuels)
  - Transportation and urbanization patterns
  - Environmental policies and regulations
- Some input variables had **missing values** that were filled with medians — this can reduce the model's ability to capture true patterns
- The **relationship may not be purely linear**, but linear regression assumes a straight-line relationship

HOW TO IMPROVE THE MODEL:

- **Include more relevant features**, such as:
  - Energy consumption by sector
  - CO_2 emissions by source (transport, industry, etc.)
  - Policy or governance indicators
- **Use more advanced algorithms** like:
  - Random Forest
  - Gradient Boosting (e.g., XGBoost)
  - Support Vector Machines
- **Use feature engineering** to create meaningful derived variables
- **Apply log transformation or polynomial regression** to capture nonlinear relationships better
- **Improve data quality**: reduce null values, and avoid imputation if possible by using more complete datasets