Project Title: Predicting CO_2 Emissions Using Machine Learning
Team ID: SWTID1749709340
Date: 4 July 2025

1. Project Overview

The goal of this project was to develop a machine learning model capable of predicting CO_2 emissions (in kilotons) based on country and year inputs. The project involved data preprocessing, exploratory data analysis (EDA), model training and evaluation, deployment via a web interface, and visualization of key trends in global CO_2 emissions.

2. Dataset Description

The dataset was obtained from the World Bank Indicators dataset, containing over 5 million rows covering multiple countries, years, and development indicators. We filtered it to focus exclusively on the **CO_2 emissions (kt)** indicator.

Key features used:

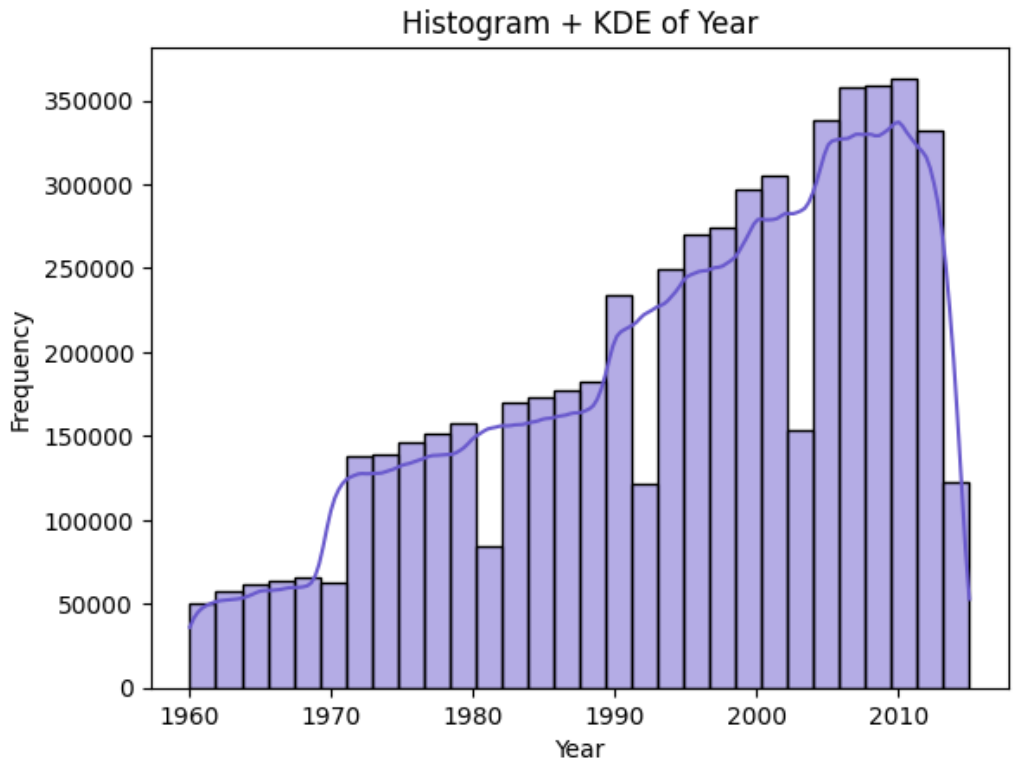
- CountryName
- CountryCode
- IndicatorName
- IndicatorCode
- Year
- Value (Target variable)

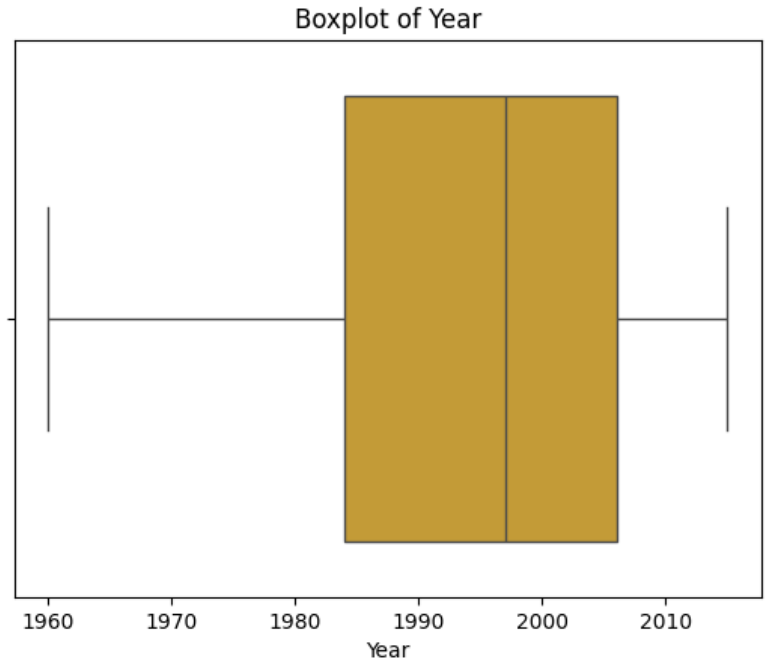
3. Exploratory Data Analysis (EDA)

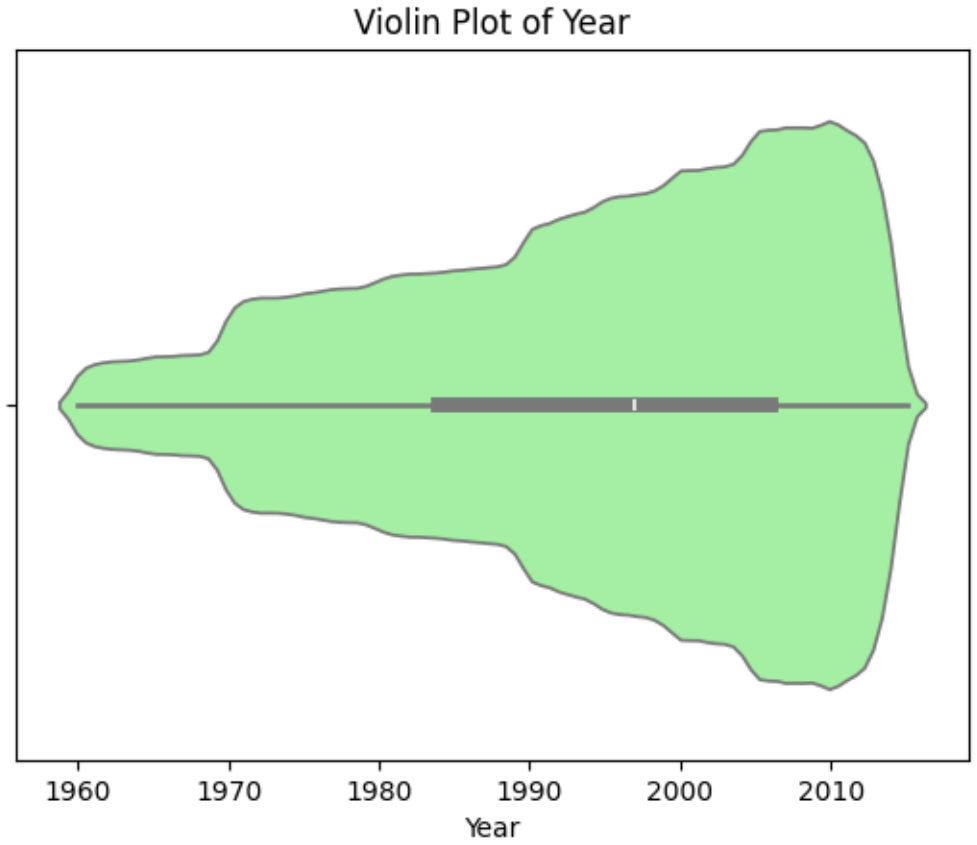
EDA was performed using **Matplotlib** and **Seaborn** to understand trends, patterns, and distributions in CO_2 emissions. The following visualizations were created:

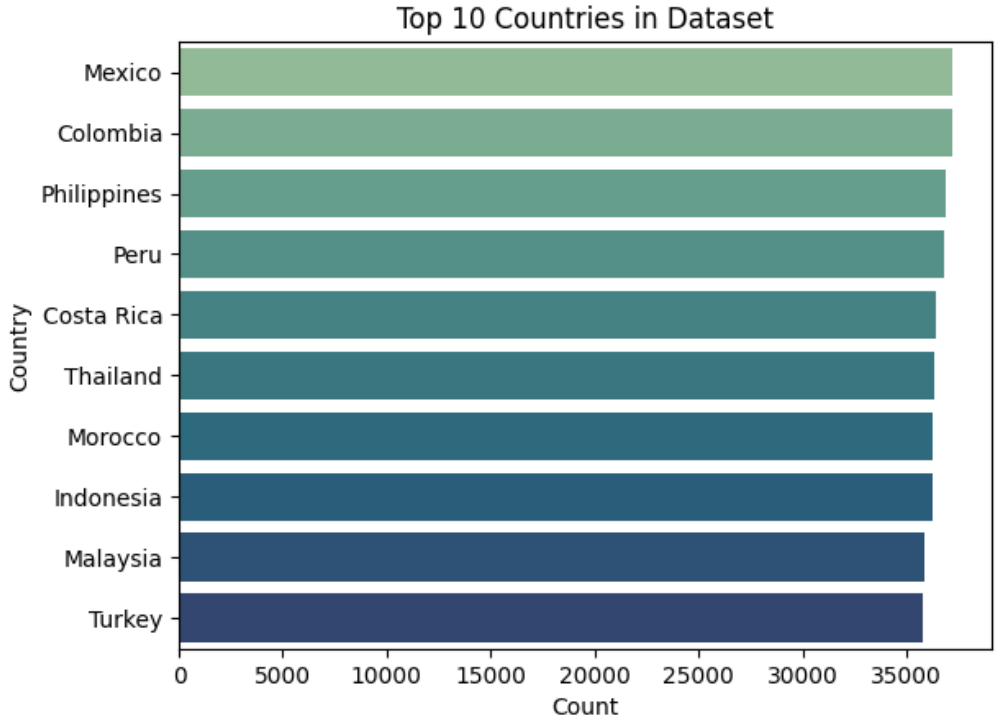
- Bar charts showing top countries by average CO_2 emissions
- Time-series plots for specific countries like India
- Histograms, KDE plots, and violin plots to show distribution
- Correlation heatmaps and pair plots for numeric analysis
- Pie charts to categorize countries based on emission levels

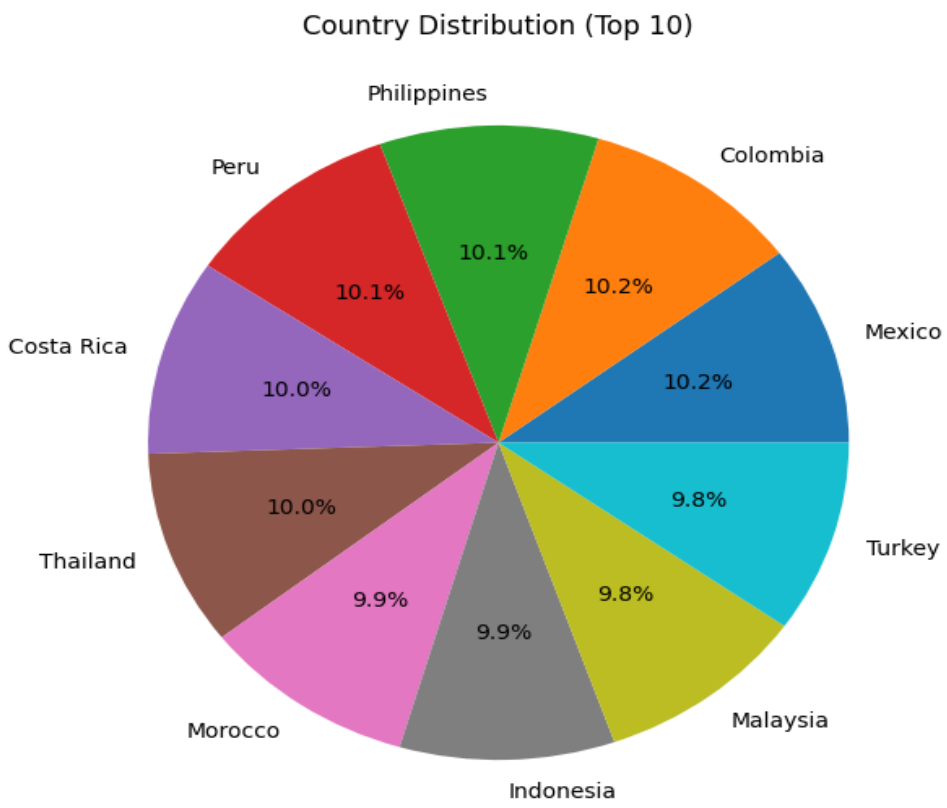
These visuals revealed clear upward trends in emissions for industrial nations and wide variability across countries.

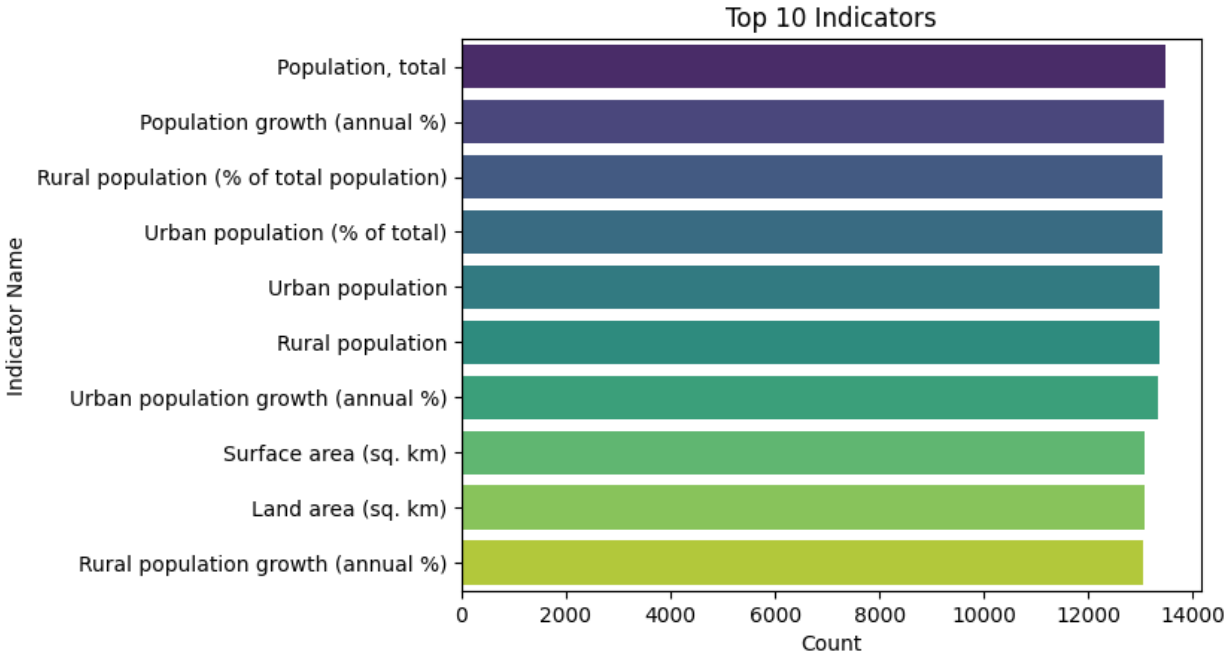

Histogram + KDE of Year
Frequency
350000
300000
250000
200000
150000
100000
50000
0
1960
1970
1980
1990
2000
2010
Year

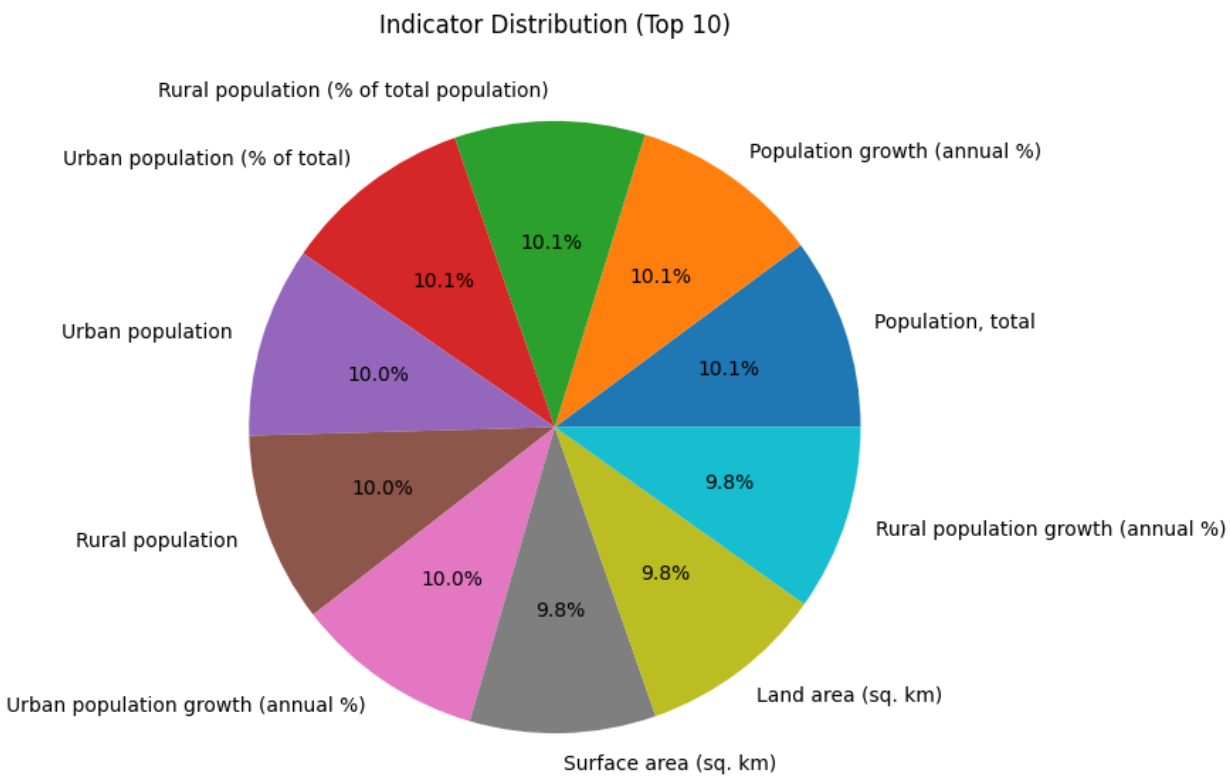

Boxplot of Year
1960
1970
1980
1990
2000
2010
Year

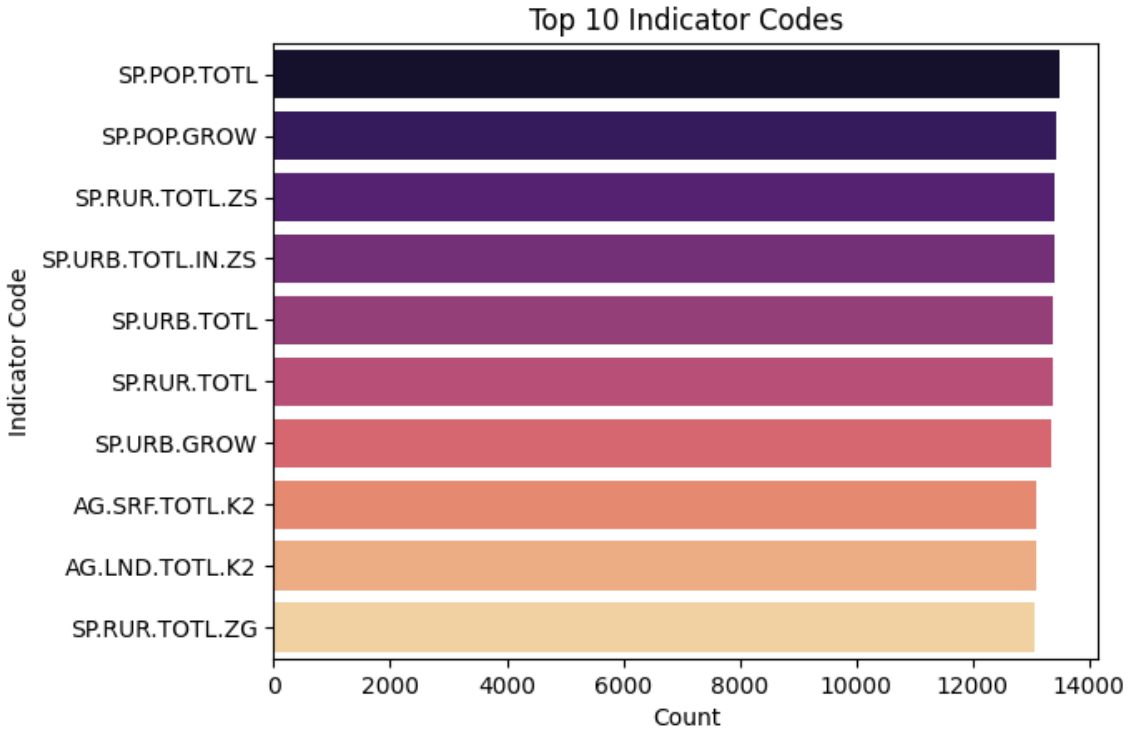

Violin Plot of Year
1960
1970
1980
1990
2000
2010
Year

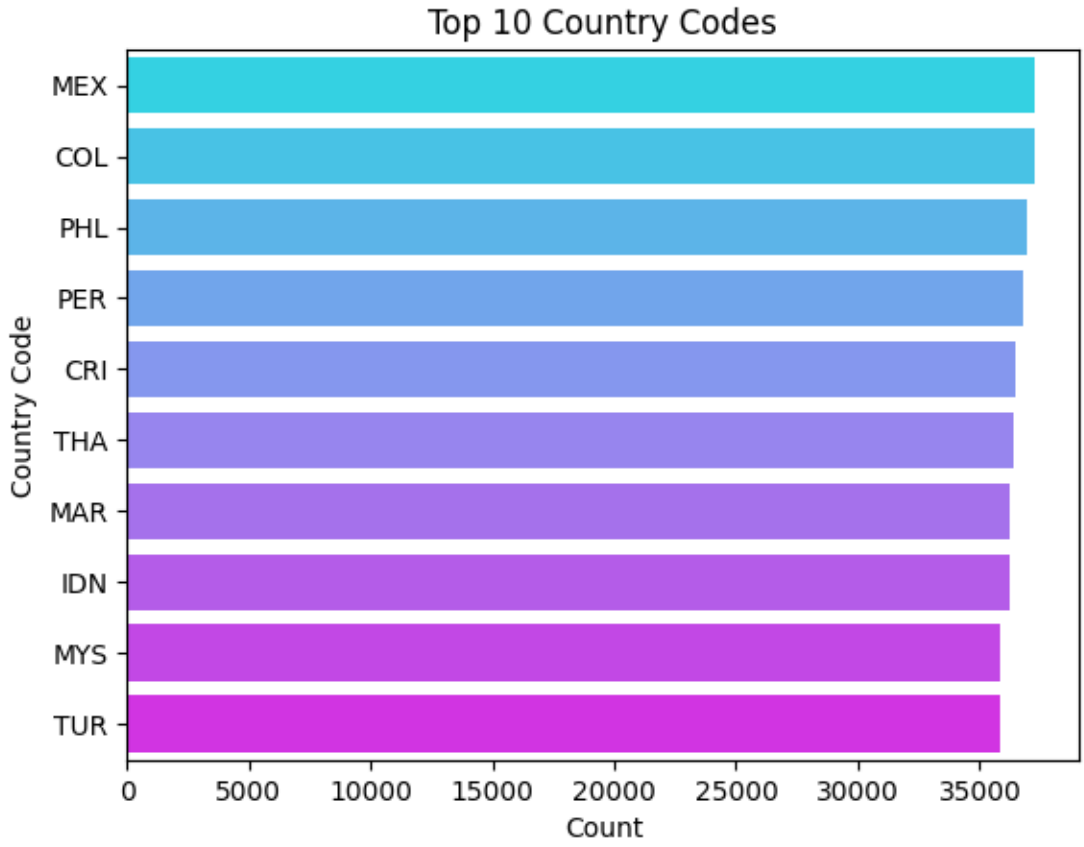

Top 10 Countries in Dataset
Mexico
Colombia
Philippines
Peru
Costa Rica
Thailand
Morocco
Indonesia
Malaysia
Turkey
Country
0
5000
10000
15000
20000
25000
30000
35000
Count

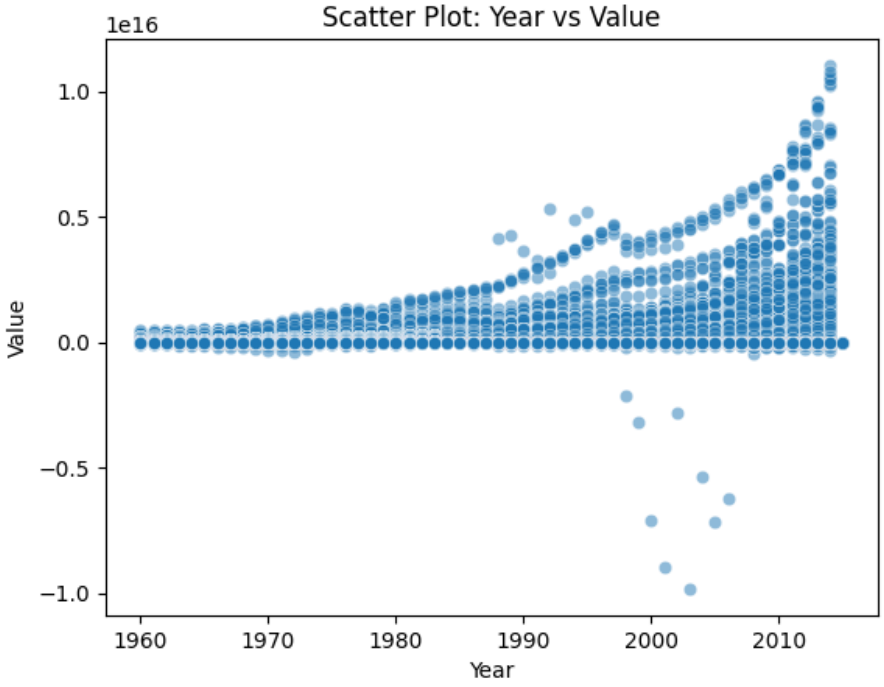

Country Distribution (Top 10)
Philippines
Colombia
Peru
Mexico
Costa Rica
Thailand
Morocco
Malaysia
Indonesia
Turkey
10.1%
10.2%
10.1%
10.2%
10.0%
10.0%
9.9%
9.9%
9.8%
9.8%

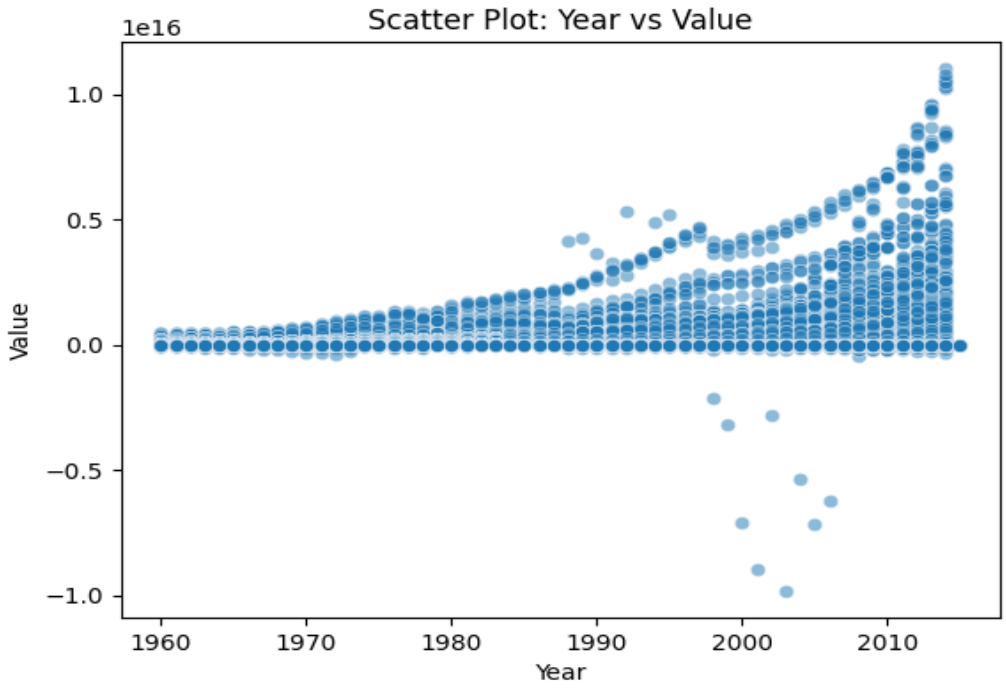

Top 10 Indicators
Indicator Name
Population, total
Population growth (annual %)
Rural population (% of total population)
Urban population (% of total)
Urban population
Rural population
Urban population growth (annual %)
Surface area (sq. km)
Land area (sq. km)
Rural population growth (annual %)
0
2000
4000
6000
8000
10000
12000
14000
Count

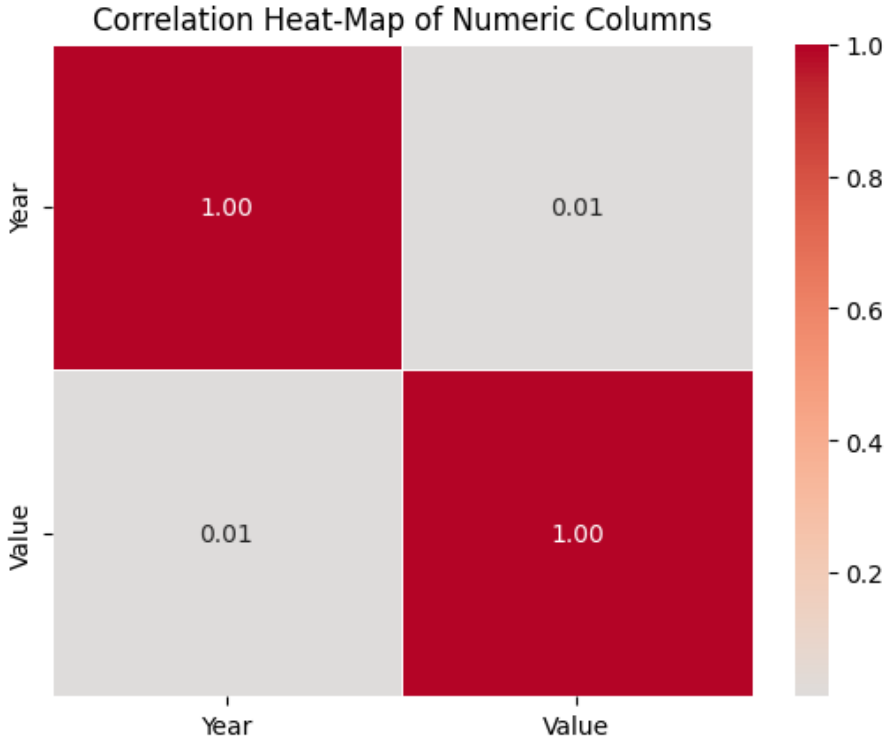

Indicator Distribution (Top 10)
Rural population (% of total population)
Urban population (% of total)
Population growth (annual %)
Urban population
Population, total
Rural population
Rural population growth (annual %)
Urban population growth (annual %)
Land area (sq. km)
Surface area (sq. km)
10.1%
10.1%
10.1%
10.1%
10.0%
10.0%
10.0%
9.8%
9.8%
9.8%

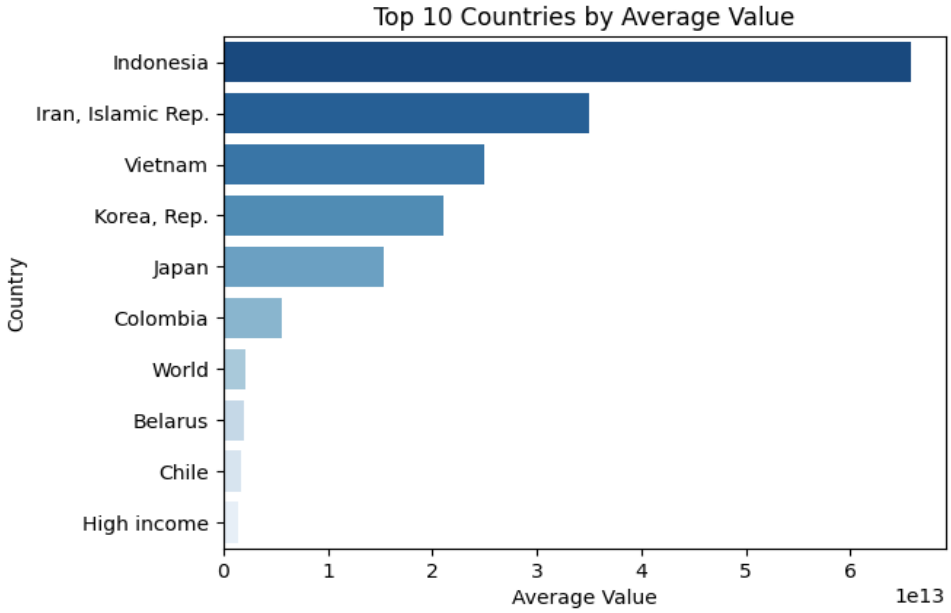

Top 10 Indicator Codes
Indicator Code
SP.POP.TOTL
SP.POP.GROW
SP.RUR.TOTL.ZS
SP.URB.TOTL.IN.ZS
SP.URB.TOTL
SP.RUR.TOTL
SP.URB.GROW
AG.SRF.TOTL.K2
AG.LND.TOTL.K2
SP.RUR.TOTL.ZG
0
2000
4000
6000
8000
10000
12000
14000
Count

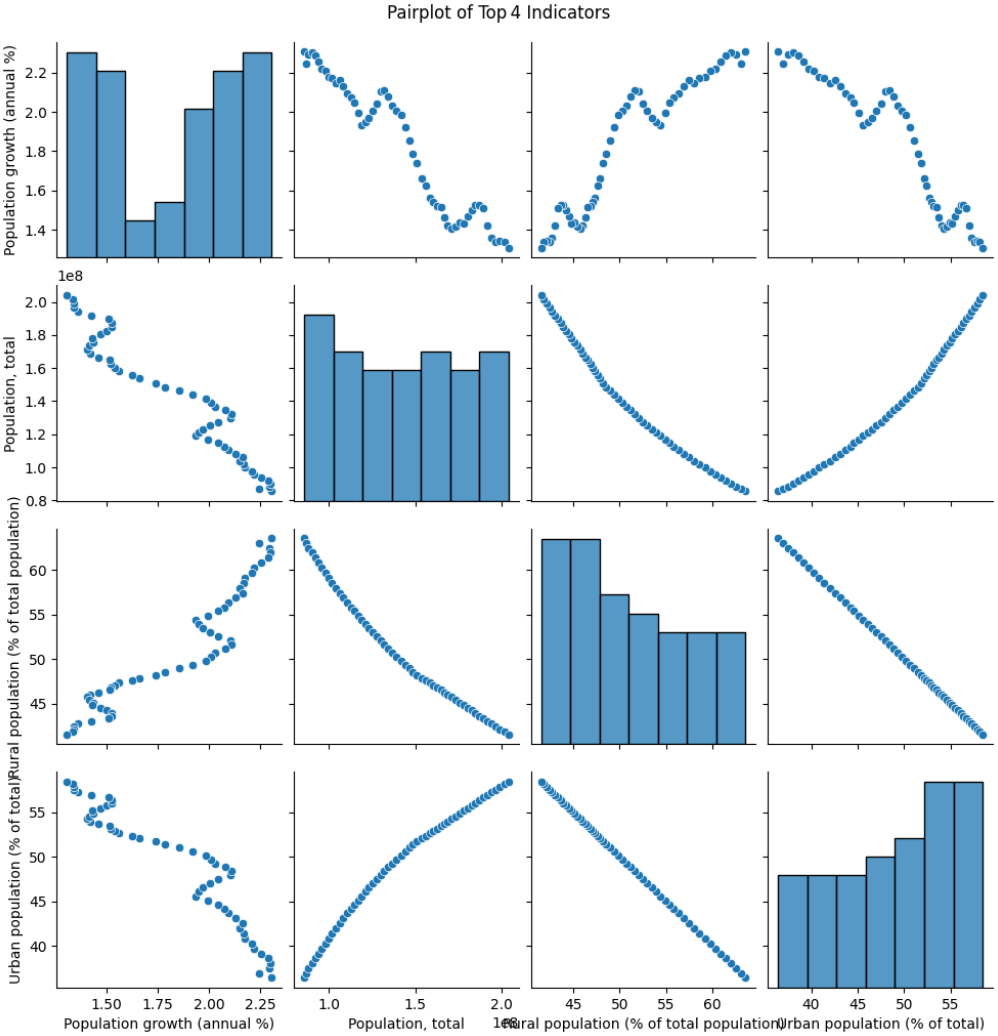

Top 10 Country Codes
MEX
COL
PHL
PER
CRI
THA
MAR
IDN
MYS
TUR
Country Code
0
5000
10000
15000
20000
25000
30000
35000
Count


Scatter Plot: Year vs Value
1e16
1.0
0.5
0.0
−0.5
−1.0
Value
1960
1970
1980
1990
2000
2010
Year


Scatter Plot: Year vs Value
1e16
1.0
0.5
0.0
−0.5
−1.0
Value
1960
1970
1980
1990
2000
2010
Year


Correlation Heat-Map of Numeric Columns
1.00
0.01
0.01
1.00
Year
Value
Year
Value
1.0
0.8
0.6
0.4
0.2


Top 10 Countries by Average Value
Indonesia
Iran, Islamic Rep.
Vietnam
Korea, Rep.
Japan
Colombia
World
Belarus
Chile
High income
Country
0
1
2
3
4
5
6
Average Value
1e13


Pairplot of Top 4 Indicators
Population growth (annual %)
Population, total
1e8
Rural population (% of total population)
Urban population (% of total)
1.50
1.75
2.00
2.25
1.0
1.5
2.0
45
50
55
60
40
45
50
55

4. Data Preprocessing

The preprocessing steps included:

- Dropping missing or irrelevant entries
- Encoding categorical variables (CountryName, IndicatorCode, etc.) using **One-Hot Encoding**
- Scaling numerical features like Year using **StandardScaler**
- Using **ColumnTransformer** to combine preprocessing steps in a clean pipeline

5. Model Deployment

We used **Pickle** to save the trained Random Forest model and the label encoder.
The deployment was implemented as a **Flask web app** with:

- HTML frontend (user inputs country and year)
- Prediction endpoint
- **Ngrok** tunneling to serve the site publicly via temporary HTTPS URL (for demonstration)

This app allowed real-time prediction of CO_2 emissions for any country-year pair.

6. Tools & Technologies Used

- **Python (Pandas, scikit-learn, XGBoost)**
- **Matplotlib, Seaborn** for visualization
- **Flask + HTML** for deployment
- **Ngrok** for hosting the web app
- **Jupyter Notebook / Google Colab** for development
- **Pickle** for model persistence

7. Conclusion & Learnings

This project demonstrated how data science and machine learning can be used to model and visualize environmental trends like CO_2 emissions. By building a deployable prediction tool, we showcased the end-to-end ML pipeline from data ingestion to real-world application. Key takeaways included:

- Importance of EDA and preprocessing
- Trade-offs between model performance and interpretability
- Real-world deployment using minimal tools (Flask + Ngrok)