Determining if Regions Accurately Predict the Per Capita CO_2 Emissions

1. Project Objectives:

- Explore trends in global CO_2 emissions over time
- Analyze per capita CO_2 emission patterns by region
- Identify regions with disproportionate contributions to climate change

Research Question: How have global CO_2 emissions changed over time, and how do regional per capita emissions compare between regions?

2. Dataset:

Carbon (CO2) Emissions

Source: https://www.kaggle.com/datasets/ravindrasinghrana/carbon-co2-emissions

3. Findings:

Findings from Graph 1 (Line Plot):

- Global CO2 emissions have increased over the last several decades
- A clear upward trend, especially in the early 2000s, is consistent with industrialization and worldwide economic growth.
- This supports the idea that human activity is accelerating climate change.

The data confirms a significant increase in emissions over time and reinforces the existence of global warming

Findings from Graph 2 (Heatmap):

- Africa has the lowest and most stable per capita emissions, indicating a minimal individual contribution to global emissions.
- Europe and Asia have had higher per capita emissions, though Europe is declining over time.
- Asia shows rising per capita emissions, reflecting growing industrialization.

From this non-trivial information, we see that not all regions contribute equally to climate change, especially on a per-person basis.

4. Tools:

Dataset Metadata:

Command-Line Tool: awk to Summarize CO_2 by Year

```
awk -F',' 'NR > 1 { split($3, d, "-"); sum[d[3]] += $4 } END { for (y in sum) print y ","
sum[y] }' Carbon_\(CO2\)_Emissions_by_Country.csv > co2_by_year.csv
```

Command for Number of Entries (Rows)

```
wc -l Carbon_\(CO2\)_Emissions_by_Country.csv
```

Command for Number of Features (Columns)

```
head -n 1 Carbon_\(CO2\)_Emissions_by_Country.csv | awk -F',' '{print NF}'
```

PySpark (Spark SQL & DataFrame API):

Spark handled large-scale data processing. We:

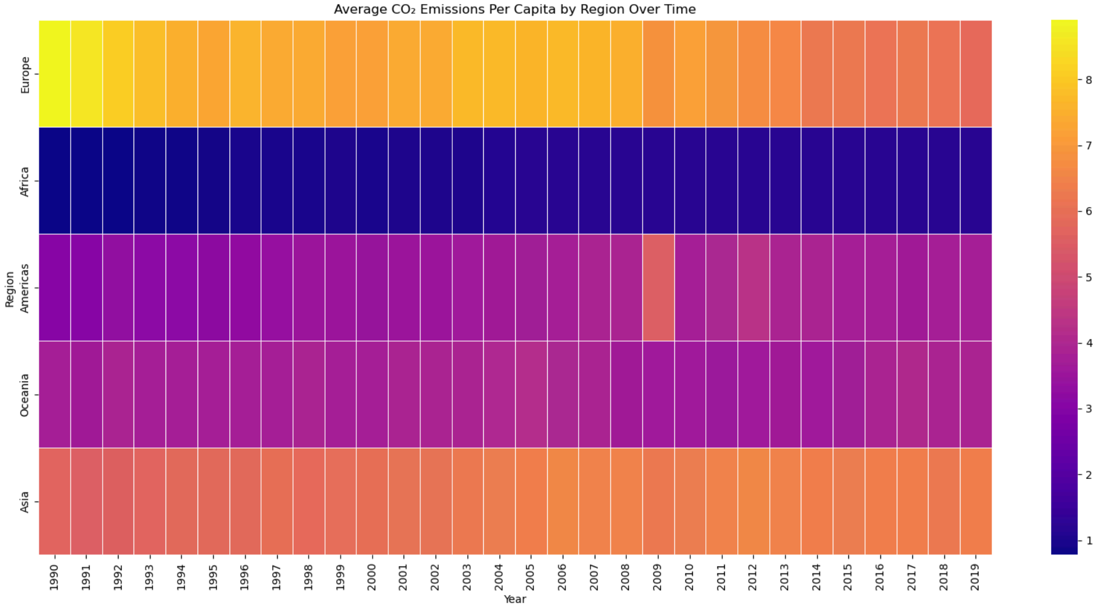
- Parsed date fields to extract the year
- Made a line plot showing the correlation between the year and CO2 emissions
- Grouped data by year to calculate total global emissions
- Grouped by region and year to compute average per capita CO2 emissions
- Pivoted the data into a matrix format for the heatmap visualization

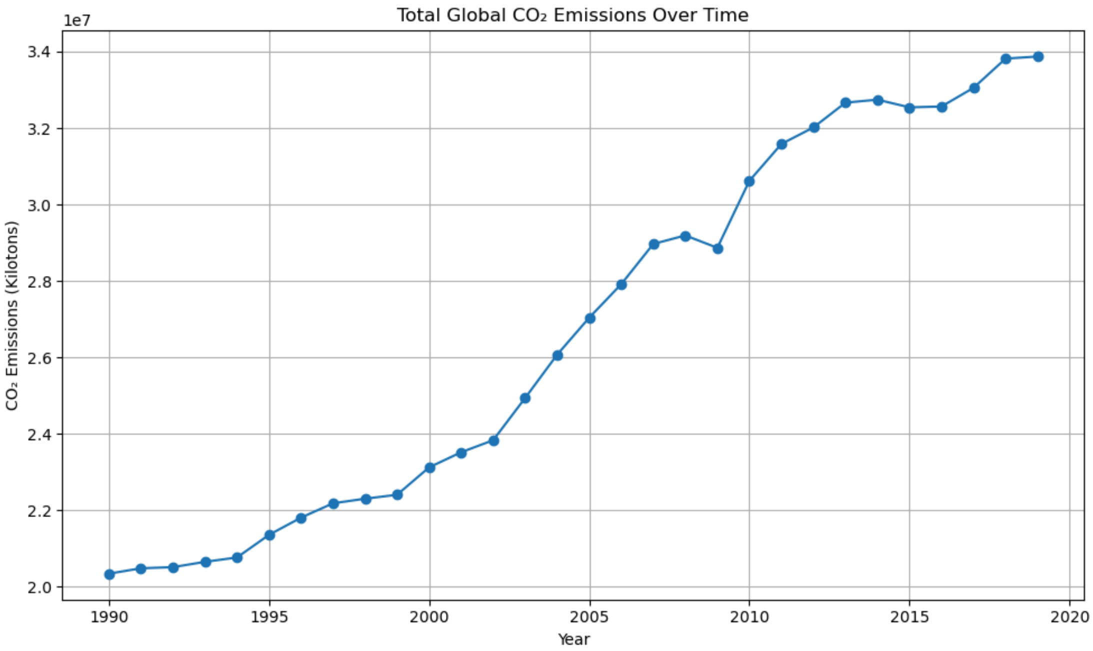
Python (Pandas, Matplotlib, Seaborn):

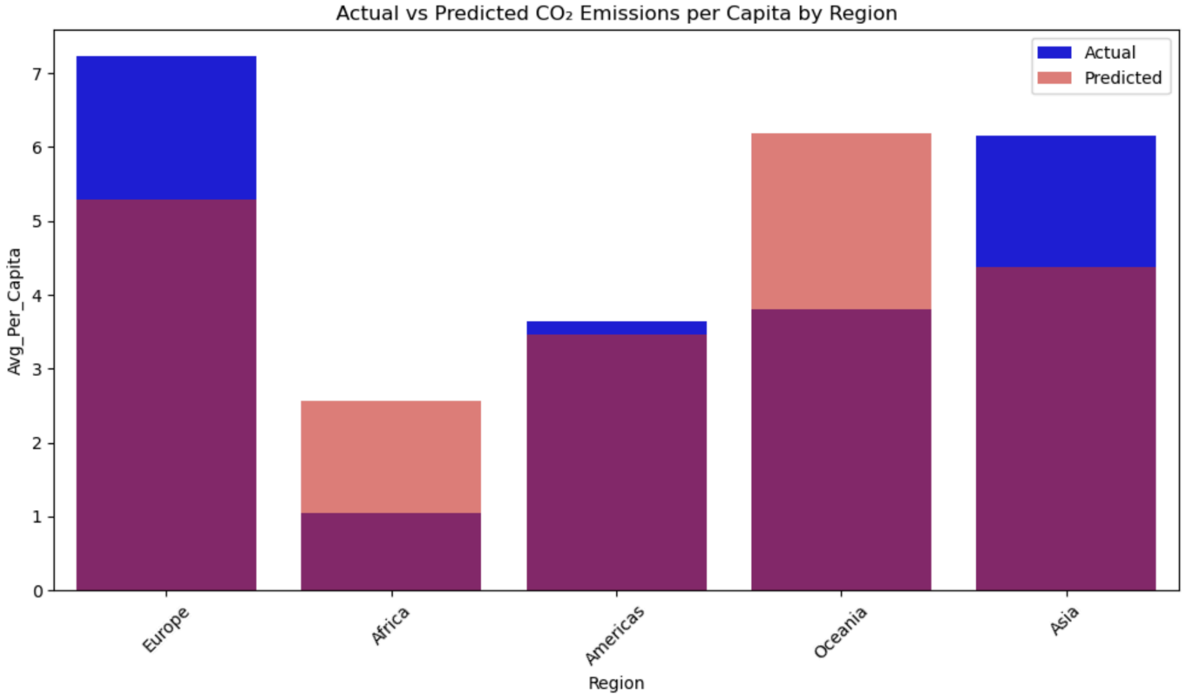
After transforming data with Spark, we converted the results to Pandas DataFrames to create:

- A line plot showing global emissions trends over time
- A heatmap of regional per capita emissions over time

5. Three plots







- Blue bars represent the actual average CO_2 emissions per capita for each region.
- Red bars represent the model’s predictions for those averages.

- The height of each bar reflects the emission value.
- If the bars align closely, the model's predictions are good.
- When the red (predicted) bar is much higher or lower than the blue (actual), that's a sign of prediction error for that region.

Summary:
Through data manipulation and modeling, we analyzed average CO_2 emissions per capita across global regions. First, we aggregated emissions data by region to calculate the mean metric tons of CO_2 emitted per person. Then, using a machine learning pipeline in PySpark, we trained a linear regression model to predict these regional averages based solely on categorical region data. The model's performance, evaluated using Root Mean Squared Error (RMSE), showed a prediction error of approximately 0.73 metric tons. The bar chart showed different model accuracy across the regions, with some predictions closely aligning (Americas, Asia), while others deviated a lot (Africa, Oceania). This indicates that regional identity alone may not be a strong predictor of per capita CO_2 emissions, suggesting that incorporating additional features like GDP or energy use could improve model accuracy. This aligns with what we originally thought, that names of regions alone can't determine emissions, but it was interesting to explore this fun idea.