INFO 201 Final Project

By: Alexander Huynh, Asmit Sinha, Camilo Montes De Haro, Jessica Xiong

Introduction

- Our work aims to provide policymakers with evidence-based insights that can guide the development of effective climate policies and strategies.
- By understanding the factors driving greenhouse gas emissions, legislators can make informed decisions to address the pressing issue of climate change and work towards a more sustainable future.
- Our target group comprises legislators and politicians who have the authority to enact policies and allocate resources to combat climate change effectively.
- The project aims to explore the correlation between CO2 emissions and relating factors such as deforestation, electricity use, and urbanization.
- By analyzing datasets containing relevant variables for different countries, we seek to identify patterns and relationships that can inform policy making decisions.

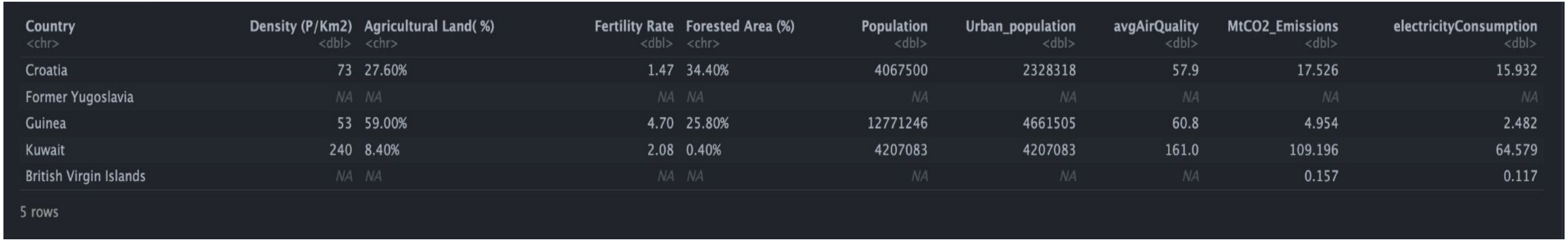
Questions:

Questions that our team chose to explore with this project are:

1. Does the amount of forested area in different countries affect the air quality of those countries?
2. Is there a relationship between a country’s electricity consumption and air quality?
3. Is there a correlation between population and/or urban population with CO2 emissions?
4. Is there a correlation between fertility rates and CO2 emissions?

Data

- We used and merged 4 different datasets containing general information about countries in the world, air quality indices for each country, global electricity statistics, and CO2 emissions per country.
- Most of our data (World Air Quality Index, Countries of the World 2023, Global Electricity Statistics) are from Kaggle. However, we got data about global CO2 emissions per country from The Global Carbon Project.
- Overall, the quality of our data was pretty good. However, we have missing NA values throughout our datasets, with about 50-60 countries containing missing values for at least some variable.
- We didn't see any prominent ethical issues regarding the usage of our data. All of the information in our data are free for educational and research purposes, and contain information widely available from reputable and ethical institutions such as World Bank.

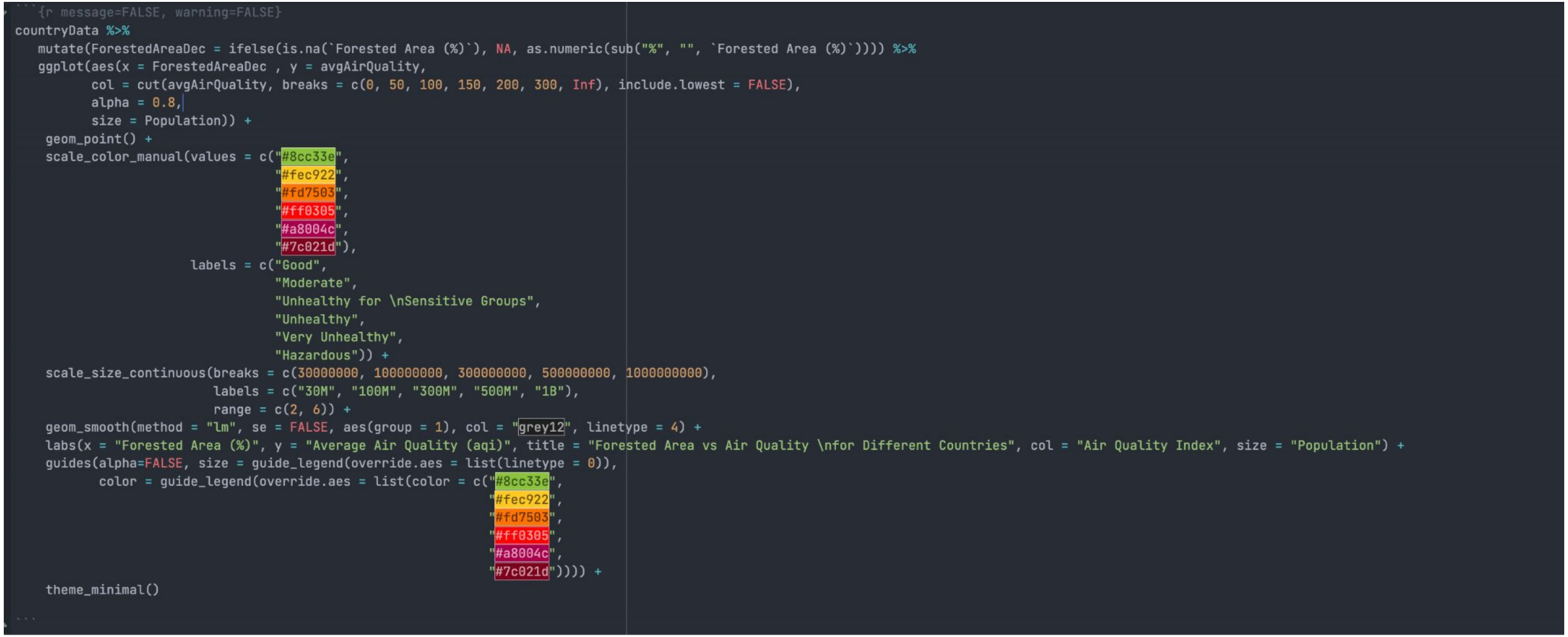


Country	Density (P/Km2)	Agricultural Land(%)	Fertility Rate	Forested Area (%)	Population	Urban_population	avgAirQuality	MtCO2_Emissions	electricityConsumption
<chr>	<dbl>	<chr>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Croatia	73	27.60%	1.47	34.40%	4067500	2328318	57.9	17.526	15.932
Former Yugoslavia	NA	NA	NA	NA	NA	NA	NA	NA	NA
Guinea	53	59.00%	4.70	25.80%	12771246	4661505	60.8	4.954	2.482
Kuwait	240	8.40%	2.08	0.40%	4207083	4207083	161.0	109.196	64.579
British Virgin Islands	NA	NA	NA	NA	NA	NA	NA	0.157	0.117

5 rows

Methods

- Samples: We excluded every country missing at least one value for the attributes we were comparing.



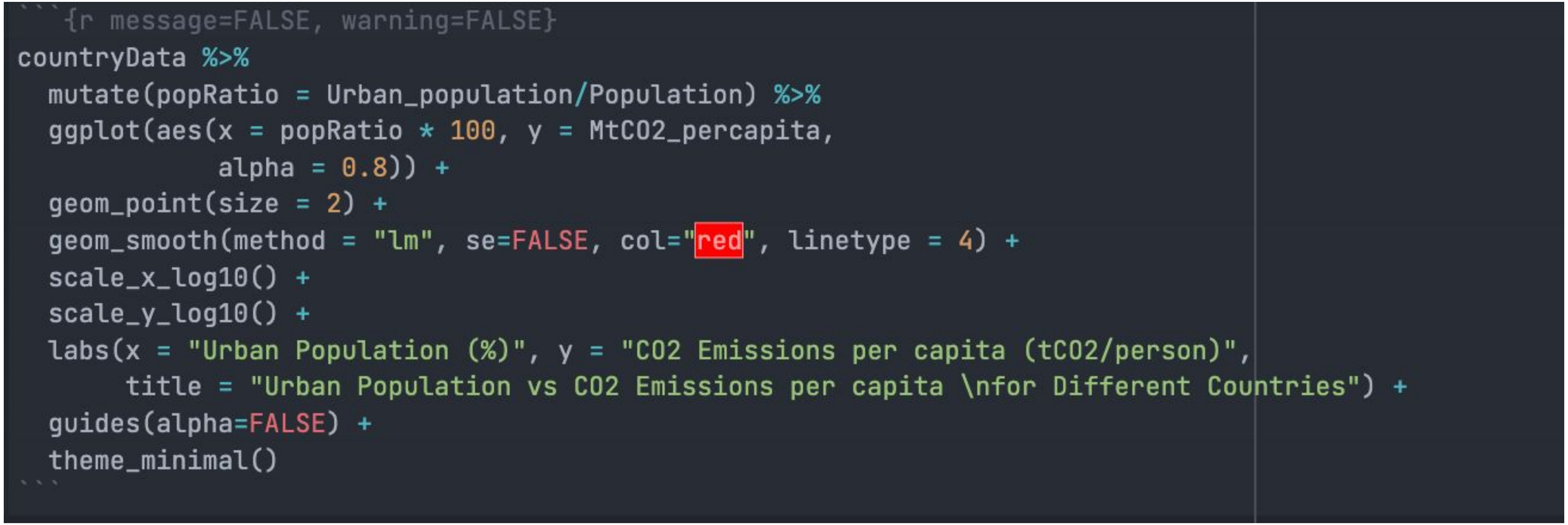
~~~
```{r message=FALSE, warning=FALSE}
countryData %>%
   mutate(ForestedAreaDec = ifelse(is.na(`Forested Area (%)`), NA, as.numeric(sub("%", "", `Forested Area (%)`)))) %>%
   ggplot(aes(x = ForestedAreaDec , y = avgAirQuality,
          col = cut(avgAirQuality, breaks = c(0, 50, 100, 150, 200, 300, Inf), include.lowest = FALSE),
          alpha = 0.8,
          size = Population)) +
    geom_point() +
    scale_color_manual(values = c("#8cc33e",
                                  "#fec922",
                                  "#fd7503",
                                  "#ff0305",
                                  "#a8004c",
                                  "#7c021d"),
                       labels = c("Good",
                                  "Moderate",
                                  "Unhealthy for \nSensitive Groups",
                                  "Unhealthy",
                                  "Very Unhealthy",
                                  "Hazardous")) +
    scale_size_continuous(breaks = c(30000000, 100000000, 300000000, 500000000, 1000000000),
                          labels = c("30M", "100M", "300M", "500M", "1B"),
                          range = c(2, 6)) +
    geom_smooth(method = "lm", se = FALSE, aes(group = 1), col = "grey12", linetype = 4) +
    labs(x = "Forested Area (%)", y = "Average Air Quality (aqi)", title = "Forested Area vs Air Quality \nfor Different Countries", col = "Air Quality Index", size = "Population") +
    guides(alpha=FALSE, size = guide_legend(override.aes = list(linetype = 0)),
           color = guide_legend(override.aes = list(color = c("#8cc33e",
                                                              "#fec922",
                                                              "#fd7503",
                                                              "#ff0305",
                                                              "#a8004c",
                                                              "#7c021d")))) +

    theme_minimal()

```
~~~

Methods

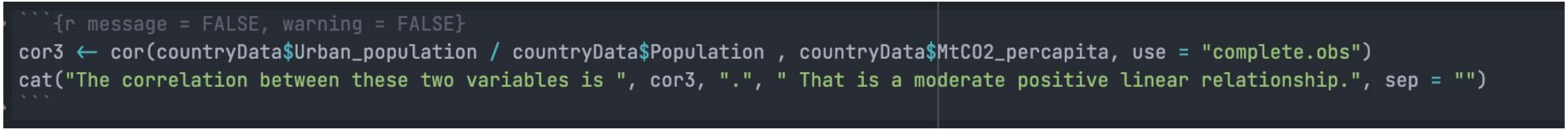
- We used per capita values for a clearer picture



~~~
```{r message=FALSE, warning=FALSE}
countryData %>%
  mutate(popRatio = Urban_population/Population) %>%
  ggplot(aes(x = popRatio * 100, y = MtCO2_percapita,
             alpha = 0.8)) +
  geom_point(size = 2) +
  geom_smooth(method = "lm", se=FALSE, col="red", linetype = 4) +
  scale_x_log10() +
  scale_y_log10() +
  labs(x = "Urban Population (%)", y = "CO2 Emissions per capita (tCO2/person)",
       title = "Urban Population vs CO2 Emissions per capita \nfor Different Countries") +
  guides(alpha=FALSE) +
  theme_minimal()
```
~~~

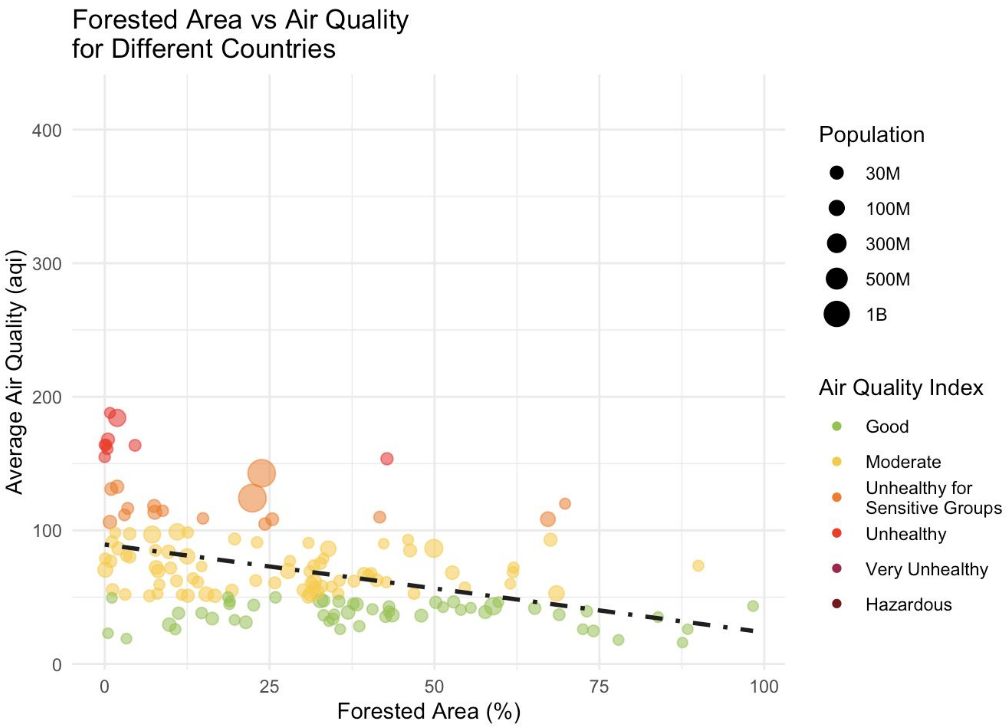
Methods

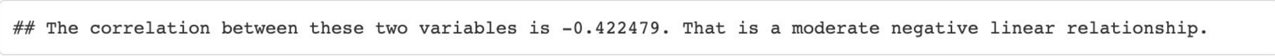
Correlation Coefficient:



~~~
```{r message = FALSE, warning = FALSE}
cor3 ← cor(countryData$Urban_population / countryData$Population , countryData$MtCO2_percapita, use = "complete.obs")
cat("The correlation between these two variables is ", cor3, ".", " That is a moderate positive linear relationship.", sep = "")
```
~~~

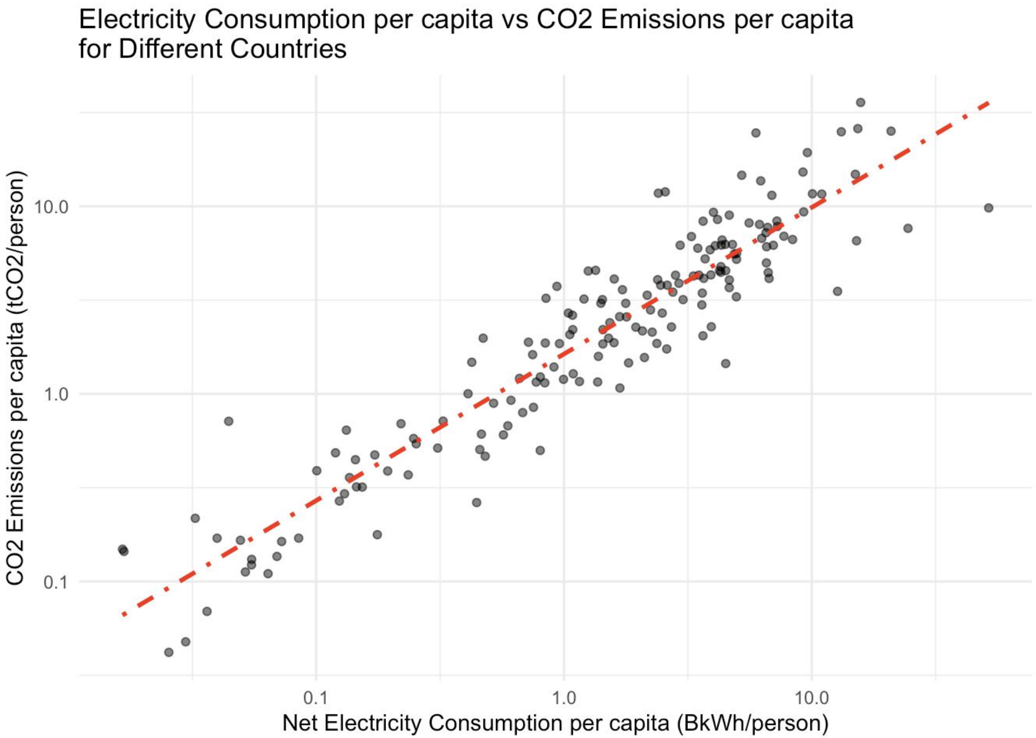
Forested Area vs. Air Quality





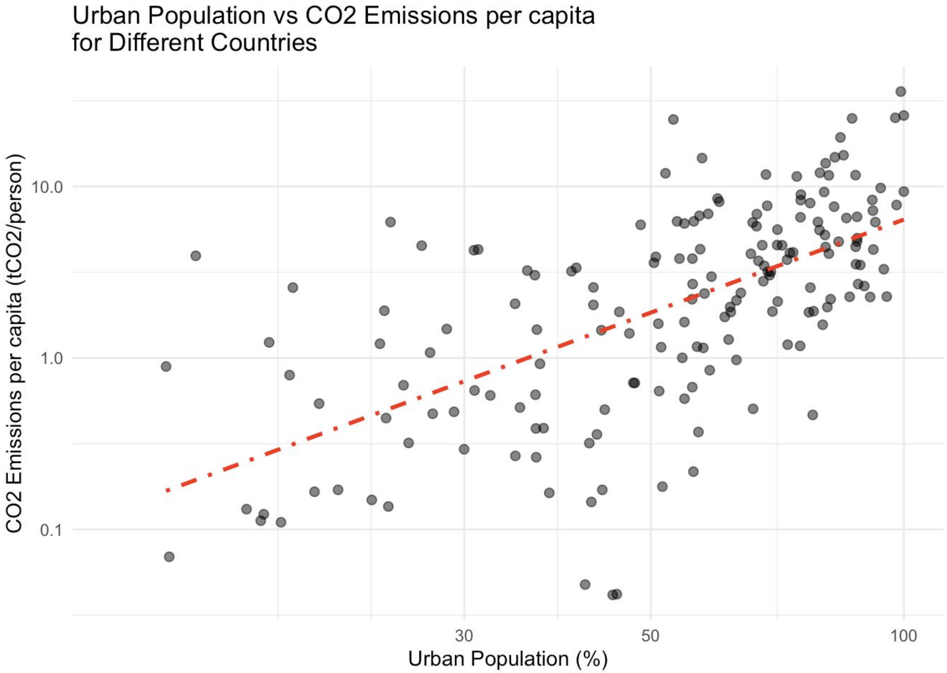
```
## The correlation between these two variables is -0.422479. That is a moderate negative linear relationship.
```

Electricity Consumption vs. CO2 Emissions



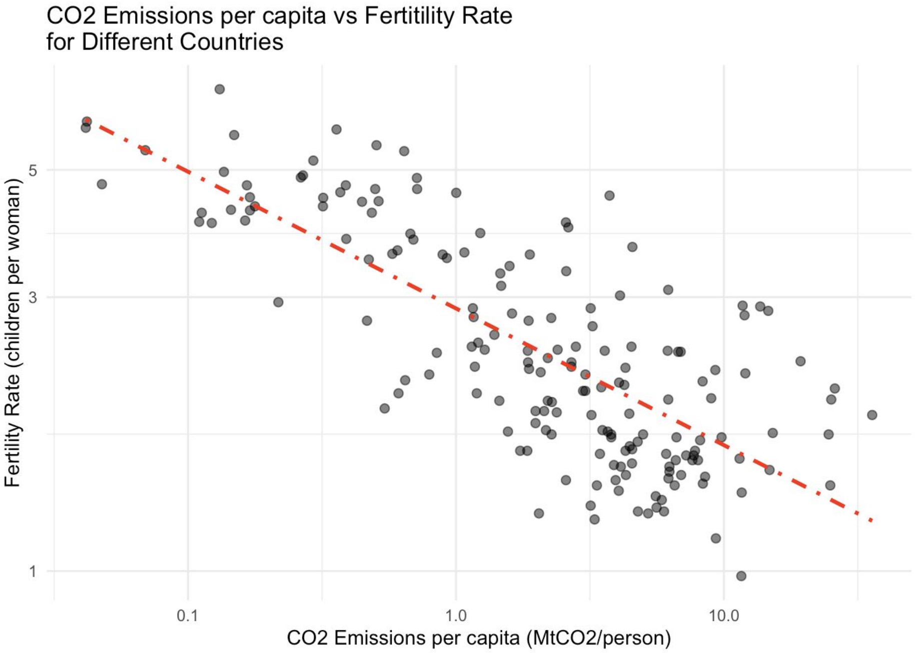
```
## The correlation between these two variables is 0.6076909. That is a relatively strong positive linear relations
hip.
```

Urban Population vs. CO2 Emissions



```
## The correlation between these two variables is 0.5084871. That is a moderate positive linear relationship.
```

CO2 Emissions vs Fertility Rate



CO2 Emissions per capita vs Fertitility Rate
for Different Countries

```
## The correlation between these two variables is -0.4318185. That is a moderate negative linear relationship.
```

Results

- The results show that using more electricity in cities and having more people living in urban areas tend to lead to higher CO2 emissions. On the other hand, places with higher birth rates and more forest areas tend to have lower CO2 emissions.
- Currently, the data connecting each variable to CO2 emissions shows correlation but not causation.
- There could be other important factors that we didn't include, which might influence the results differently.
    i. Industrial processes
    ii. Transportation
    iii. Policy
- Using our existing variables in conjunction with more specific variables would help reinforce our claims.
- To extend our understanding, we may require access to more datasets covering a wider range of variables and regions.
- Our analysis demonstrates the complex relationship between greenhouse gas emissions and various socio-economic and environmental factors. While our findings provide valuable insights, further research is needed to understand the reasons behind these correlations and identify targeted solutions to combat CO2 emissions effectively.