Research Document: Analysis of CO2 Emissions and Influencing Factors

1. Introduction to the Topic

1.1 Background and Objectives

Climate change remains one of humanity's greatest challenges, with CO2 emissions from industrial activities, transportation, and urbanization playing a central role. This study focuses on analyzing CO2 emission trends in major countries such as China, India, and the United States from 1990 to 2018, aiming to assess environmental impacts and driving factors. The primary objectives are:

- Analyze CO2 emission trends across decades.
- Utilize population, GDP, and vehicle sales data as supporting factors to understand the causes of emission growth or control.
- Propose mitigation strategies based on data analysis.

1.2 Significance

The rise in CO2 emissions in developing nations like China and India reflects rapid economic growth, while the United States demonstrates control through technology and policy. Understanding these dynamics helps shape global sustainable strategies, especially given the increasing role of transportation (vehicle sales) and population growth.

1.3 Scope of Research

- **Timeframe**: From 1990 to 2018 for CO2, with population data from 1970–2021 and vehicle sales from 1990–2018 (if available).
- **Location**: Focus on China, India, the United States, with additional countries like Indonesia and Pakistan for population data.
- **Methodology**: Trend analysis based on historical Kaggle data and visualization to derive conclusions.

---

2. Selected Datasets

2.1 Data Sources

The datasets used in this research are sourced from Kaggle, a platform providing open datasets, and are stored in the project directory at C:\Users\PC\jupyternotebook\project\finalproject\data\historical\. The datasets include:

- **historical_emissions.csv (Global CO2 Emissions)**
  - **Description**: Contains annual CO2 emission data from 1990 to 2018, measured in million metric tons of CO2 equivalent ($MtCO_2e$), serving as the core of the study.
  - **Structure**: Columns such as Country, Data source, Sector, Gas, Unit, and year columns (1990, 1991, ..., 2018).
  - **Source**: Kaggle dataset, possibly "Global Carbon Dioxide Emissions" or similar (specific link to be confirmed).
  - **Role**: Primary data for analyzing CO2 emission trends.
- **csvData.csv (World Population Density)**
  - **Description**: Contains population data from 1980 to 2021, including China, India, United States, Indonesia, and Pakistan, with 1970–1979 estimated based on 1980 values.
  - **Structure**: Columns such as name (country name), pop1980, pop1990, ..., pop2021 (population by year, in millions).
  - **Source**: Kaggle dataset, possibly "World Population Dataset" or similar (specific link to be confirmed).
  - **Role**: Supporting data to assess population impact on emissions.
- **Countries GDP 1960-2020.csv (World GDP)**
  - **Description**: Contains annual GDP data from 1960 to 2020, in USD, with 1990–2018 utilized.
  - **Structure**: Columns such as Country Name, Country Code, and year columns (1990, 1991, ..., 2020).
  - **Source**: Kaggle dataset, possibly "World Bank GDP Data" or similar (specific link to be confirmed).
  - **Role**: Supporting data to analyze the relationship between economic growth and CO2.
- **scraped_data.csv (Vehicle Sales Data)**
  - **Description**: Contains estimated or actual data on vehicle sales by country and year, from 1990 to 2018 (if available).
  - **Structure**: Expected to include columns such as Country, Year, and Vehicle_Sales (or similar names), with units in number of vehicles sold.
  - **Source**: Kaggle dataset, possibly "Global Vehicle Sales" or data scraped from online sources (specific link to be confirmed).
  - **Role**: Supporting data to evaluate the impact of transportation on CO2 emissions, requiring verification of actual structure.

2.2 Data Characteristics

- **CO2 Emissions**: Data from 1990–2018 reflects emission trends from industry and transportation, ranging from hundreds to tens of thousands of $MtCO_2e$.
- **Population**: Data from 1980–2021 is accurate, with 1970–1979 estimated.
- **GDP**: Data from 1990–2018, ranging from millions to trillions of USD.
- **Vehicle Sales**: Data requires verification but is assumed to reflect vehicle sales quantities, potentially linked to CO2.

2.3 Data Limitations

- **CO2 Emissions**: May lack detailed sectoral data (only aggregated).
- **Population**: Missing data from 1970–1979, leading to estimates based on 1980.
- **Vehicle Sales**: Structure of scraped_data.csv is unclear, requiring column verification (Year, Vehicle_Sales).
- **Consistency**: Datasets may not fully align in terms of years or countries.

2.4 Methodology

- Data is processed using Python with the Pandas library for cleaning and formatting.
- Visualization is performed using Plotly to create line charts, facilitating the comparison of CO2 trends with supporting factors.