Global Carbon Emission Patterns (COMP3125 Individual Project)

School of Computing and Data Science

***Abstract* – Nowadays, carbon emissions is one of the most talked global concerns, which cause climate change, influence environmental policies and economic sustainability. This project analyzes carbon emissions across various countries using datasets that contain carbon emissions and GDP figures. The goal of this project is to analyze patterns between economic growth and emissions, and categorize countries based on their carbon footprint.**

With the help of Python, we will use statistical analysis, data visualization and machine learning techniques such as linear regression and K-means clustering to analyze results. The study highlights a big difference in emission per capita, the relationship between GDP and CO2 emissions. The results will offer great source of knowledge for policymakers, researchers and economists.

Keywords: CO_2 emissions, GDP, climate change, data analysis, machine learning

Introduction (Heading 1)

The Industrial revolution changed everything in the world. With the invention of internal combustion engines, as wood, peat and coal were not efficient enough, need for more efficient and versatile fuel types arose. As petroleum was more efficient and easily transportable, it became the main fuel type all over the world very soon. With the oil boom countries with huge oil reserves started to extract the oil in huge amounts. However, even though it was very efficient as a fuel type, it was very dangerous for nature. This individual project aims to analyze the relationship between the biggest carbon emitters and their GDP.

Datasets

A. Source of dataset(Heading 2)

In this project I used 2 datasets. I obtained CO2 emissions dataset from Kaggle, and GDP per country dataset from MarketWatch. Both of these sources are credible and widely used. Here is more information:

1. Carbon Emissions by country – From 1990 to 2019
2. Gross Domestic Product by country – From 1960 to 2023

B. Characteristics of datasets

Both datasets were downloaded in .csv format. GDP by country dataset needed some preprocessing.

Dataset	Variables	Units	Processing steps
CO_2 emissions by country	Country, Region, Date	Kilotons of Co2, Metric Tons Per Capita	None
GDP per country	Country name and code, Indicator name and code, years	Dollar amount	Skipped first 4 rows

Used Pandas to merge CO2 emissions dataset with the GDP dataset on the Country and Date columns. I also used Seaborn hue for visualization. I did not perform any unit conversions, as all the values were used in their original form. The data cleaning process involved cleaning metadata rows from GDP dataset where I skipped first four rows and I also handled missing values ensuring consistency across the dataset before analysis.

Methodology

A. K-Means Clustering Analysis

K-Means clustering was used to identify patterns in **CO_2 emissions across countries**. This method helps categorize countries into distinct groups based on their average emissions.

Implementation:

```
X_cluster = df.groupby('Country')[['Kilotons of Co2']].mean()

kmeans = KMeans(n_clusters=3, random_state=42, n_init=10)

kmeans.fit(X_cluster)

X_cluster['Cluster'] = kmeans.labels_
```

The **advantages** of K-Means clustering include its **simplicity, efficiency, and ability to uncover patterns in emissions data**. However, a **key limitation** is that the number of clusters must be specified in advance, rather than being determined automatically.

B. Correlation Analysis

This analysis examines the relationship between **GDP and CO_2 emissions** to understand how economic activity influences environmental impact.

Implementation:

```
merged_df = df.merge(df_gdp, on=['Country', 'Date'], how='inner')
X = merged_df[['GDP']]
y = merged_df['Kilotons of Co2']
model = LinearRegression()
model.fit(X, y)
```

The **advantages** of this method include its **ease of implementation and clear visualization of economic-environmental trends**. However, its **limitations** include detecting only **linear relationships** and being **sensitive to outliers** in the data.