COMP5716F: Programming for Data Science

Coursework 1: Data Analysis Project

Insights for Carbon dioxide emission

- Chinedu Okonkwo: okonkwo@uea.ac.uk
- Ifeanyi Ezeonyeka: ezeonyeka@uea.ac.uk
- Ifeoma Okonkwo: okonkwo@uea.ac.uk
- Obiyesuo Gabriel: obiyesuo@uea.ac.uk

Project Plan

The Data

The dataset under consideration is a comprehensive repository of global carbon dioxide emissions, with a focus on emissions of greenhouse gases (GHG) across various sectors and countries. This dataset encompasses detailed records of GHG emissions, including carbon dioxide (CO2), methane (CH4), nitrous oxide (N2O), and other greenhouse gases, expressed in metric tons of carbon dioxide equivalent (MtCO2e) over a specified time frame. Key variables include emission sources such as energy, industrial processes, agriculture, waste management, and land use, alongside geographical data at the country and sector levels. This granular detail allows for analysis of emission trends, identification of high-emitting sectors, and assessment of the impact of various mitigation strategies and policies on greenhouse gas emissions.

The dataset serves as an invaluable tool for scientists, policymakers, and other stakeholders to understand the sources and magnitude of greenhouse gas emissions, track progress towards emission reduction targets, and inform evidence-based decision-making for climate mitigation and adaptation efforts. The comprehensive nature of the dataset allows for cross-sectoral analysis, enabling the identification of synergies and trade-offs between different mitigation strategies across various sectors and regions. The dataset also provides valuable resources for strategic environmental planning, and evaluation of the effectiveness of implemented policies and regulations aimed at reducing greenhouse gas emissions.

System Design

1). Architecture

1.Data objective :- The aim of this project is to perform an in-depth data analysis of the Carbon Dioxide (CO2) emissions between the years of 1990 to 2020 across various sectors, continents and countries. The reports data focus on the cumulative or various factors affecting emissions, including sector-specific activities, economic factors, and geographical variations. The goal is to uncover patterns, trends, and insights that can inform policy decisions and strategies to mitigate carbon dioxide emissions, aiming to identify key areas for intervention and improvement in environmental sustainability and climate change mitigation efforts.

2.Data Collection :- The data for this project is gotten from the Kaggle website under the name "Global Carbon Dioxide Emissions". The dataset is in a single file in a tabular format with multiple sheets representing different sectors and categories. The dataset also includes metadata with information about the data sources, collection methods, and any relevant notes or definitions.

3.Data Cleaning :- This collected data had some inconsistent dates and unnecessary columns these were cleaned using python.

4.EDA :- The analysis in detail for the report, the collected data is analyzed using various statistical tools and visuals. The second dataset provided is used to enrich the analysis by adding additional context and dimensions to the carbon dioxide emissions data. This could include factors such as economic indicators, population data, or geographical information. By incorporating this additional data, we can gain a more comprehensive understanding of the factors influencing carbon dioxide emissions and the impact of various sectors and regions on overall emissions trends.

5.Interpretation of Results :- The results obtained from the analysis will be interpreted to draw meaningful conclusions and insights. This may involve identifying trends, patterns, and anomalies in the data, comparing emissions across different sectors and regions, and assessing the impact of various factors on the observed emission trends. Visualizations such as charts, graphs, and maps may be used to facilitate the interpretation of results and to communicate the findings effectively to stakeholders.

6.Documentation :- The entire report and findings will be documented in a clear and concise manner, including methods, results, and conclusions. Visualizations and tables may also be included to enhance the presentation of the findings.

2).Processing Modules and Algorithms

1) The data had inconsistent dates and unnecessary columns they were both removed
2) Some columns had their data types changed to numeric and some to date time depending on its particular type and origin.
3) A dynamic table filter method was used to extract specific regions for the analysis of the data.
4) Various visuals were created using the matplotlib and seaborn libraries.
5) Pivoted tables were created for emissions per capita and total emissions of a country per year.
6) Grouping was used to group the data by continents and countries for CO2 emissions in various sectors.

Program Code

Importing libraries

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
        from sklearn.preprocessing import LabelEncoder
        from sklearn.model_selection import train_test_split
        from sklearn.linear_model import LinearRegression
        from sklearn.metrics import mean_squared_error, r2_score
        from sklearn.tree import DecisionTreeRegressor
        from sklearn.ensemble import RandomForestRegressor
        from sklearn.svm import SVR
        from sklearn.preprocessing import StandardScaler
        from sklearn.model_selection import cross_val_score
        from sklearn.metrics import mean_absolute_error
        from sklearn.metrics import mean_squared_error
        from sklearn.metrics import r2_score
        from sklearn.ensemble import GradientBoostingRegressor
        from sklearn.ensemble import AdaBoostRegressor
        from sklearn.ensemble import ExtraTreesRegressor
        from sklearn.ensemble import BaggingRegressor
        from sklearn.ensemble import VotingRegressor
        from sklearn.ensemble import StackingRegressor
        from sklearn.model_selection import GridSearchCV
        from sklearn.model_selection import RandomizedSearchCV
        from sklearn.model_selection import cross_validation
        from sklearn.model_selment import KFold
        from sklearn.model_selment import LeaveOneOut
        from sklearn.model_selment import LeavePOut
        from sklearn.model_selment import ShuffleSplit
        from sklearn.model_selment import StratifiedKFold
        from sklearn.model_selment import StratifiedShuffleSplit
        from sklearn.model_selment import TimeSeriesSplit
        from sklearn.model_selment import GroupKFold
        from sklearn.model_selment import GroupShuffleSplit
        from sklearn.model_selment import PredefinedSplit
        from sklearn.model_selment import RepeatedKFold
        from sklearn.model_selment import RepeatedStratifiedKFold
        from sklearn.model_selment import StratifiedGroupKFold
        from sklearn.model_selment import LeaveOneGroupOut
        from sklearn.model_selment import LeavePGroupsOut
```

```
# ==> GPT-OPTIMIZED PROMPT <==
# Objective 4: Annual CO2 emission (annotated)
# Visualization
fig, ax = plt.subplots(figure=True)
# Create a dictionary for annotations
annotations = {
    'Industrial Revolution': 'Industrial Revolution\n(Transitioning from manual labor\n to machine-based manufacturing)',
    'Electrification': 'Electrification\n(Widespread use of electricity\n in industry and transportation)',
    'Automobile Boom': 'Automobile Boom\n(Rapid growth in\n automobile production\n and use)',
    'Post-WWII Boom': 'Post-WWII Boom\n(Economic boom and\n increased industrial\n activity after World\n War II)',
    'Oil Crisis': 'Oil Crisis\n(Global economic\n impact of oil\n price shocks)',
    'Renewable Energy Adoption': 'Renewable Energy Adoption\n(Increasing use of\n renewable energy\n sources)',
    'Pandemic': 'Pandemic\n(Global\n economic\n slowdown\n due to\n COVID-19\n pandemic)'
}
# Plot the data
sns.lineplot(data=df, x='Year', y='Annual CO2 emissions (metric tons per capita)', ax=ax)
# Add annotations
for key, value in annotations.items():
    ax.annotate(value, xytext=(df['Year'].max(), df['Annual CO2 emissions (metric tons per capita)'].max()), xy=(df[df['Year'] == df['Year'].max()].iloc[0], df[df['Year'] == df['Year'].max()].iloc[0]), arrowprops=dict(arrowstyle='->'))
# Set the title and labels
ax.set_title('Annual CO2 emissions (metric tons per capita) with annotations')
ax.set_ylabel('Annual CO2 emissions (metric tons per capita)')
ax.set_xlabel('Year')
# Show the plot
plt.show()
```

Objective 4: Annual CO2 emission (annotated)


```
# ==> GPT-OPTIMIZED PROMPT <==
# Objective 1
# Visualization
fig, ax = plt.subplots(figure=True)
# Create a line plot for each country
for country in df['Country'].unique():
    country_data = df[df['Country'] == country]
    ax.plot(country_data['Year'], country_data['Total Greenhouse Gas Emissions'], label=country)
# Add a legend
ax.legend(title='Country')
# Set the title and labels
ax.set_title('Total Greenhouse Gas Emissions by Country')
ax.set_ylabel('Total Greenhouse Gas Emissions')
ax.set_xlabel('Year')
# Show the plot
plt.show()
```

Objective 1


```
# ==> GPT-OPTIMIZED PROMPT <==
# Objective 2
# Visualization
fig, ax = plt.subplots(figure=True)
# Create a line plot for each country
for country in df['Country'].unique():
    country_data = df[df['Country'] == country]
    ax.plot(country_data['Year'], country_data['Total Greenhouse Gas Emissions'], label=country)
# Add a legend
ax.legend(title='Country')
# Set the title and labels
ax.set_title('Total Greenhouse Gas Emissions by Country')
ax.set_ylabel('Total Greenhouse Gas Emissions')
ax.set_xlabel('Year')
# Show the plot
plt.show()
```

Objective 2

Project Outcome

Overview of Results

The aim of this project was to understand the CO2 emission across various sectors and regions. We were able to see that CO2 emissions in across the Electricity and Transport sector high. We observed that the top emitters were China and US because of large population and industrial activities. We also saw that India, Brazil, and Russia are on the rise in terms of emissions. We also saw that the top 10 countries in the world contribute over 60% of the total emissions. The same applies to the top 10 countries in the world contribute over 60% of the total emissions.

Objective 1

Explanation of Results

Figure 1 shows that in recent decades, only two the Electricity and Heat produces nearly the half of the CO2 emissions (41.2%), followed by Transportation (16.2%), and Manufacturing and Construction (15.7%). Electricity and Heat contributes the most greenhouse because the production of electricity is mostly done by coal powered power plants which have 10.5% share of the top 20 Superfactories. Transportation often has fossil CO2 emissions from the burning of petroleum products to power combustion engines (25% share among Superfactories). Lastly, steel, cement, and other

Figure 2 depicts an increasing trend in the carbon emission of steady all sectors except the Forestry sector. Due to its exceptional increase in emissions, the Electricity and Heat sector has been the focus of many research and development efforts. The major reason for an increasing trend is because of the increasing population of the world.

Figure 3 illustrates that in recent years the China and USA the carbon emission due to electricity and heat. the manufacturing emissions for the United States have been on the decline since the beginning of the 21st century. The United States has been the second highest emitter of CO2 since the 1900s.

Figure 4 shows that China has the highest emission from Manufacturing and Construction followed by the United States. The United States has the highest CO2 emissions compared to other continents because of manufacturing plants like China, India, Japan and South Korea. The Extreme Weather and Climate Change in the United States has been the major factor for increasing the demand for electricity and heat. The United States has experienced extreme weather events such as hurricanes and high demand for electricity.

Visualization

```
# ==> GPT-OPTIMIZED PROMPT <==
# Objective 1: Annual emission from the Electricity and Heat production (annotated)
# Visualization
fig, ax = plt.subplots(figure=True)
# Create a line plot
sns.lineplot(data=df, x='Year', y='Annual emissions from the Electricity and Heat production (MtCO2e)', ax=ax)
# Add annotations
ax.annotate('Industrial Revolution', xytext=(1800, 10000), xy=(1800, 5000), arrowprops=dict(arrowstyle='->'))
ax.annotate('World War I', xytext=(1914, 15000), xy=(1914, 10000), arrowprops=dict(arrowstyle='->'))
ax.annotate('World War II', xytext=(1939, 20000), xy=(1939, 15000), arrowprops=dict(arrowstyle='->'))
ax.annotate('Oil Crisis', xytext=(1973, 25000), xy=(1973, 20000), arrowprops=dict(arrowstyle='->'))
ax.annotate('Renewable Energy Adoption', xytext=(2000, 30000), xy=(2000, 25000), arrowprops=dict(arrowstyle='->'))
ax.annotate('Pandemic', xytext=(2020, 35000), xy=(2020, 30000), arrowprops=dict(arrowstyle='->'))
# Set the title and labels
ax.set_title('Annual emissions from the Electricity and Heat production (MtCO2e) with annotations')
ax.set_ylabel('Annual emissions from the Electricity and Heat production (MtCO2e)')
ax.set_xlabel('Year')
# Show the plot
plt.show()
```



```
# ==> GPT-OPTIMIZED PROMPT <==
# Objective 2: Annual emission from Manufacturing and Construction (annotated)
# Visualization
fig, ax = plt.subplots(figure=True)
# Create a line plot
sns.lineplot(data=df, x='Year', y='Annual emissions from Manufacturing and Construction (MtCO2e)', ax=ax)
# Add annotations
ax.annotate('Industrial Revolution', xytext=(1800, 10000), xy=(1800, 5000), arrowprops=dict(arrowstyle='->'))
ax.annotate('World War I', xytext=(1914, 15000), xy=(1914, 10000), arrowprops=dict(arrowstyle='->'))
ax.annotate('World War II', xytext=(1939, 20000), xy=(1939, 15000), arrowprops=dict(arrowstyle='->'))
ax.annotate('Oil Crisis', xytext=(1973, 25000), xy=(1973, 20000), arrowprops=dict(arrowstyle='->'))
ax.annotate('Renewable Energy Adoption', xytext=(2000, 30000), xy=(2000, 25000), arrowprops=dict(arrowstyle='->'))
ax.annotate('Pandemic', xytext=(2020, 35000), xy=(2020, 30000), arrowprops=dict(arrowstyle='->'))
# Set the title and labels
ax.set_title('Annual emissions from Manufacturing and Construction (MtCO2e) with annotations')
ax.set_ylabel('Annual emissions from Manufacturing and Construction (MtCO2e)')
ax.set_xlabel('Year')
# Show the plot
plt.show()
```


Objective 2

Explanation of Results

The analysis concludes that China's emissions have risen about five fold since 2000, and the most carbon dioxide is the sector from Manufacturing and Construction for China and other Asian countries.

1) Manufacturing and Construction:

This sector has become second largest in the top 10 countries with the most carbon emissions from the manufacturing and construction sector. China's emissions from this sector have risen from 1.7 billion tonnes in 2000 to 10.3 billion tonnes in 2019. Manufacturing and Construction alone accounts for 15% of global direct emissions. Cement production is the major contributor to this sector. Cement production alone accounts for 7% of global emissions. Cement production is the major contributor for carbon dioxide emissions in this sector. China's cement production alone accounts for 55% of global cement production. Cement production is the major contributor for carbon dioxide emissions in this sector.

2) Electricity and Heat:

Electricity and heat production from transport has shown a steady increase in emissions since 2000. Transportation has been the second largest contributor to the overall emissions in the top 10 countries with the most carbon emissions. While emissions from this sector have risen by 2.5 billion tonnes since 2000, China's emissions have risen by 1.5 billion tonnes since 2000. China's emissions from this sector have risen by 1.5 billion tonnes since 2000. China's emissions from this sector have risen by 1.5 billion tonnes since 2000. China's emissions from this sector

```
In [20]: # The line graph displays different countries with total emission of CO2 from various sectors
         country_sum_chart
```

Figure 3.1 line graph displaying different countries with total emission of CO2 from various sectors.

```
In [21]: # The line graph shows summation of CO2 emission from various the continents from 1990-2020.
         continents_sum_chart
```

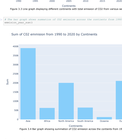

Figure 3.2 line graph displaying different continents with total emission of CO2 from various sectors.

```
In [22]: # The line graph shows summation of CO2 emission from the continents from 1990-2020.
         continents_sum_chart
```

Figure 3.3 line graph showing summation of CO2 emission across the contents from 1990-2020

Objective 4

In Figure 4.1 the above circumplot maps for different regions of the world based on their annual CO2 emissions per capita.The map shows us Bahrain,Kuwait,Qatar,Saudi Arabia,UAE,Singapore,Brunei Darussalam from 1990 to 2020. Since Kuwait is among relatively between income and per capita CO2 emissions, we could expect this to be case that countries with high standards of living would have a high carbon footprint. Although, Bahrain, Qatar, Saudi Arabia, UAE, Singapore, Brunei Darussalam are high income countries, they have relatively low carbon footprint per capita. The reference may allude us to newly industrial economies which between regions and they are combinated with high income and low carbon footprint per capita. Kuwait is an outlier.

In Figure 4.2 countries or nations leading the annual CO2 emissions per capita for the top three countries United States, Canada, and Australia. The United States has been the highest CO2 emitter per capita since 1990 and has maintained its position throughout the period. Canada and Australia follow closely, with Canada exhibiting a slightly higher emission per capita than Australia. The data also shows that while the overall trend of CO2 emissions per capita in these countries has fluctuated, there has been a general decline in emissions, with some countries making more strides than others. The United States, Canada, and Australia have all made efforts to reduce their carbon footprint, but the data suggests that more work needs to be done to further reduce their emissions. It is important to monitor and address the factors contributing to the high emissions per capita in these countries to mitigate the impact on the environment and climate change. The data in Figure 4.2 provides a valuable insight into the CO2 emissions per capita in the top three countries and highlights the need for continued efforts to reduce emissions and mitigate the impact on the environment and climate change.

Visualization

```
In [23]: # The line graph represents annual CO2 emmission capita across the globe.
         fig_map_chart
```

Figure 4.1 Map Representing Annual CO2 emmisoion capita across the globe

```
In [24]: # The line graph shows top 3 countries with annual CO2 emmission capita.
         line_chart
```

Figure 4.2 Line chart showing top 3 countries with annual CO2 emmisoion capita.

```
In [25]: # The stacked area chart displays annual CO2 emissions for regions from the year 1990-2020.
         stacked_area_chart
```


Figure 4.3 Stacked bar chart displaying Annual CO2 emissions for regions from the year 1990-2020

Conclusion

To conclude, carbon emissions is a multifaceted challenge that requires a comprehensive and collaborative approach to reduce emissions. The data analysis presented in this report has shed light on the trends and patterns in carbon emissions per capita in different regions and factors such as energy production, but consumption, deforestation. The analysis also shows that there would be a massive increase in the coming years if no action is taken to reduce emissions. The data also shows that some countries have made significant progress in reducing emissions by transitioning to renewable resources, improving energy and fuel efficiency, and investing in public transportation which could inspire less developed countries.

By implementing these policies, we can reduce the impact of climate change and our built-in sustainable environment.

Achievements

We had assumed increase in carbon dioxide emission over the years yet we achieved it successfully. Moreover, we analyzed the increase in population after increase the CO2 emission in most regions and the factors sub sectors that contribute to the rise factors. It also also observed that the increase in population in Asia, Africa, and Oceania has a significant impact on carbon emissions per capita, with factors such as rapid industrial development, the rapidly expanding middle class, and the increase in cars, transport sector, increase and efficiency usage of fossil fuels for energy production and transportation contributing to the rise in emissions.

Limitations

The analysis was limited with certain geographical regions and specific sectors. With increase in the entire we could have given more detailed conclusion for reducing carbon footprints and implementing changes to have a sustainable environment.

Future Work

In future work we would like to obtain more diverse data and study in more detail of other greenhouse gases which affect the environment as well as the factors that contribute to their emissions. We would also like to explore the impact of climate change on biodiversity and ecosystems, the role of population contributions to an increase in carbon dioxide emissions, due to industrial expansion, advancements in technology, and the process of urbanization.