VISUALIZATION OF WORLD GDP AND CARBON - DIOXIDE EMISSION

Objective

The objective of this project is to analyze and visualize the relationship between GDP per capita and CO2 emissions per capita using the World Development Indicators dataset from the World Bank. The goals include:

1. **Exploring** the dataset to understand its structure and content.
2. **Visualizing** CO2 emissions and GDP data to identify trends and patterns.
3. **Comparing** CO2 emissions and GDP per capita for the United States and other countries.
4. **Assessing** the relationship between GDP and CO2 emissions through scatter plots and correlation analysis.

Summary

The World Development Indicators dataset contains annual economic development indicators from countries worldwide. Key indicators include CO2 emissions per capita and GDP per capita. The dataset allows for detailed exploration and visualization of economic and environmental data.

Results

1. **Dataset Exploration**:
   - **Shape and Size**: The dataset is large, with thousands of rows representing different indicators, countries, and years.
   - **Unique Values**: There are numerous unique country names, country codes, indicators, and years in the dataset.
2. **CO2 Emissions Analysis**:
   - **For the USA**:
     - A line plot of CO2 emissions per capita over time shows a general decrease.
     - A histogram of CO2 emissions values reveals a high concentration around 19-20 metric tons per capita, with some outliers.
   - **For All Countries (2011)**:
     - A histogram of CO2 emissions per capita in 2011 shows that the USA has relatively high emissions compared to other countries.
3. **GDP Analysis**:
   - **For the USA**:
     - A line plot of GDP per capita shows growth over time without a corresponding decrease aligned with CO2 emissions.

   - **Comparison with CO2 Emissions**:

        - A scatter plot comparing GDP per capita and CO_2 emissions per capita shows a weak relationship.
        - The correlation coefficient between GDP and CO_2 emissions is approximately 0.07, indicating a very weak correlation.

Conclusion

The analysis and visualization of CO_2 emissions and GDP per capita reveal the following:

- **Trends**: CO_2 emissions per capita in the USA have decreased over time, while GDP per capita has generally increased.
- **Comparison**: The USA's CO_2 emissions are relatively high compared to other countries.
- **Relationship**: There is a weak correlation between GDP per capita and CO_2 emissions, suggesting that economic growth does not strongly correlate with changes in emissions.

Code :

```
'''
World Development Indicators :
The World Development Indicators dataset obtained from the World Bank containing over a
thousand annual indicators of economic development from hundreds of countries around the world.
'''


# Initial exploration of the Dataset
import pandas as pd
import numpy as np
import random
import matplotlib.pyplot as plt
import matplotlib.cbook
import zipfile
import bz2
import warnings


warnings.filterwarnings("ignore", category=matplotlib.MatplotlibDeprecationWarning)
#Let us read the dataset
data = pd.read_csv("D:\\Aditya's Notes\\Aditya's Data Science Notes\\Projects and Other
Datasets\\ML PROJECTS\\data\\Indicators.bz2")
#data = pd.read_csv('data/Indicators.bz2', compression = 'bz2')
print("data.shape: ", data.shape)
#This is a really large dataset, at least in terms of the number of rows.
print("Sample Data: \n",data.head())
print("Columns: \n",data.columns)


#From the above dataset, it looks like it has different indicators for different
# countries with the year and value of the indicator.


#How many UNIQUE country names are there ?
countries = data['CountryName'].unique().tolist()
```

```
print("Number of countries: ",len(countries))
#Are there same number of country codes ?
#How many unique country codes are there?
#It should be the same as number of unique countries.
countryCodes = data['CountryCode'].unique().tolist()
print("Number of country codes: ",len(countryCodes))
#Are there many indicators or few ?
#How many unique indicators are there?
indicators = data['IndicatorName'].unique().tolist()
print("Number of indicators: ",len(indicators))
#How many years of data do we have ?
years = data['Year'].unique().tolist()
print("Number of years: ",len(years))
#What's the range of years?
print(min(years)," to ",max(years))

#################
## Data Visualization
#################
'''
#Let us pick a country and an indicator to explore CO2 Emissions per capita
and the USA.
#To select CO2 emissions for the United States, We will take the intersection
# of two masks, one with all the rows that contains the string,
# "C02 emissions" and the other which contains all the rows containing the
string, "USA".
'''

hist_indicator = 'CO2 emissions \\(metric'
hist_country = 'USA'
mask1 = data['IndicatorName'].str.contains(hist_indicator)
```

```
mask2 = data['CountryCode'].str.contains(hist_country)

stage = data[mask1 & mask2]

# stage dataset contain indicators matching the USA for country code & CO2emissions over time.

print (stage.shape)

stage.head()

print("Indicator Name: ", stage['IndicatorName'].iloc[0])


#Let us see how emissions have changed over time using MatplotLib

years = stage['Year'].values # get the years

co2 = stage['Value'].values # get the values

# Plot the Histogram

plt.bar(years,co2)

plt.show()



#It is seen that emissions per capita have dropped a bit over time,

# but let us make this graph a bit more appealing before we continue to explore it.

#Let us create a line plot.

plt.plot(stage['Year'].values, stage['Value'].values)

# Label the axes

plt.xlabel('Year')

plt.ylabel(stage['IndicatorName'].iloc[0])

# Label the figure

plt.title('CO2 Emissions in USA')

# Start the y axis at 0 and x axis from 1959

plt.axis([1959, 2011,0,25])

plt.show()


#Using Histograms to explore the distribution of values

#We could also visualize this data as a histogram to better explore

# the ranges of values in CO2 production per year.
```

```
# If we want to just include those within one standard deviation fo the mean, you could do the following

# lower = stage['Value'].mean() - stage['Value'].std()

# upper = stage['Value'].mean() + stage['Value'].std()

# hist_data = [x for x in stage[:10000]['Value'] if x>lower and x<upper ]

# Otherwise, let's look at all the data

hist_data = stage['Value'].values

print(hist_data)

print(len(hist_data))

# Histogram of the data

plt.hist(hist_data, 10, density=False, facecolor='green') # 10 is the number of bins

plt.xlabel(stage['IndicatorName'].iloc[0])

plt.ylabel('# of Years')

plt.title('Histogram Example')

plt.grid(True)

plt.show()



#USA has many years where it produced between 19-20 metric tons per capita

#with outliers on either side.

#But how do the USA's numbers relate to those of other countries?

# select CO2 emissions for all countries in 2011

hist_indicator = 'CO2 emissions \\(metric'

hist_year = 2011

mask1 = data['IndicatorName'].str.contains(hist_indicator)

mask2 = data['Year'].isin([hist_year])


# apply our mask

co2_2011 = data[mask1 & mask2]

co2_2011.head()

#For how many countries do we have CO2 per capita emissions data in 2011
```

```
print(len(co2_2011))



# Let us plot a histogram of the emmissions per capita by country

# subplots returns a touple with the figure, axis attributes.

fig, ax = plt.subplots()

ax.annotate("USA",xy=(18, 5), xycoords='data',xytext=(18, 30),

textcoords='data',

 arrowprops=dict(arrowstyle="->",connectionstyle="arc3"))

plt.hist(co2_2011['Value'], 10, density=False, facecolor='green')

plt.xlabel(stage['IndicatorName'].iloc[0])

plt.ylabel('# of Countries')

plt.title('Histogram of CO2 Emissions Per Capita')

plt.grid(True)

plt.show()



#USA, at ~18 CO2 emissions (metric tons per capital) is quite high among all countries.

#3. Matplotlib: Basic Plotting Part 2

#Relationship between GDP and CO2 Emissions in USA

# Select GDP Per capita emissions for the United States

hist_indicator = 'GDP per capita \\(constant 2005'

hist_country = 'USA'

mask1 = data['IndicatorName'].str.contains(hist_indicator)

mask2 = data['CountryCode'].str.contains(hist_country)

# Stage is just those indicators matching the USA for country code and CO2 emissions over time.

gdp_stage = data[mask1 & mask2]

# Plot gdp_stage vs stage

print("GDP: ",gdp_stage.head())

stage.head()

# Switch to a line plot
```

```
plt.plot(gdp_stage['Year'].values, gdp_stage['Value'].values)
# Label the axes
plt.xlabel('Year')
plt.ylabel(gdp_stage['IndicatorName'].iloc[0])
#Label the figure
plt.title('GDP Per Capita USA')
plt.show()

#Although we have seen a decline in the CO2 emissions per capita,
# it does not seem to translate to a decline in GDP per capita
#ScatterPlot for comparing GDP against CO2 emissions (per capita)
#First, we will need to make sure we are looking at the same time frames.
print("GDP Min Year = ", gdp_stage['Year'].min(), "max: ",
gdp_stage['Year'].max())
print("CO2 Min Year = ", stage['Year'].min(), "max: ", stage['Year'].max())

#We have 3 extra years of GDP data, so let's trim those off so the scatterplot
# has equal length arrays to compare (this is actually required by scatterplot)
gdp_stage_trunc = gdp_stage[gdp_stage['Year'] < 2012]
print(len(gdp_stage_trunc))
print(len(stage))

import matplotlib.pyplot as plt
fig, axis = plt.subplots()
# Grid lines, Xticks, Xlabel, Ylabel
axis.yaxis.grid(True)
axis.set_title('CO2 Emissions vs. GDP (per capita)',fontsize=10)
axis.set_xlabel(gdp_stage_trunc['IndicatorName'].iloc[0],fontsize=10)
axis.set_ylabel(stage['IndicatorName'].iloc[0],fontsize=10)
X = gdp_stage_trunc['Value']
Y = stage['Value']
```

```
axis.scatter(X, Y)

plt.show()

#This does not look like a strong relationship. We can test this by looking at correlation.

print(np.corrcoef(gdp_stage_trunc['Value'],stage['Value']))

#A correlation of 0.07 is very weak.
```