SDG_Analysis

July 8, 2025

[1]:
```
import csv
import pandas as pd
import seaborn as sns
import numpy as np
import os
import plotly.express as px
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from matplotlib.pyplot import figure
import matplotlib.pyplot as plt
```

1 Executive Summary

I analysed a dataset on sustainable development in all countries between the 1990s and 2020. The data analysis suggests that urban populations have risen consistently in the 1990s, however, when dividing the average between states, one finds significant deviations in off growth of urban populations between countries, which may have significantly skewed the total average. Moreover, we found a positive correlation between the number of flights and GDP growth, suggesting that an increase in flights is merely an externality of economic growth. The same thing cannot be necessarily stated about the relationship between the percentage of 'green' areas in a country and economic growth, as one can find that the largest economic growing economies have had positive forestation efforts, while other growing economies had negative forestation efforts.

More broadly, I checked the composition of the data itself, I cheched how much data was collected throughout the years and I tried to explain why we had positive or negative trends. Further, I checked which categories have the most collected data and which income group those those countries belong to.

2 Introduction

The Sustainable development goals dataset contains various data, collected over thirty years, about countries from all over the world. In an age where climate change presents a severe problem to our society, sustainable development is the answer. I chose to analyze this dataset because I think it is the most relevant topic and I am interested in the statistics behind our collective progress or failure. I am expecting to see some patterns behind the growth of energy consumption, gas emission, and other relevant data like traffic or GPP.

3 Data Cleaning

Firstly, I am going to check how many empty values I have in each column.

[2]:

```
# Load data
data = pd.read_csv("SDGData.csv")

# Count missing values
na_counts = data.isna().sum()

# Drop columns with 0 missing values
na_counts = na_counts[na_counts > 0]

# Transpose for horizontal view
na_table_wide = pd.DataFrame(na_counts).T
na_table_wide.index = ['Missing Values']

# Display in wide format
na_table_wide
```

[2]:

```
                 1990   1991   1992   1993   1994   1995   1996   1997   1998  \
Missing Values  82491  77866  76175  77041  76595  74313  74866  73487  74118

                 1999  ...   2012   2013   2014   2015   2016   2017   2018  \
Missing Values  71668  ...  53390  55608  52788  53103  53346  51517  54078

                 2019   2020  Unnamed: 35
Missing Values  60308  71630       106488

[1 rows x 32 columns]
```

We are cleaning the code by deleting the rows with empty values.

[3]:

```
# First check if the column exists before trying to drop it
if "Unnamed: 35" in data.columns:
    delete = ["Unnamed: 35"]
    data.drop(delete, inplace=True, axis=1)  # Drop the column if it exists

deleteNull = data.dropna()  # Create new dataframe by dropping rows with missing␣
 ↪values

# Count missing values
na_counts = deleteNull.isna().sum()

# Transpose for horizontal view
na_table_wide = pd.DataFrame(na_counts).T
na_table_wide.index = ['Missing Values']
```

```
# Display in wide format
na_table_wide
```

```
[3]:                 Country Name  Country Code  Indicator Name  Indicator Code  \
     Missing Values             0             0               0               0

                     1990  1991  1992  1993  1994  1995  ...  2011  2012  2013  \
     Missing Values     0     0     0     0     0     0  ...     0     0     0

                     2014  2015  2016  2017  2018  2019  2020
     Missing Values     0     0     0     0     0     0     0

     [1 rows x 35 columns]
```

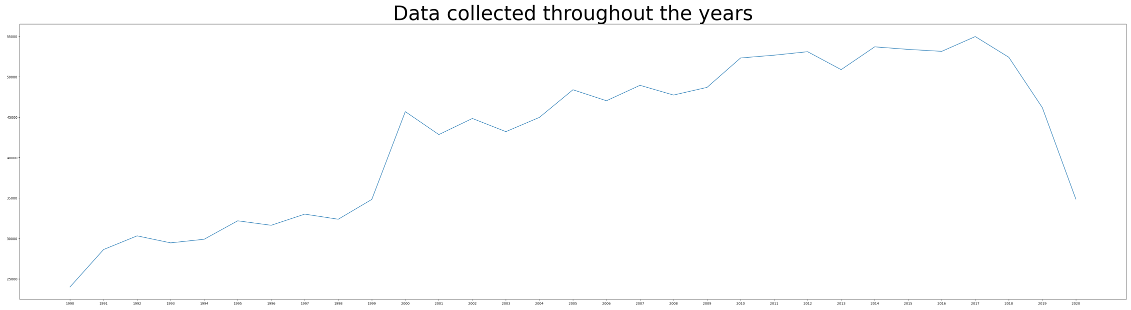
We deleted all the missing values and Unnamed: 35.

4 Exploratory Data Analysis

Let’s try to show how much data was collected throughout the years

```
[4]: plt.rcParams["figure.figsize"] = (60,15) #to spread out the years

     plt.plot(data.count()[4:]) #we removed first 4 columns because those are our␣
      ↪categories and we are interested in the data
     plt.title('Data collected throughout the years', fontsize = 60) #we put label
```

```
[4]: Text(0.5, 1.0, 'Data collected throughout the years')
```



We can see that we were more and more diligent when collecting the data as years went exception being in the last couple of years. I think this was due to Covid-19 Virus.

Now I am going to check which categories have the most data. I did that by checking how many countries have the data in each category.

```
[5]: table = deleteNull.groupby(["Indicator Name"]).count().sort_values("Country␣
      ↪Code") #we are grouping categories with country codes
```

```
df_new = table.iloc[:,0:1] #we want to show only "Indicator Name" and "Country"␣
 ↪columns
display(df_new) #we displayed table
```

```
                                                    Country Name
Indicator Name
Compulsory education, duration (years)                         1
Preprimary education, duration (years)                         1
Poverty headcount ratio at $1.90 a day (2011 PP...             1
Proportion of people living below 50 percent of...             1
Share of youth not in education, employment or ...             1
...                                                          ...
Primary education, duration (years)                          248
Forest area (sq. km)                                         249
Urban population growth (annual %)                           257
Urban population                                             258
Urban population (% of total population)                     259

[147 rows x 1 columns]
```

5 Descriptive Analytics

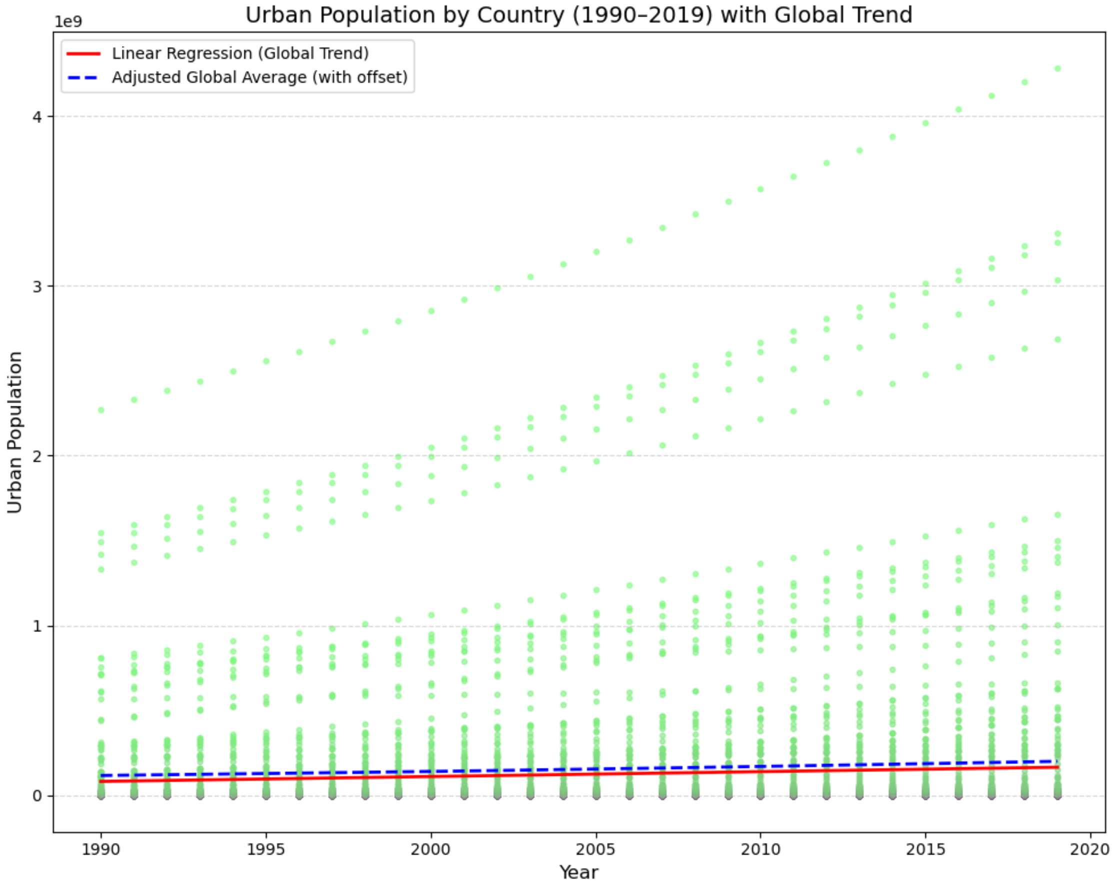
6 What was urban population growth from 1990 to 2020?

The first question asked was as follows; "what was urban population growth from 1990 to 2020?" this is a relevant point because, as an increase in urban populations suggests a higher concentration of individuals within high industry areas and economic zones of the state.

[6]:

```
# Filter for urban population indicator
pop = deleteNull[deleteNull["Indicator Code"] == "SP.URB.TOTL"]

# Prepare data for regression
X = np.array(sum([[year for j in range(len(pop))] for year in range(1990,␣
 ↪2020)], [])).reshape((-1, 1))
y = np.array(sum([list(pop[str(year)]) for year in range(1990, 2020)], []))

# Fit linear regression model
reg = LinearRegression().fit(X, y)
y_pred = reg.predict(np.array(range(1990, 2020)).reshape(-1, 1))

# Plot regression and scatter
plt.figure(figsize=(10, 8), dpi=100)
plt.plot(range(1990, 2020), y_pred, color="red", linewidth=2, label="Linear␣
 ↪Regression (Global Trend)")

# Scatter points for each country
ytemp = [sorted(list(pop[str(year)])) for year in range(1990, 2020)]
```

```
for i in range(len(pop)):
    plt.scatter(range(1990, 2020),
                [ytemp[year][i] for year in range(2020 - 1990)],
                color=(0.5, i / len(pop), 0.5),
                alpha=0.6,
                s=10)

# Plot average with artificial offset
plt.plot(range(1990, 2020),
         [sum(pop[str(year)]) / len(pop) + 10**7.5 for year in range(1990,␣
 ↪2020)],
         color="blue", linestyle="--", linewidth=2,
         label="Adjusted Global Average (with offset)")

# Add labels and title
plt.xlabel("Year", fontsize=12)
plt.ylabel("Urban Population", fontsize=12)
plt.title("Urban Population by Country (1990-2019) with Global Trend",␣
 ↪fontsize=14)
plt.legend()
plt.grid(axis='y', linestyle='--', alpha=0.5)
plt.tight_layout()
plt.show()

# Print predictions
print("Predicted number for average people in 2050 and 2100, respectively:",
      reg.predict(np.array([2050, 2100]).reshape(-1, 1)))
```



```
Predicted number for average people in 2050 and 2100, respectively:
[2.54829257e+08 3.98826961e+08]
```

In reference to the first graph, one can see that urban population growth is linear in the last 30 years, with a steady growth of 666 million individuals per year. However, it must be recognized that different states have different rates of urban population growth which is highlighted in the second graph. It underscored that different states have significantly different urban population growth rates, where the world averages can be skewed by some states such as China and India. The red line represents a model which it attempts to predict the aggregate urban growth in the future, between all states using data from the 1990s to 2020.

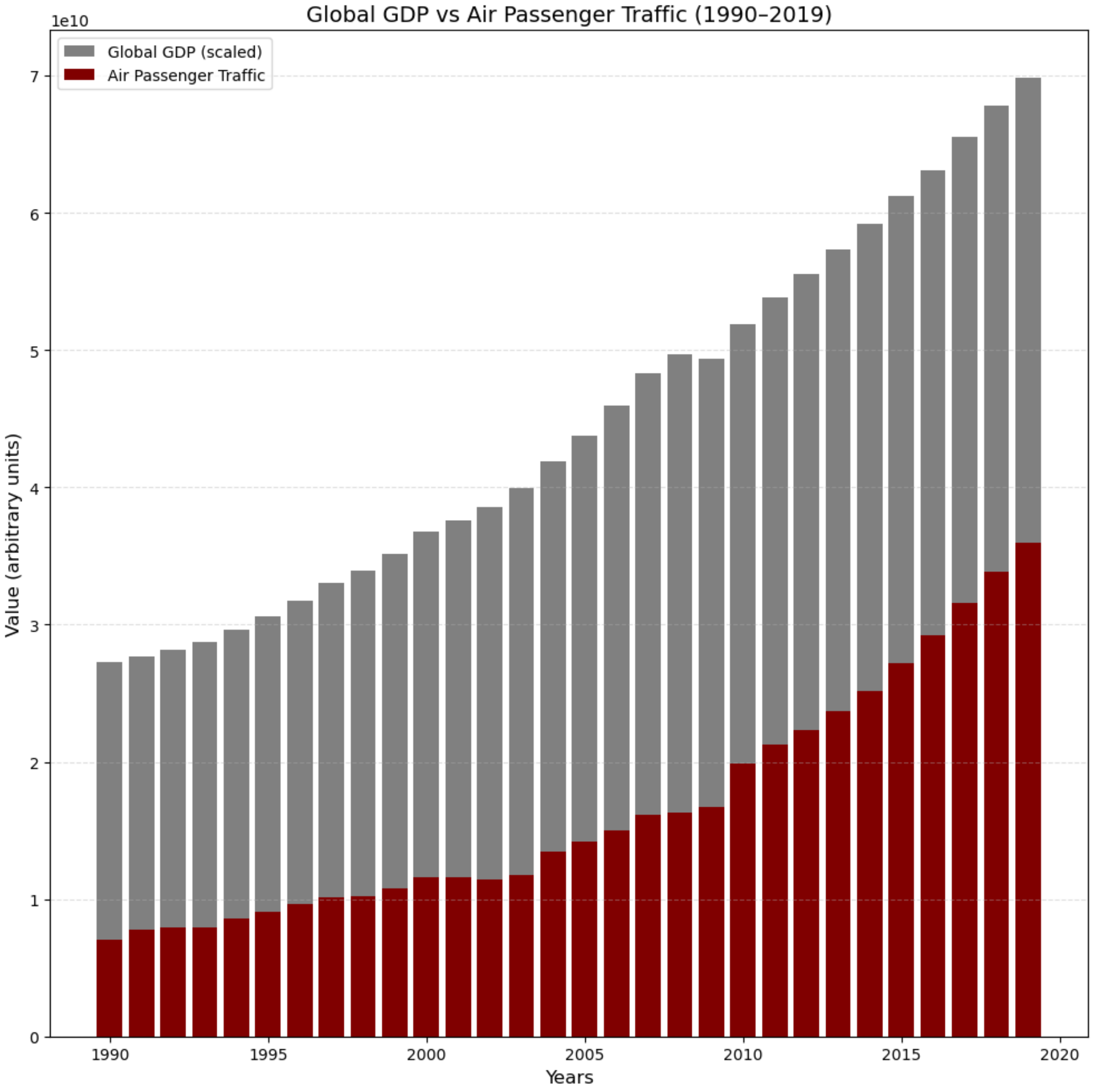
By using a linear regression model, and assuming axioms in relation to other data points, the model was so accurate, that the blue line which represents the actual average of urban growth of states is overlapping with our prediction line (the red line). For visibility's sake, the blue line has to be scaled up by (10**7.5). From the linear regression model, we can thus approximate the world average urban population growth in 2050 and in 2100, and that is (2.54829257e+08 3.98826961e+08) respectively.

7 How is the growth of GDP correlated to the number of flights?

The second question asked was as follows; “how is the growth of GDP correlated to the number of flights?” this is an incredibly significant data point because while economic growth is paramount and is something that must always be thrived for, it can however have negative externalities, mostly manifesting itself with regressive changes to the climate. Airplanes are one of the most pollutant ways of transportation.

[7]:

```python
# Filter data
flight = deleteNull[deleteNull["Indicator Code"] == "IS.AIR.PSGR"]
gdp = deleteNull[deleteNull["Indicator Code"] == "NY.GDP.MKTP.KD"]

# Prepare years
years = list(range(1990, 2020))

# Create figure
plt.figure(figsize=(10, 10))  # same size

# Plot GDP (arbitrarily scaled down for visual comparison)
plt.bar(
    years,
    [sum(gdp[str(year)]) / (10**4) for year in years],
    color=(0.5, 0.5, 0.5),
    label="Global GDP (scaled)"
)

# Plot Air Passengers (raw sum, assumed in millions or billions)
plt.bar(
    years,
    [sum(flight[str(year)]) for year in years],
    color="maroon",
    label="Air Passenger Traffic"
)

# Add labels and legend
plt.xlabel("Years", fontsize=12)
plt.ylabel("Value (arbitrary units)", fontsize=12)
plt.title("Global GDP vs Air Passenger Traffic (1990-2019)", fontsize=14)
plt.legend()
plt.grid(axis='y', linestyle='--', alpha=0.4)
plt.tight_layout()
plt.show()
```



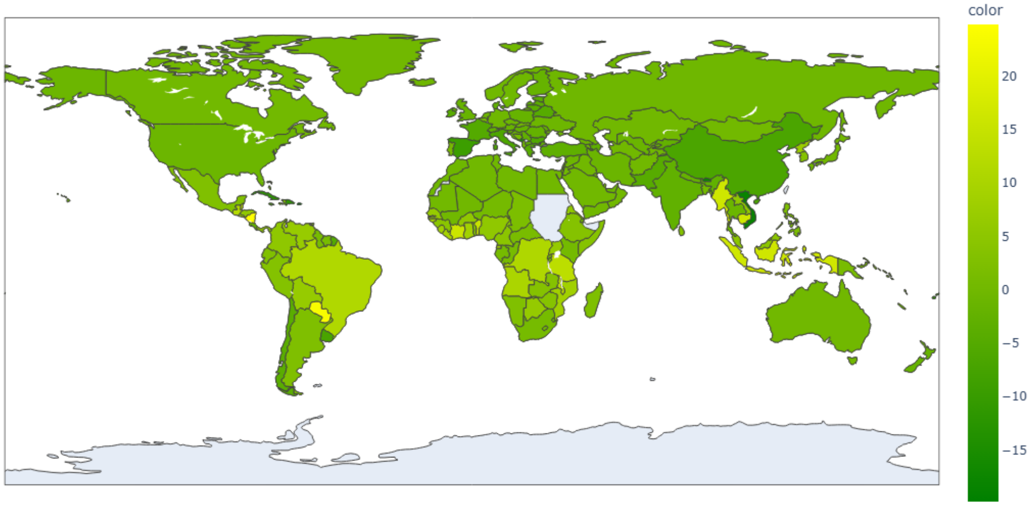
In reference to the third graph, there is a positive correlation between the number of flights and increases in GDP growth. An interesting observation was made on the graph that the rate of increase in the number of flights flattened correlated to the 2008 financial crisis, only further proving the positive correlation between the number of flights had, and economic growth. Where the number of flights started to increase again as soon as the economy started to recover in 2010

8 What is the change in ‘green’ areas in countries between 1990 and 2020?

The third question asked was as follows; “what is the change in ‘green’ areas in countries between 1990 and 2020?”. This is relevant because, one would assume that in the context of economic growth, and industrialization, one would see a significant decrease in the number of ‘green’ areas, where green areas are defined as the % of the state land as forests.

[8]:

```
forest1 = deleteNull[deleteNull["Indicator Code"] == "AG.LND.FRST.ZS"]
fig = px.choropleth(forest1, locations="Country Code",
                    color=forest1["1990"]-forest1["2020"],
                    hover_name="Country Name",
                    color_continuous_scale=["green", "Yellow"],
                    width=1100,
                    height=650)


# Display interactive version in notebook
fig.show()
```



The relevance of forests can not be undermined, not only do they have the positive externality of making the world a little less bleak, but they are also the best natural CO2 suppressor. One can assume that the largest economic powers would have the largest deficit in such areas, but as highlighted in the map. Countries like China, which have experienced the largest rates of economic growth in the past 30 years (the time period within our data set), had negative deforestation rates of 6.66727%. In other words, the 'green' space actually increased. This is a fascinating point because economic development and growth can be achieved without excessive deforestation. Especially when one considers that African states had the opposite reaction where an increase in economic activity lead to higher rates of deforestation. One can attribute this to incompetent policy-making, using India and China as references (developing states with positive reforestation rates) however this only suggests that China and India are entering secondary and tertiary sectors of economies' rather than primary sector ones like Africa.

9 Suggestions

The main problem with this data set was having multiple null values which resulted in losing many entries after the data clearing.

One of the ways of dealing with this problem would be through the use of trained artificial intelligence by machine learning. The AI will get trained on our unimpacted (not deleted/whole) data and then further be trained on existing data with the agenda of restoring the missing values. With our new stronger data we will be able to do more modeling and analysis which we previously could not.

Further, we could use our trained AI to approximate future data. With future data, we could be more aware of incoming problems and challenges and adjust ourselves accordingly. Furthermore, the key to having a rich and neat data set comes down to diligent and precise measuring and data collection. Newly collected data on top of our previously collected data will further enrich the overall data our IA is using to create its weights and that way even our previous data will become more precise.

Furthermore, we could use outsource the data from the other correlated and trustworthy data sets to make our pool of information that much bigger and more pottent to predictions.

To conclude, we live in world of information where we use data daily to analyze information, create models, make predictions, and come to conclusions about the world around us. More data enriches our capabilities to predict the seasonalities and trends in the data. Ergo it is key to obtain as much data as we can and some ways of generating more data are machine learning AI, outsourcing the data, and being more precise when measuring.

Bibliography List all sources you have utilised in the making of this report here. https://www.w3schools.com/colors/colors_hex.asp
https://stackoverflow.com/questions/51161620/linear-regression-python
https://stackoverflow.com/questions/7095441/defining-a-color-in-python
https://stackoverflow.com/questions/43934864/using-matlab-in-python
https://stackoverflow.com/questions/1514553/how-do-i-declare-an-array-in-python
https://www.w3schools.com/python/numpy/default.asp https://www.w3schools.com/python/pandas/default.asp
https://www.db.com/what-next/digital-disruption/better-than-humans/how-artificial-intelligence-is-changing-banking/index