Phone Usage vs. CO_2 Emissions - Zoe Brentzel

https://colab.research.google.com/drive/1zvRXS1eK2oR6vvIFlY9R-eXaT8lWbbFq?usp=sharing

Individual Question

From the question, **how does development influence carbon dioxide emissions in countries across the world?** I decided to explore specifically the correlation between phone usage and CO_2 emissions. As countries develop, they tend to have more access to infrastructure, including phone lines, and people have more access to this technology. CO_2 emissions also tend to increase over time for many countries. Infrastructure is just one aspect of a country's overall development, and I believe it is necessary to analyze how just this aspect relates to the overall CO_2 production of a country, if at all. Combining this question with others regarding different aspects of development will lead to a better understanding of how global development correlates with global emissions.

Approach

For the initial exploratory analysis, I would like to try bivariate analysis between several columns of data, including seeing the relationship between total telephone lines and total carbon emissions and the relationship between mobile cell phone subscriptions and total carbon emissions. I would also like to see if there is a correlation between phone usage and carbon emissions over time using bivariate analysis.

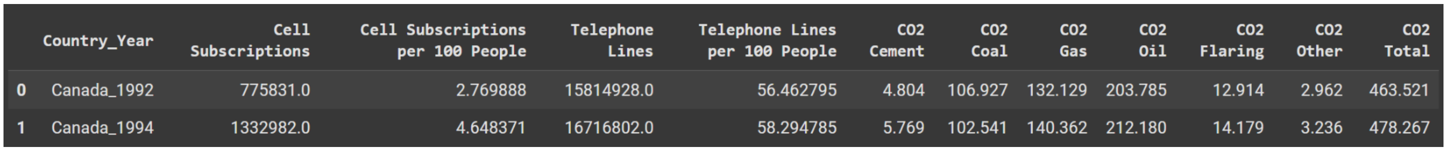
Unit of Analysis

Each data point will represent the data in one particular country during one given year.

Data Querying

```
# from the merged dataset, create a new dataset that contains only necessary columns
df_new = df[['Country_Year', 'Cell Subscriptions', 'Cell Subscriptions per 100 People', 'Telephone Lines', 'Telephone Lines per 100 People', 'CO2 Cement', 'CO2 Coal', 'CO2 Gas', 'CO2 Oil', 'CO2 Flaring', 'CO2 Other', 'CO2 Total']]
```

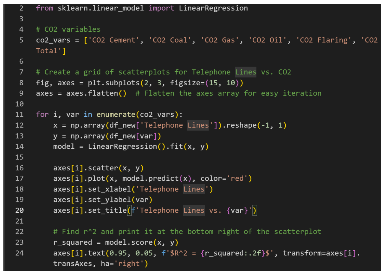
Technically, I got the queried dataset before shortening the column names, but they took up quite a lot of space on this page. I decided on this query because the columns were already categorized, and I chose to analyze the `Data.Infrastructure` variables provided by the dataset. Additionally, I decided to look at all forms of CO_2 emissions rather than just the total amount of CO_2 to see if I could determine which is most correlated to phone usage. Querying just the columns below without reducing the number of rows finished in 0 seconds. It returned these 12 columns for 785 rows.



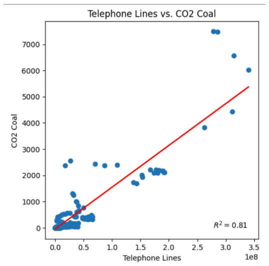
	Country_Year	Cell Subscriptions	Cell Subscriptions per 100 People	Telephone Lines	Telephone Lines per 100 People	CO2 Cement	CO2 Coal	CO2 Gas	CO2 Oil	CO2 Flaring	CO2 Other	CO2 Total
0	Canada_1992	775831.0	2.769888	15814928.0	56.462795	4.804	106.927	132.129	203.785	12.914	2.962	463.521
1	Canada_1994	1332982.0	4.648371	16716802.0	58.294785	5.769	102.541	140.362	212.180	14.179	3.236	478.267

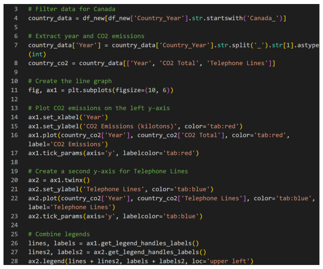
Visualizations and Relevant Code

I enjoyed using Google Colab to generate many different visualizations of the above variables, so there are more visualizations in my Colab notebook (linked at the top of the page) than there will be explained here. I found it extremely useful to generate several scatterplots between the columns 'Cell Subscriptions' and 'Telephone Lines' and each column related to CO_2 because I could see which methods of producing CO_2 had the strongest correlation to phone usage. I found that the strongest correlation was between 'Telephone Lines' and 'CO2 Coal', and the plot is given below (left). This shows that there is a correlation, though not an extremely strong one, between the global development factor of telephone line usage and global carbon dioxide emissions. The other scatterplots are in the Google Colab. I was also able to see how the number of telephone lines in a country correlated with the total CO_2 emissions over time for a given country, and it does appear to have positive correlation. The plot for Canada is given below (right).

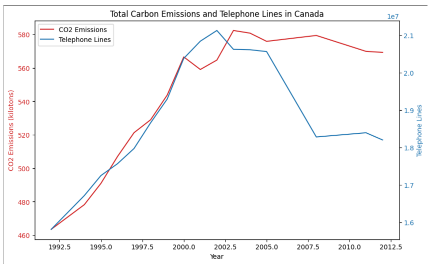


```
from sklearn.linear_model import LinearRegression

# CO2 variables
co2_vars = ['CO2 Cement', 'CO2 Coal', 'CO2 Gas', 'CO2 Oil', 'CO2 Flaring', 'CO2 Total']

# Create a grid of scatterplots for Telephone Lines vs. CO2
fig, axes = plt.subplots(2, 3, figsize=(15, 10))
axes = axes.flatten()  # Flatten the axes array for easy iteration

for i, var in enumerate(co2_vars):
    x = np.array(df_new['Telephone Lines']).reshape(-1, 1)
    y = np.array(df_new[var])
    model = LinearRegression().fit(x, y)

    axes[i].scatter(x, y)
    axes[i].plot(x, model.predict(x), color='red')
    axes[i].set_xlabel('Telephone Lines')
    axes[i].set_ylabel(var)
    axes[i].set_title(f'Telephone Lines vs. {var}')

    # Find r^2 and print it at the bottom right of the scatterplot
    r_squared = model.score(x, y)
    axes[i].text(0.95, 0.05, f'$R^2 = {r_squared:.2f}$', transform=axes[i].transAxes, ha='right')
```





```
# Filter data for Canada
country_data = df_new[df_new['Country_Year'].str.startswith('Canada_')]

# Extract year and CO2 emissions
country_data['Year'] = country_data['Country_Year'].str.split('_').str[1].astype(int)
country_co2 = country_data[['Year', 'CO2 Total', 'Telephone Lines']]

# Create the line graph
fig, ax1 = plt.subplots(figsize=(10, 6))

# Plot CO2 emissions on the left y-axis
ax1.set_xlabel('Year')
ax1.set_ylabel('CO2 Emissions (kilotons)', color='tab:red')
ax1.plot(country_co2['Year'], country_co2['CO2 Total'], color='tab:red', label='CO2 Emissions')
ax1.tick_params(axis='y', labelcolor='tab:red')

# Create a second y-axis for Telephone Lines
ax2 = ax1.twinx()
ax2.set_ylabel('Telephone Lines', color='tab:blue')
ax2.plot(country_co2['Year'], country_co2['Telephone Lines'], color='tab:blue', label='Telephone Lines')
ax2.tick_params(axis='y', labelcolor='tab:blue')

# Combine legends
lines, labels = ax1.get_legend_handles_labels()
lines2, labels2 = ax2.get_legend_handles_labels()
ax2.legend(lines + lines2, labels + labels2, loc='upper left')
```



Urban Development vs. Carbon Dioxide Emissions - Ashwini Jha:

https://colab.research.google.com/drive/1T7wsBfqpqE11ESHCs3SUpopF0SKZScFQ?usp=sharing

Individual Question:

For this project, our preliminary question of interest questions **how development influences carbon dioxide emissions across the globe.** I have chosen to look specifically at how the population in urban areas impacts carbon emissions. I find this to be a relevant area of research since people living in urban areas impact CO2 emissions through increased energy consumption for transportation, buildings, and industrial activities, and a growing population can cause changes in land use and infrastructure development. Not only that, changes in population would affect changes in areas of vegetation and it would also increase landfill use, which both have an impact on overall CO2 emissions. I believe it is necessary to analyze this part of development to see what kind of an impact on CO2 it has, and combined with the other questions our team has asked, I believe it will help us gain incredible insight into how development and carbon dioxide emissions have been related historically.

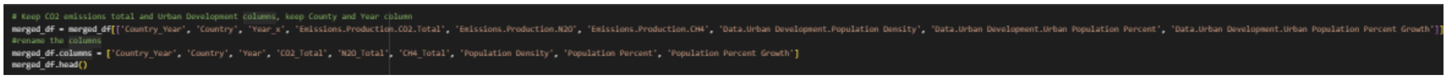
Approach:

I would like to try bivariate analysis for Population Density vs CO2 Emissions Total, Population Percent vs CO2 Emissions Total, and Population Percent Growth vs CO2 Emissions Total. I would also like to perform a multivariate analysis by creating a correlation heatmap for all the columns.

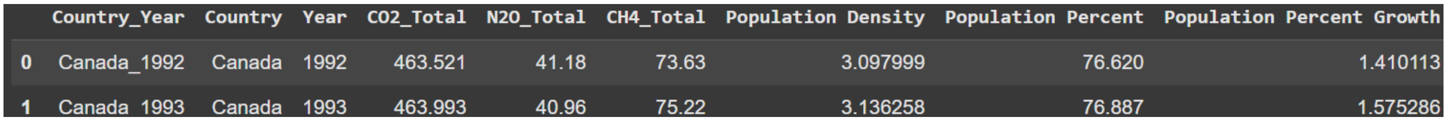
Unit of Analysis: Each row represents the data in one country in a given year.

Data Querying:

The reason I decided to use this specific query is to allow me to identify trends across different years and countries for the total emissions vs urban population. This separates all the columns that are useful for analysis from any irrelevant columns that are in the main merged dataset, thus making the dataset easier to work with. The data frame has 788 entries and 6 columns in total. The Python kernel states that this query took 0s to run.



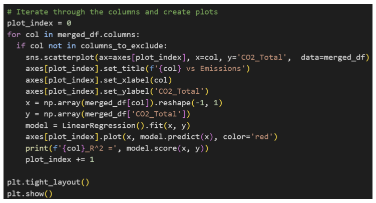
```
# Keep CO2 emissions total and Urban Development columns, keep County and Year column
merged_df = merged_df[['Country_Year', 'Country', 'Year_x', 'Emissions.Production.CO2.Total', 'Emissions.Production.N2O', 'Emissions.Production.CH4', 'Data.Urban Development.Population Density', 'Data.Urban Development.Urban Population Percent', 'Data.Urban Development.Urban Population Percent Growth']]
#rename the columns
merged_df.columns = ['Country_Year', 'Country', 'Year', 'CO2_Total', 'N2O_Total', 'CH4_Total', 'Population Density', 'Population Percent', 'Population Percent Growth']
merged_df.head()
```



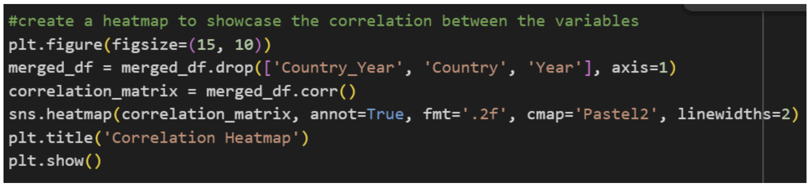
	Country_Year	Country	Year	CO2_Total	N2O_Total	CH4_Total	Population Density	Population Percent	Population Percent Growth
0	Canada_1992	Canada	1992	463.521	41.18	73.63	3.097999	76.620	1.410113
1	Canada_1993	Canada	1993	463.993	40.96	75.22	3.136258	76.887	1.575286

Exploratory Data Analysis and Visualization:

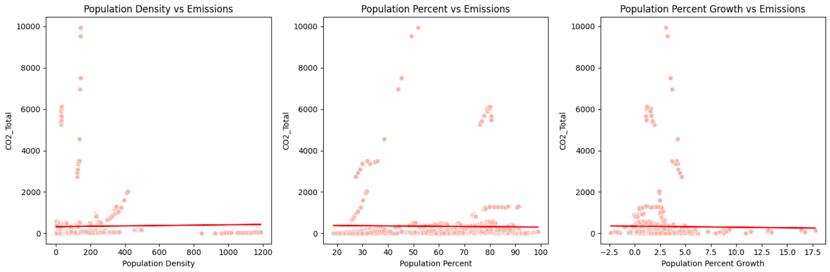
This exercise was really helpful in allowing me to learn more about how the population in urban areas impacts carbon emissions. Visualizing the data through the heatmap and the scatterplot allowed me to see and understand the trends and patterns in the data. Based on these visualizations, there doesn't seem to be any significant correlation between the variables mapped. There do seem to be quite a few outliers, so there are probably a few lurking variables in the dataset that I did not take into account before I mapped them. It was surprising to realize that there wasn't as much of a correlation between columns I was expecting to have a high correlation, and I'm excited to see what more I can learn from this dataset.

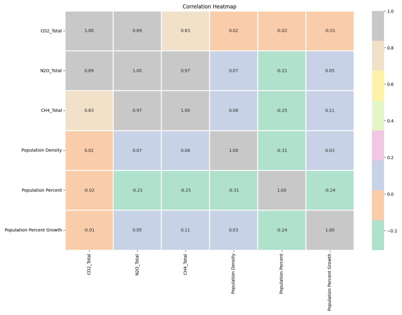


```
# Iterate through the columns and create plots
plot_index = 0
for col in merged_df.columns:
  if col not in columns_to_exclude:
    sns.scatterplot(ax=axes[plot_index], x=col, y='CO2_Total',  data=merged_df)
    axes[plot_index].set_title(f'{col} vs Emissions')
    axes[plot_index].set_xlabel(col)
    axes[plot_index].set_ylabel('CO2_Total')
    x = np.array(merged_df[col]).reshape(-1, 1)
    y = np.array(merged_df['CO2_Total'])
    model = LinearRegression().fit(x, y)
    axes[plot_index].plot(x, model.predict(x), color='red')
    print(f'{col}_R^2 =', model.score(x, y))
    plot_index += 1

plt.tight_layout()
plt.show()
```



```
#create a heatmap to showcase the correlation between the variables
plt.figure(figsize=(15, 10))
merged_df = merged_df.drop(['Country_Year', 'Country', 'Year'], axis=1)
correlation_matrix = merged_df.corr()
sns.heatmap(correlation_matrix, annot=True, fmt='.2f', cmap='Pastel2', linewidths=2)
plt.title('Correlation Heatmap')
plt.show()
```





	CO2_Total	N2O_Total	CH4_Total	Population Density	Population Percent	Population Percent Growth
CO2_Total	1.00	0.89	0.83	0.02	-0.02	-0.01
N2O_Total	0.89	1.00	0.97	0.07	-0.21	0.05
CH4_Total	0.83	0.97	1.00	0.08	-0.25	0.11
Population Density	0.02	0.07	0.08	1.00	-0.31	0.03
Population Percent	-0.02	-0.21	-0.25	-0.31	1.00	-0.24
Population Percent Growth	-0.01	0.05	0.11	0.03	-0.24	1.00

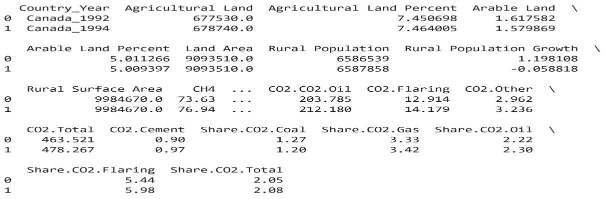
Rural Development Vs. Emissions - Divya Palasamudram
https://colab.research.google.com/drive/18AtDWpHGE8DZkN-iVvp55mZ8vHWEQ-38?usp=sharing

Individual Question

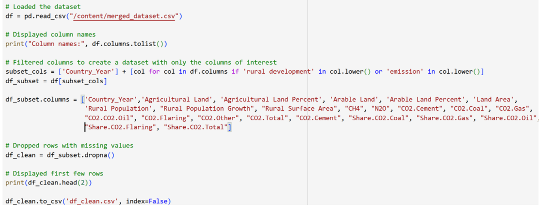
Our preliminary question for this project is **how development impacts/influences emissions across the globe.** In my part, I am exploring specifically how rural areas affect emissions. This is relevant because certain agricultural practices related to agricultural land and arable land can contribute to emissions. Similarly, as population increases and the need for food grows, we may see changes in emission production.

Data Querying:

In order to analyze the relationship between rural development and emissions across the globe, I queried the data so that it contained only the variables of interest (26 columns, 786 rows, run time: 0 seconds). These were the Country_Year variable (unit of analysis) as well as the rural development variables and emissions variables. By isolating these key variables, I made sure the dataset contained only the necessary information to explore the relationship between rural development and emissions across different countries.



```
  Country_Year  Agricultural Land  Agricultural Land Percent  Arable Land  \
0  Canada_1992           677530.0                   7.450698     1.617582
1  Canada_1994           678740.0                   7.464005     1.579869

   Arable Land Percent  Land Area  Rural Population  Rural Population Growth  \
0             5.011266  9093510.0           6586539                 1.198108
1             5.009397  9093510.0           6587858                -0.058818

   Rural Surface Area    CH4  ...  CO2.CO2.Oil  CO2.Flaring  CO2.Other  \
0           9984670.0  73.63  ...      203.785       12.914      2.962
1           9984670.0  76.94  ...      212.180       14.179      3.236

   CO2.Total  CO2.Cement  Share.CO2.Coal  Share.CO2.Gas  Share.CO2.Oil  \
0    463.521        0.90            1.27           3.33           2.22
1    478.267        0.97            1.20           3.42           2.30

   Share.CO2.Flaring  Share.CO2.Total
0               5.44             2.05
1               5.98             2.08
```



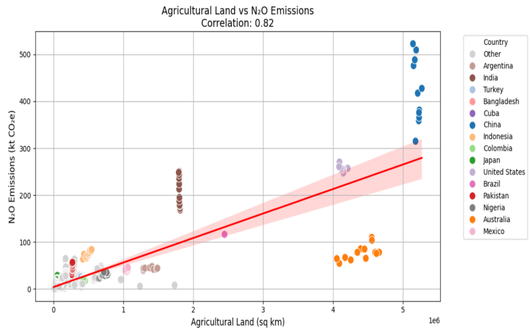
```
# Loaded the dataset
df = pd.read_csv("/content/merged_dataset.csv")

# Displayed column names
print("Column names:", df.columns.tolist())

# Filtered columns to create a dataset with only the columns of interest
subset_cols = ['Country_Year'] + [col for col in df.columns if 'rural development' in col.lower() or 'emission' in col.lower()]
df_subset = df[subset_cols]

df_subset.columns = ['Country_Year','Agricultural Land', 'Agricultural Land Percent', 'Arable Land', 'Arable Land Percent', 'Land Area',
                     'Rural Population', "Rural Population Growth", "Rural Surface Area", "CH4", "N2O", "CO2.Cement", "CO2.Coal", "CO2.Gas",
                     "CO2.CO2.Oil", "CO2.Flaring", "CO2.Other", "CO2.Total", "CO2.Cement", "Share.CO2.Coal", "Share.CO2.Gas", "Share.CO2.Oil",
                     "Share.CO2.Flaring", "Share.CO2.Total"]

# Dropped rows with missing values
df_clean = df_subset.dropna()

# Displayed first few rows
print(df_clean.head(2))

df_clean.to_csv('df_clean.csv', index=False)
```

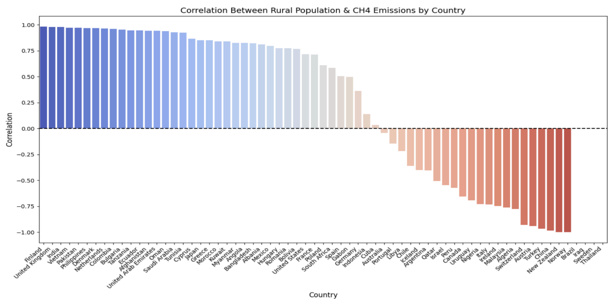
Approach

I will be doing a bivariate analysis on N2O emissions vs Data.Rural Development.Agricultural Land as well as on CH4 emissions and Data.Rural Development.Rural Population. In order to do this, I am first creating a heatmap to see the correlations between these variables and then creating a scatter plot and grouped bar chart to further explore these correlations.

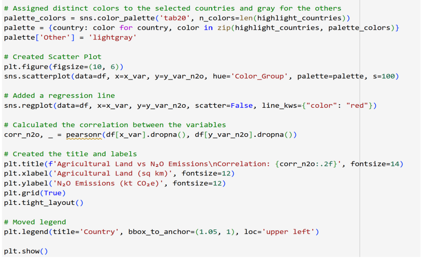
Exploratory Data Analysis and Visualization

In the left visualization below, we see there is a strong positive correlation between agricultural land and N2O emissions. This means that, in general, countries with larger agricultural land areas tend to have higher N2O emissions. Specifically, we see that India and China have relatively high agricultural land areas and high N2O emissions. This suggests that these countries may be more involved in certain agricultural activities such as fertilizer use, livestock management, and soil tillage, as these practices play a large role in N2O emissions. Additionally, as population goes up and the demand for food increases, countries may need to expand their agricultural land, again leading to higher N2O emissions.



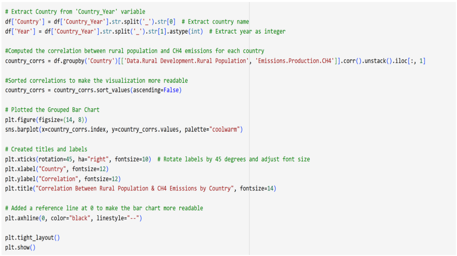


On the right visualization above, on the left side, we see countries that have a strong positive correlation between rural population and CH4 emissions (As their rural population went up, so did CH4 emissions). This suggests that these countries may have engaged more in certain activities including, livestock farming, rice cultivation, biomass burning, and may have less developed waste management systems, as their rural populations went up. All of these activities are significant sources of CH4 emissions. On the right side, we see countries with a negative correlation between rural population and CH4 emissions. This suggests that these countries may have become more reliant on renewable energy and have switched to more sustainable agriculture practices as their rural population went up.



```
# Assigned distinct colors to the selected countries and gray for the others
palette_colors = sns.color_palette('tab20', n_colors=len(highlight_countries))
palette = {country: color for country, color in zip(highlight_countries, palette_colors)}
palette['Other'] = 'lightgray'

# Created Scatter Plot
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x=x_var, y=y_var_n2o, hue='Color_Group', palette=palette, s=100)

# Added a regression line
sns.regplot(data=df, x=x_var, y=y_var_n2o, scatter=False, line_kws={"color": "red"})

# Calculated the correlation between the variables
corr_n2o, _ = pearsonr(df[x_var].dropna(), df[y_var_n2o].dropna())

# Created the title and labels
plt.title(f'Agricultural Land vs N₂O Emissions\nCorrelation: {corr_n2o:.2f}', fontsize=14)
plt.xlabel('Agricultural Land (sq km)', fontsize=12)
plt.ylabel('N₂O Emissions (kt CO₂e)', fontsize=12)
plt.grid(True)
plt.tight_layout()

# Moved legend
plt.legend(title='Country', bbox_to_anchor=(1.05, 1), loc='upper left')

plt.show()
```

Scatter Plot code



```
# Extract Country from 'Country_Year' variable
df['Country'] = df['Country_Year'].str.split('_').str[0]  # Extract country name
df['Year'] = df['Country_Year'].str.split('_').str[1].astype(int)  # Extract year as integer

#Computed the correlation between rural population and CH4 emissions for each country
country_corrs = df.groupby('Country')[['Data.Rural Development.Rural Population', 'Emissions.Production.CH4']].corr().unstack().iloc[:, 1]

#Sorted correlations to make the visualization more readable
country_corrs = country_corrs.sort_values(ascending=False)

# Plotted the Grouped Bar Chart
plt.figure(figsize=(14, 8))
sns.barplot(x=country_corrs.index, y=country_corrs.values, palette="coolwarm")

# Created titles and labels
plt.xticks(rotation=45, ha="right", fontsize=10)  # Rotate labels by 45 degrees and adjust font size
plt.xlabel("Country", fontsize=12)
plt.ylabel("Correlation", fontsize=12)
plt.title("Correlation Between Rural Population & CH4 Emissions by Country", fontsize=14)

# Added a reference line at 0 to make the bar chart more readable
plt.axhline(0, color="black", linestyle="--")

plt.tight_layout()
plt.show()
```

Grouped Bar Chart Code

Analysis Sketch:

Github Repo: https://github.com/zpenguin19/Intro-to-Data-Science/tree/main/Data%20Science%20Project/Homework%206

Goals and subgoals of the dataset:

- Main Goal: determine how different factors of global development contribute to emissions
- Subgoal 1: determine which factors have the highest correlation with emissions
- Subgoal 2: determine if development over time for a given country leads to a change in emissions
- Subgoal 3: predict trends for emissions in upcoming years

Dates and Responsibilities for Each Team Member:

- Before our next meeting, we will each do additional EDA as necessary to determine correlations between necessary variables
- Meet on **April 7th** to discuss our findings and discuss the final week of the project
- By **April 13th**, we will have finished writing our presentation to be submitted, and do any final reviews
- Finalize the presentation on **April 14th** and submit the assignment

Plan for Further Analysis:

Here are some of the ideas for what we want to discuss and for what kinds of models we may want to develop.

- Given labeled data regarding different factors of country development for a given country over a given year (including Total Population, Population Density, Rural Land, Telephone Lines, etc.), we can create a model to predict the total CO_2 emissions by giving a weight to each category based on the correlations we determined during this assignment and additional EDA with other variables. This would be appropriate given our original question because we wanted to know how each factor of development correlated with CO_2 emissions, and with the correlations we found during EDA, we can determine which factors correlate the most.
- Add a column stating whether the country is “developed” or “developing” based on the values for each column regarding development, and create a model to check if the development status of a country impacts its emissions. This seems like a lurking variable for our analysis and could help us understand our data a lot better.
- Create a time series analysis for certain countries that seem to have a significant relation between emissions and any of the factors mentioned earlier.