Economic Impact on Co2 Emissions

DSCI 510

Prof. Jose Luis

Sandeep Kahlon

I. Motivation surrounding project topic

The project aimed to perform an analysis of per capita economic features and their impact and significance on Co2 emissions on a country by country and global scale. As reported by Dartmouth University and multiple sources over the previous decade, 'climate liability is a direct responsibility of the economically affluent nations that contribute significantly to global GDP'. (Dartmouth, 2022). In order to further investigate these claims in a statistical sense, I scraped and gathered data surrounding per capita GDP, automobile count, and emissions statistics by nation. Through descriptive statistics and data aggregation with visuals, I aspired to identify countries with disparities and influxes of GDP and automobile per capita measures and decipher the emissions produced. In addition, independent t-testing was performed in order to statistically examine the significance of the predictive features (GDP and automobile count per capita) effect on emissions per capita. A linear regression was also performed in order to develop a time-series model on emissions per capita for an input country with the ability to predict rates in the future with set confidence. Through completion of these statistical tasks, the final python file can provide intuitive insight into emission rates globally and by nation alongside the economic significance respecting the emission(s) rate.

Ii. Brief description of data sources

The response feature for the study (emissions per capita), was retrieved via download from the Kaggle Website (Gosh, 2021). The dataset was originally gathered through the United Nations over the timespan of 1979-2019 and contained 215 instances respecting each world nation and 35 columns detailing their per capita emission rate for the detailed years. The dataset required transposing in order to optimally aggregate and model the data. Predictive features inclusive of GDP and automobile count per capita were retrieved through scraping via Python packages BeautifulSoup and Requests. GDP per capita statistics by country were retrieved from the World Bank API (World Bank API Documentation, 2023).

To scrape the data a request to the API source was made following the World Bank API documentation in Python and converted to JSON in order to initiate a dataframe. The structure of the link request is detailed in line 10 of Figure 1 and includes an API key, an indicator representing the type of requested data, and the data was filtered while in JSON format to include only 2019 statistics in line 19. The indicator pointing to the GDP data

```
1  # GDP Data from api --> show 20000 per page and match with country name data below --> looks good w country
2  import requests
3  import json
4  import pandas as pd
5  import numpy as np
6
7  form = 'json'
8  indicator = 'NY.GDP.PCAP.CD'
9
10 url = f'http://api.worldbank.org/v2/country/all/indicator/{indicator}?format={form}&per_page=20000'#&source=2'
11
12 # Send a GET request to the API endpoint and store the response
13 response = requests.get(url)
14
15 data = response.json() #Status Code = 200
16
17 country_dic = {}
18 for index in range(len(data[1])):
19     if data[1][index]['date'] == '2019':
20         country_dic[data[1][index]['country']['value']] = [data[1][index]['date'], data[1][index]['value']]
```

Figure 1 - Scraping of 2019 global GDP statistics by nation from the World Bank API

Automobile counts per capita was scraped from the following Wikipedia source (Wikipedia, 2023) which contains a table respecting all world nations and their automobile count per capita statistics for the year 2019. The BeautifulSoup library was leveraged in order to filter and select HTML tags corresponding to the data held within the tables. The scraping process is highlighted in Figure 2 below.

```
from bs4 import BeautifulSoup
import requests
response = requests.get('https://en.wikipedia.org/wiki/List_of_countries_by_vehicles_per_capita')
response.status_code
```

```
200
```

```
#Parse Wiki Data
soup = BeautifulSoup(response.content, 'html.parser')
table = soup.find_all('table')[0]
headers = [val.text.strip() for val in table.find_all('th')] #Headers
```

```
table_rows = table.find_all('tbody')[0].find_all('tr')  #Aggregate Rows
row_data = []
for val in table_rows[1:]:
    for index in range(len(val.find_all('td'))):
        row_data.append(val.find_all('td')[index].text.strip())
```

```
#Convert to DataFrame
import pandas as pd
auto_df = pd.DataFrame()
for col_num in range(len(headers)):
    auto_df[headers[col_num]] = row_data[col_num::4]
#sorted(auto_df['Country or region'].unique())
auto_df.head(3)
```

Figure 2 - Webscraping automobile per capita data in HTML from Wikipedia source.

The data was merged via the primary field ‘Country’ that exists in all three tables with slightly different column names. The merge type was inner and the below carindality and ER diagram is shown below. After the merge two datasets were initiated from the merged data, one was filtered to include 2019 data for analysis of all features and the other including time-series emission values for all nations for the modeling task.

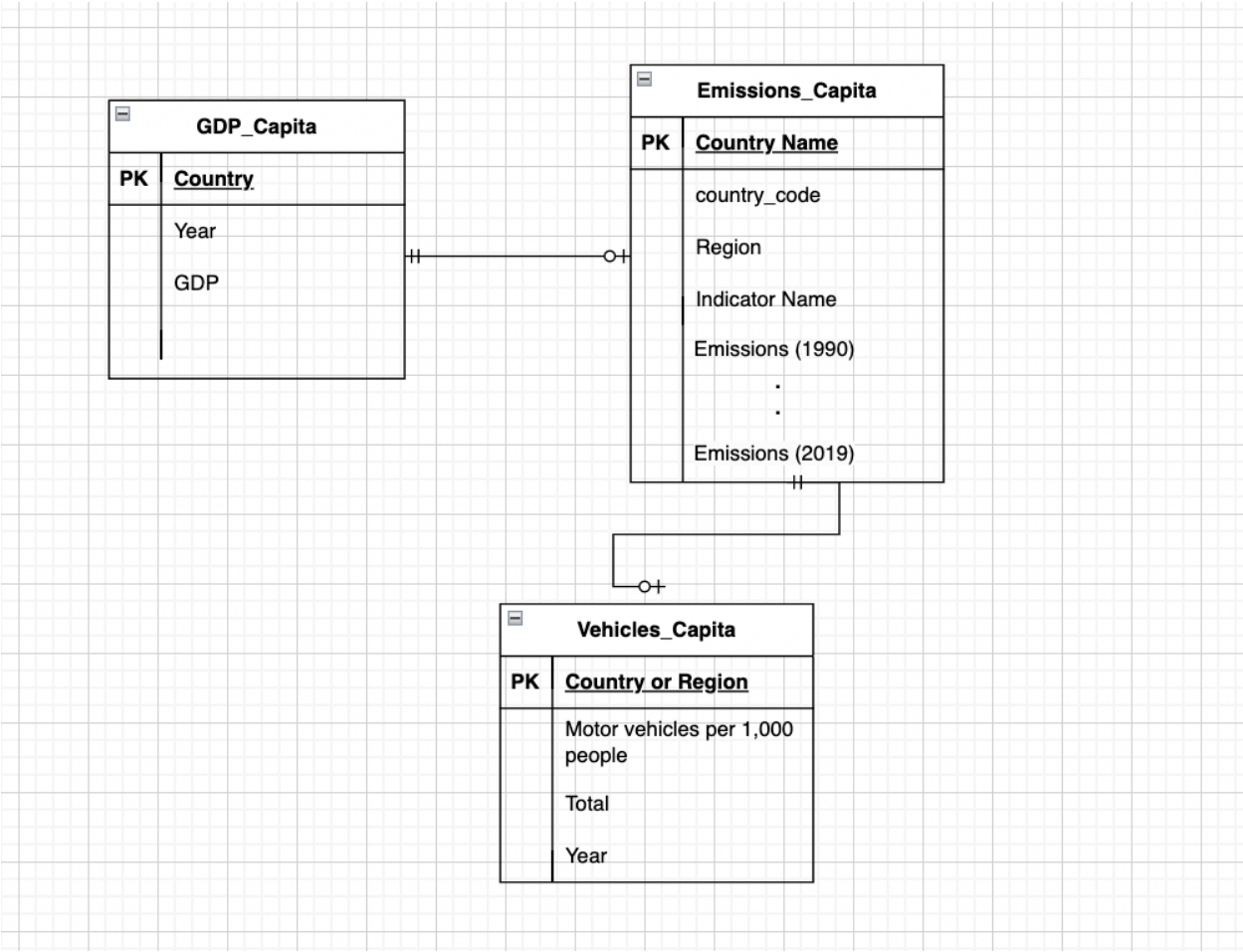


Figure 3 - ER Diagram detailing the three scraped and downloaded datasets along with primary key and cardinality relationship.

Iii. Analysis performed

Initially to view descriptive statistics regarding the predictive and observed variables I developed a class labeled ‘descriptivestats’ that is able to identify numeric fields and perform a thorough descriptive and distribution analysis of the data. Through visualizing the descriptive statistics alongside boxplots and distribution with respective 68%, 95%, and 99.7% bounds, one is able to identify trends within the data. The class has the following functions ‘stats’ and ‘plot’, ‘stats’ is able to output the descriptive statistics regarding the numeric columns and the ‘plot’ function plots the respective columns in a boxplot and

histogram to analyze the statistical distribution. The execution of the class and following functions is displayed below in Figure 4.

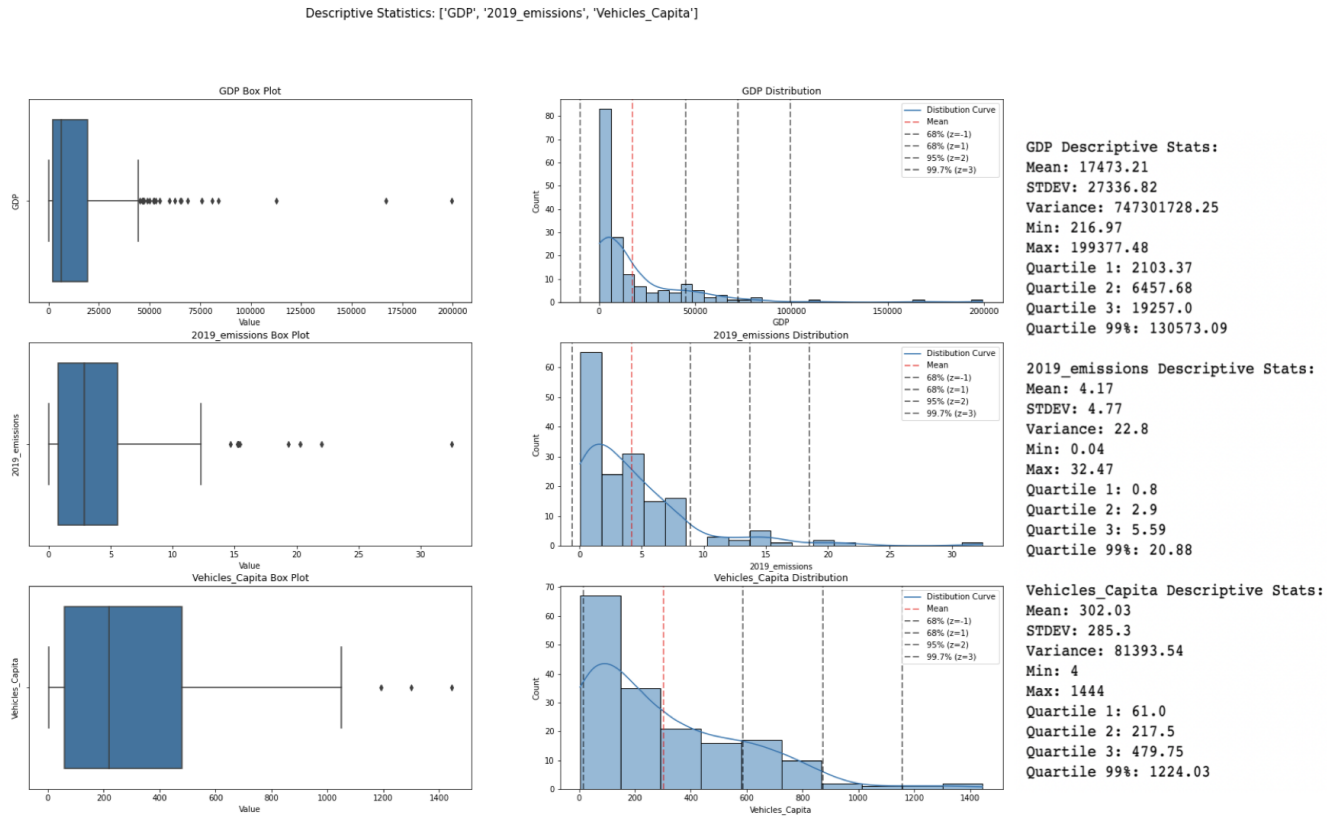


Figure 4 - Output of the execution of the ‘descriptivestats’ class and associated functions, consisting of detailed distribution and descriptive statistics analysis for numeric columns (GDP, vehicle count, and emissions per capita).

To determine highly correlated values that could be associative to each other respecting trend, a Pearson correlation was executed on the predictive and outcome features with an associated heatmap. The highest correlated features were determined to be GDP and automobiles per capita with a Pearson r value of .71.

```
Highest Correlated Features: ('GDP', 'Vehicles_Capita')
Correlation: 0.7145500977834999
```

Figure 5 - Correlation matrix heatmap respecting the analyzed features: GDP, automobile count, and emissions per capita.

User input-driven classes ‘topn’ and ‘worstn’ were developed in order to further investigate the economic impacts on Emissions. The classes both consist of 4 parameters inclusive of the data frame for analysis, the year selected for analysis (can contain the input ‘all’), the integer ‘n’ number of nations to view, and a comma-separated specification of countries to view (can contain the input ‘all’). Both classes contain two functions labeled ‘view’ and ‘plot’, the function ‘view’ provides tabular aggregated results of per capita emissions and the function ‘plot’ projects the data in a boxplot along with the respective ‘n’ nations GDP per capita measures. I leveraged these classes and the associated functions in order to view the top 5 and worst 5 nations' emission per capita and GDP rates via bordering bar plots in Figures 6 and 7.

	Country	Emissions per capita (Year: all)
25	Burundi	0.034940
29	Central African Republic	0.050767
141	Somalia	0.061578
94	Malawi	0.065943
114	Niger	0.072276

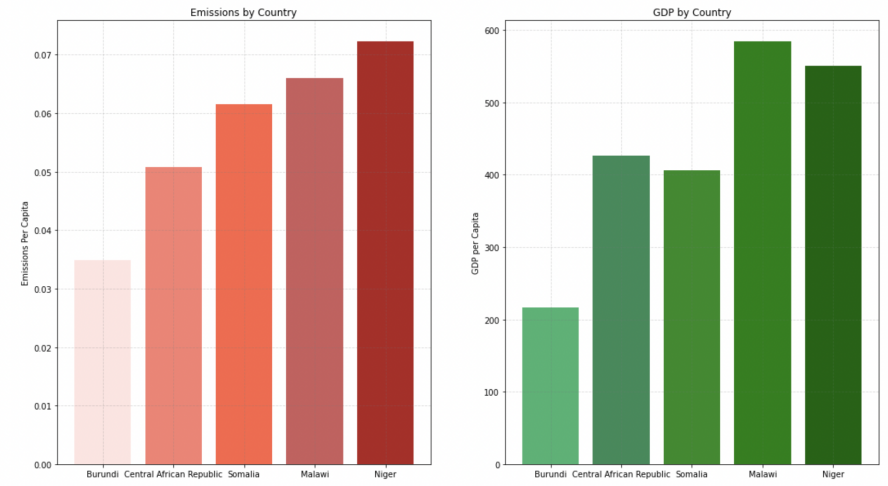


Figure 6 - Execution of the 'topn' class and associated functions with the parameters (emissions_df, 'all', 5, 'all') passed into the class upon initialization.

	Country	Emissions per capita (Year: all)
127	Qatar	38.366840
161	United Arab Emirates	25.056277
84	Kuwait	23.936485
11	Bahrain	22.482269
92	Luxembourg	21.559688

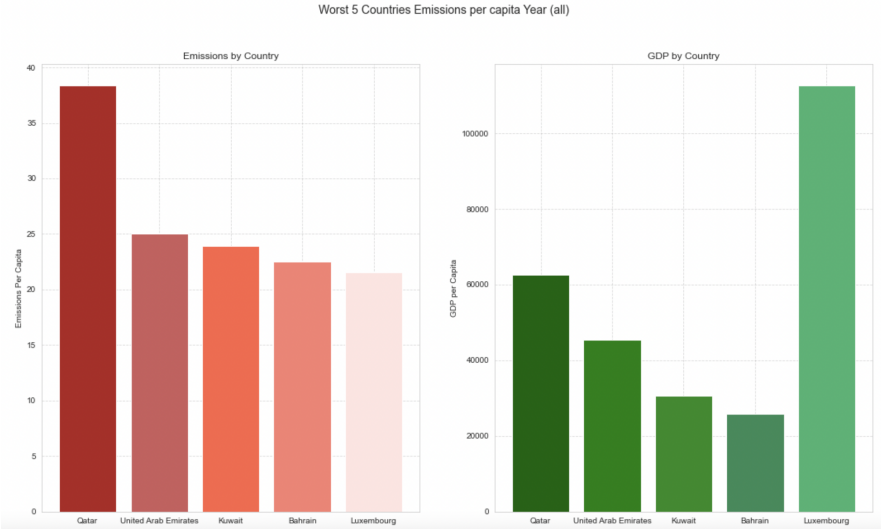


Figure 7 - Execution of the 'worstn' class and associated functions with the parameters (emissions_df, 'all', 5, 'all') passed into the class upon initialization.

In addition to the visual and aggregated analysis, I performed an independent sample t-test to confirm the significance of GDP per capita upon emissions per capita rates. Two groups were tested with group A consisting of nations with above average GDP and group B consisting of nations with below average GDP. The t table value associated with the statistical test was calculated through the Scipy library ‘t’ that required an alpha value for confidence respecting a two-tailed test alongside a degree of freedom value. The t-static value for comparison with the t-table value was calculated manually following the respective formula for independent sampling t-testing as shown in markdown notation in Figure 8. The testing compared the means of both groups with the null hypothesis claiming no significance in the mean value of emissions between the two groups and the alternative hypothesis suggesting significance. As the t-statistic value was calculated to be considerably higher than the t-table value the null hypothesis was rejected at an alpha value of .05 (95% confidence). The calculations for t-table and t-statistic value are scripted in the first code block with testing results, t-statistic validated in the second code block via the Scipy ttest_ind package, and analysis visualized in the third code block.

```
Ho: Mean_a == Mean_b
H1: Mean_a != Mean_b

T table Value= 1.974808091744976
T stat Value= 6.467983251848575
Alpha=0.5

We reject the null hypothesis with 95% confidence and conclude there is a significance between the average number of
Emissions per capita with countries higher thanaverage GDP v.s. lower than average GDP per capita (tstat > ttable)
```

```
#Validate
t_statistic, p_value = ttest_ind(group_a.emissions, group_b.emissions, equal_var=False
                                 )
t_statistic, t_stat
```

```
(6.467983251848575, 6.467983251848575)
```

```
#Visualize conclusion
plt.figure(figsize=(8,4))
sns.scatterplot(x=df_2019.GDP, y=df_2019.emissions, hue=df_2019.Region)
plt.legend(loc='best')
plt.title('GDP vs. Emissions per Capita (2019)')
plt.ylabel('Emissions per Capita')
plt.xlabel('GDP per Capita');
```

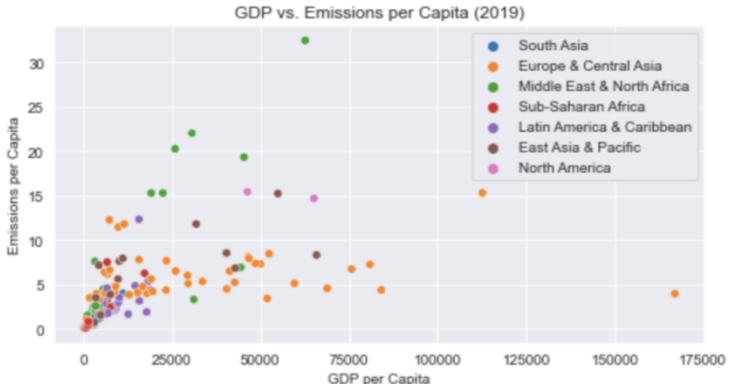


Figure 8 - Independent sample t-test respecting two groups (higher than average GDP nations) + below average GDP nations). Null hypothesis claiming both groups average emissions are the same is rejected (t statistic > t table value).

A global analysis of emissions per capita was displayed through visualizations leveraging Python's interactive visualization library 'Plotly'. I leveraged this library in order to aggregate per capita emission statistics by continent with pandas and further display them in a pie-chart form in order to determine the average percent number of global emissions each continent is responsible for throughout all years of data collection (1957-2019) as shown in Figure 9. In addition, a map was developed using the Plotly choropleth function to highlight global nations and per capita emissions throughout the years respecting a color scale, visualized in Figure 10.

The time-series nature of the dataset allows the map to be ‘played’ or configured over the years of data collection for elevated insight towards emissions per capita.

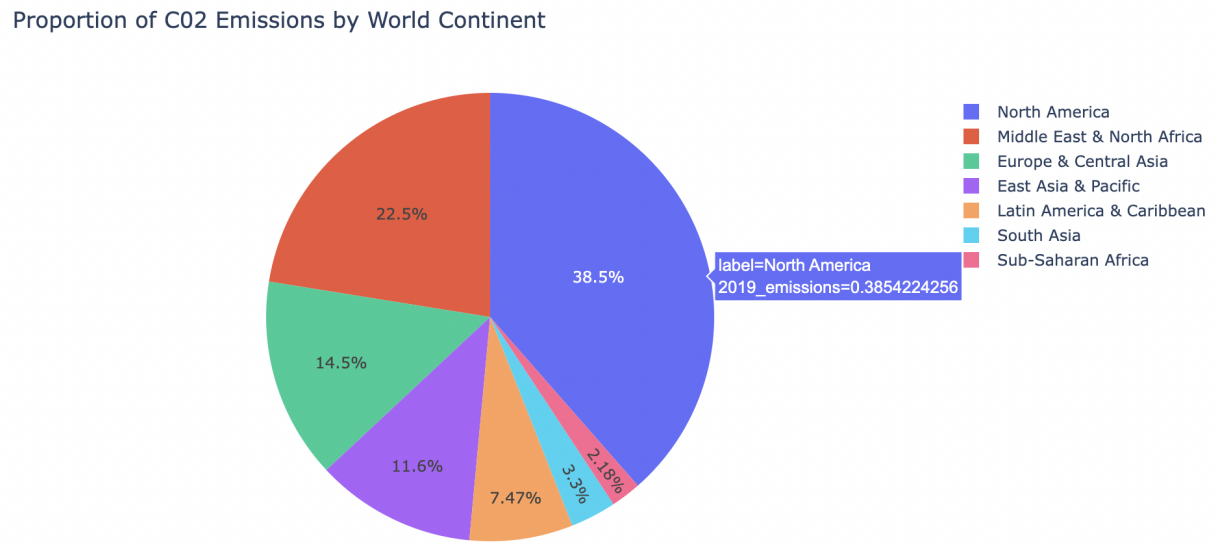


Figure 9 - Plotly pie chart detailing the mean percentage of global emissions that each global continent is responsible for contributing towards.

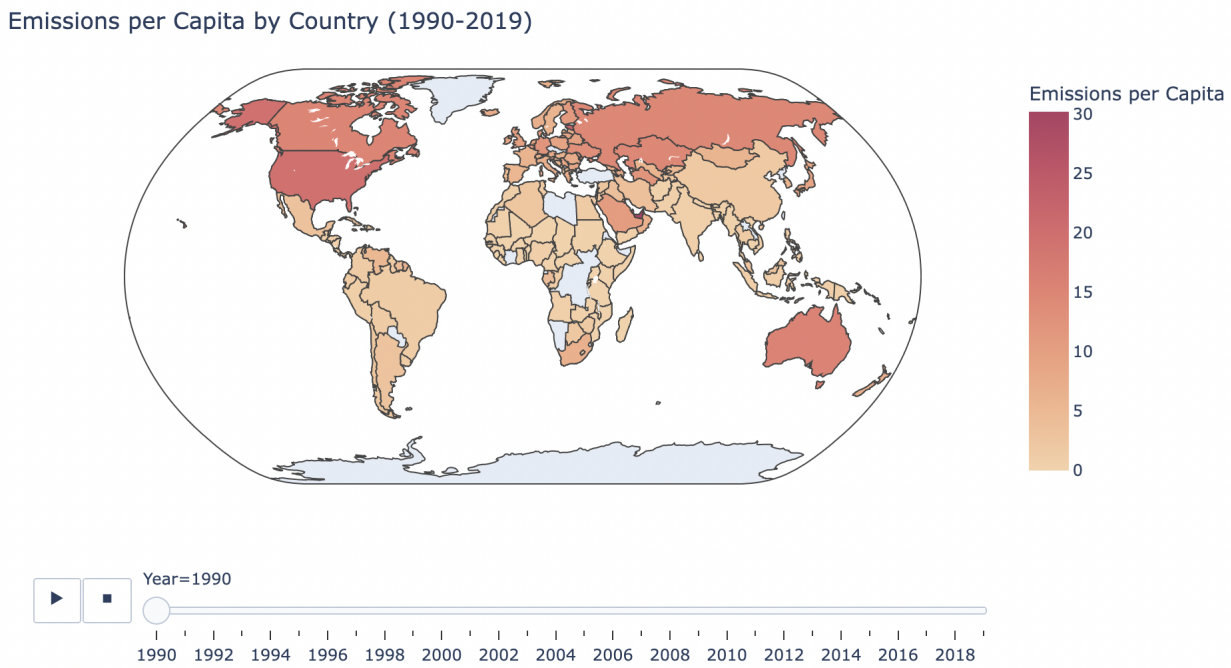
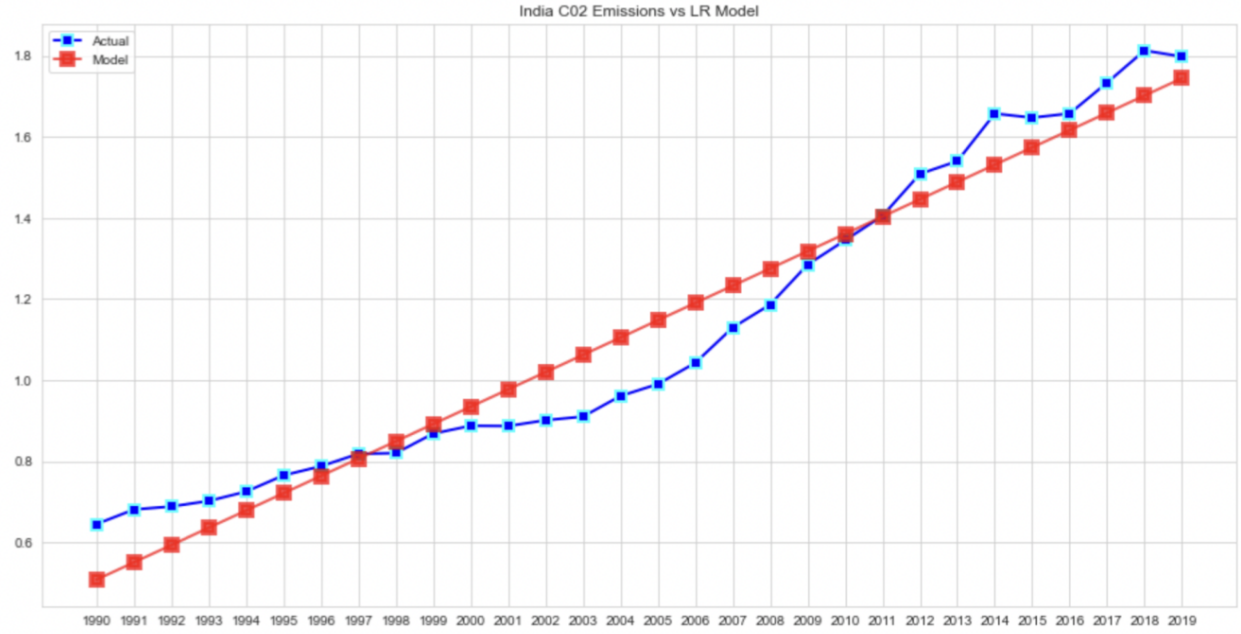


Figure 10 - Choropleth map developed in Plotly to detail in a time-series manner the per capita emissions by country on a global scale.

Lastly, in order to model emission per capita by country a class labeled 'LinearRegression' was developed with the functions 'model', 'performance', and 'predict'. The class contains one input parameter which is the country of selection to model. As most nations contained a linear trend respecting emissions per capita, I believed a time series linear regression would fit the data and allows us to forecast values into the future with high accuracy. The 'model' function of the class is able to manually calculate the slope and intercept parameters required for linear regression upon the specified country, the 'performance' function of the model allows us to view common statistical performance measures of the model inclusive of ME, MAPE, MSE, and MAD, alongside a visualization of the actual emission values for the country and the model. The 'predict' function requires an integer input detailing a future year value to predict, the function will return the same visualization forecasted into the future with a 95% confidence bound of emissions prediction along with the labeled prediction of emissions per capita and 95% confidence interval for the predicted year. The 'LinearRegression' class was initialized to model the nation 'India' emission per capita measures and execution of the associative functions was completed with the following output displayed in Figure 11.

```
Enter a Country to Model: India
bo: -84.3, b1: 0.04262

MAPE: 0.077572908198414
ME: -4.916752190572045e-12
MSE: 0.008008738657996279
MAD: 0.07666409603021039
```




```
Prediction Year 2024: 1.9573358389493052 | 95% conf: (1.5640742260925036, 2.350597451806107)
```

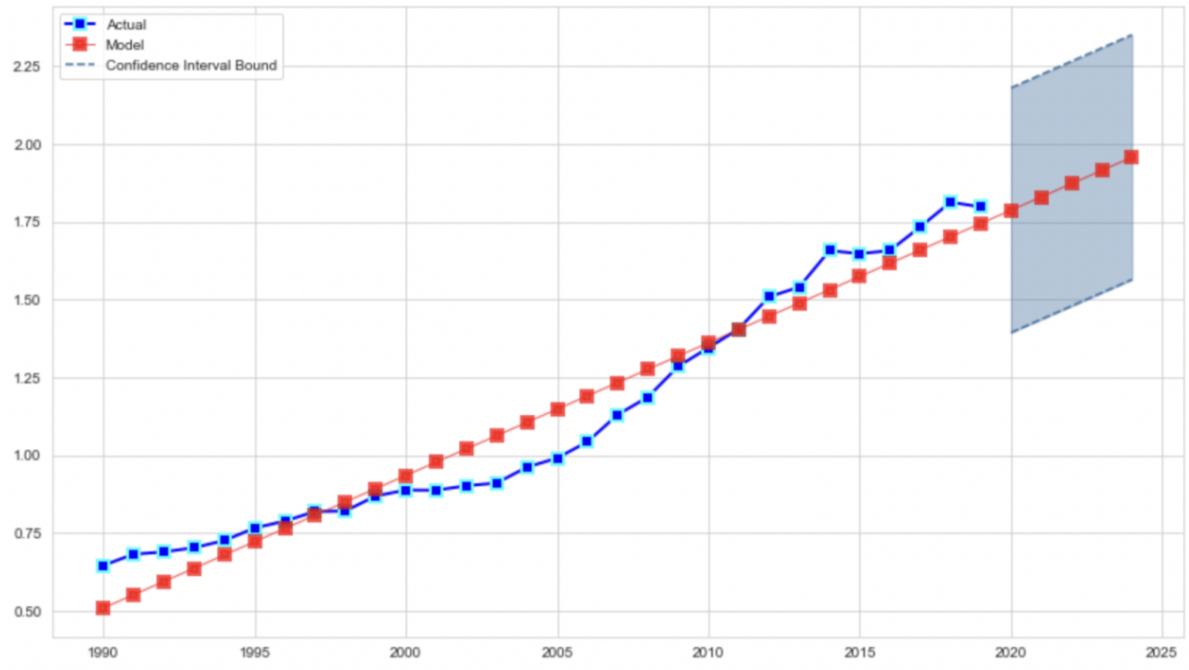


Figure 11 - Output respecting the execution of the ‘LinearRegression’class and associated functions ‘model’, ‘performance’, and ‘predict’ for the nation India. Parameters for the regression, performance metrics, and predictions for the future year of 2024 (with 95% conf. bounds) are returned.

Iv. Conclusions drawn

From the initial descriptive statistics regarding the analyzed features in Figure 1, there remained a spark disparity between the number of nations with higher than average GDP, automobiles, and emissions per capita. A few outliers of nations were observed as outliers in these three categories alongside a long right-tailed distribution for all the columns suggesting a small number of economically advanced nations in the world alongside a small portion of nations contributing a significant rate to global emissions.

The correlation matrix suggested the strongest association features to be between GDP and automobile counts per capita as initially assumed with strong economic nations as visualized via the heatmap in Figure 5. However, a moderate association was revealed respecting the Pearson R value for GDP and automobiles per capita with emissions per capita.

The classes ‘topn’ and ‘worstn’ provided visual output regarding the top 5 and worst 5 nations in a global pool respecting emissions per capita. Viewing the Y axis for GDP of the barplot visualizations, it is evident that in Figure 6 which details the best nations respecting emissions the visualized countries' GDP is considerably lower than the worst 5 emission nations shown in Figure 7.

The independent sampling t-test further confirmed the evidence suggesting that emissions per capita for a given nation is heavily dependent on economic factors. The t-test deployed in Figure 8 suggested significance between the means of per capita emissions of higher and below-average GDP per capita nations at a 95% confidence level. As the calculated t-statistic value was greater than the t-table value respecting 95% confidence (alpha=0.5, two-tailed) the null hypothesis claiming no significance

between the means of two groups was rejected, the associated visual displays a relatively linear trend between GDP and emissions per capita.

On a global scale, we can view an increase in global emissions with a significant rise in the early 2000s as seen in Figure 10 through configuring the time-series icon on the map. A high increase is seen in the middle east during this time suggesting that oil production by nation could also serve as a beneficial predictive feature in analysis or modeling of per capita emissions. The pie chart in Figure 9 also confirms these results and places high emphasis on North America, Europe, and the Middle East for global emission production. Northern America itself, consisting of three nations (United States, Canada, and Mexico) in the leveraged data are collectively responsible for over a third of global emissions.

The Linear Regression model developed worked to model and predict emissions by a user input nation. The model was beneficial and highly accurate respecting the incorporated performance measures in predicting emissions per capita for nations with a linear trend, however this accuracy was often lowered while predicting nations that contained an arbitrary trend. The results in Figure 11, displayed high accuracy in modeling India's emission per capita rate with values forecasted into the future with 95% confidence bounds. One can leverage this model to predict values for any given world nation and can return performance metrics and visualizations for the model fit.

Overall, the completed project provided enhanced knowledge and analysis of the impact of economic factors on emission rates. The project leveraged several statistical and visualization libraries to provide analytic results on emission rates and economic figures while successfully highlighting nations and regions that are highlight contributable to global emissions. Statistical testing identified a significance between GDP and emission values at a per capita metric between nations consisting of below and above-average GDP. The linear regression provided insight into predicting emission values into the future which can be resourceful for resource allocation, prevention, and other analysis. In conclusion, I would suggest a relatively linear trend in global emissions with economically affluent nations in North America, European, and the Middle East responsible for a large portion of emissions.

Sources

Ghosh, Koustav. (2021). CO2 Emission Around The World. Kaggle. Retrieved from https://www.kaggle.com/datasets/koustavghosh149/co2-emission-around-the-world?resource=download

Gibson, S. (2022, July 19). Study shows economic impacts of greenhouse gas emissions. Dartmouth News. Retrieved from https://home.dartmouth.edu/news/2022/07/study-shows-economic-impacts-greenhouse-gas-emissions#:~:text=According%20to%20the%20study%2C%20emissions,each%20for%20the%20same%20years.

Wikipedia. (n.d.). List of Countries by Vehicles per Capita. Retrieved from https://en.wikipedia.org/wiki/List_of_countries_by_vehicles_per_capita

World Bank. (2023). Documents and Reports. Retrieved from https://documents.worldbank.org/en/publication/documents-reports/api