Algonquin College of Applied Arts and Technology

Business Intelligence System Infrastructure

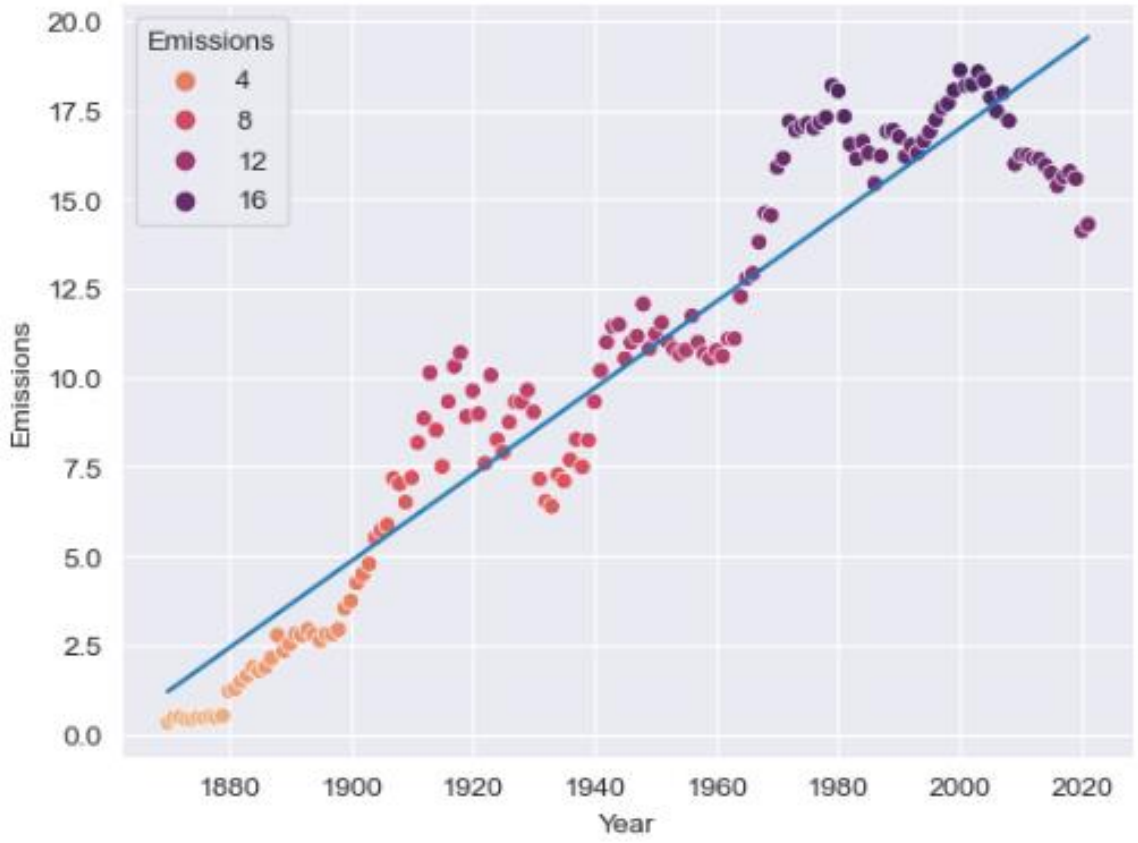
22F_CST2107 – Data Science Foundation for BISI

Final Project

on

Annual CO_2 Emissions (Per capita), its impact on Climate and correlation with GDP

Data Analysis through Machine Learning using Python & PowerBI



Prepared By

First Name	Last Name	Student ID
Esha	Esha	041041612
Ravneet Kaur	Malhi	041083648
Sumita	Dhull	041076073
Wali	Hyder	041057663

Under the kind guidance of Jim Myronyk, MBA, MSc.ISS | Senior Professor, BISI Program Coordinator

Submission Date: 09th December 2022

TABLE OF CONTENTS

1 EXECUTIVE SUMMARY

Group 02 is pleased to present a report on the Topic: ***"Annual CO_2 Emissions (Per capita), its impact on Climate and correlation with GDP - Data Analysis through Machine Learning using Python & PowerBI"***. This work comes along with the IPython Notebook (.ipynb) file, data visualization file (.pbix) and the relevant datasets prepared in a collaborative effort by Esha Esha, Ravneet Kaur Malhi, Sumita Dhull & Wali Hyder (Group 02) to fulfill the requirement of the final project for the CST2107 - Data Science Foundation for BISI course.

Through meticulous qualitative & quantitative data analysis and machine learning models, efforts have been made to find interesting insights and an attempt to predict the Canada's future CO_2 emissions. The problem statements put forward here and explicitly answered in the background & analysis sections of this document are:

1) *"Canada's Per capita CO_2 Emissions and its correlation with the GDP "*
2) *"Impact of rising CO_2 Emissions on the Climate"*

As the primary driver of global climate change, carbon dioxide (CO_2) emissions are widely recognized as compared to the other greenhouse gases (GHGs). Evidently, to avoid the worst impacts of climate change, the world needs to urgently reduce emissions[1]. Through this study we have also shown its correlation with a measure of economy i.e., Gross Domestic Product (GDP).

In the past, CO_2 emissions were strongly correlated with the money we have. This held true particularly for booming economies, The richer we were, the more CO_2 we emit. This is because we use more energy – which often comes from burning fossil fuels. But this relationship no longer holds accurate and we confidently validate this and have demonstrated it in our study[1].

The data democratization tool used for the dataset munging and preparation prior to analysis and final data visualizations is Microsoft PowerBI 2022 Desktop Version 2.111.590.0.

JupiterLab has been used to write the code for the Machine Learning model and the resulting linear regression visualizations.

Key Conclusions

- CO_2 along with other the greenhouse gases (GHGs) are crucial for sustaining a habitable temperature on this planet. Without them,the *Average temperature of Earth's surface would be about −18 °C (0 °F) rather than the present average of 15 °C (59 °F) [2].*
- With the industrial revolution and the massive increase in fossil fuel consumption, CO_2 emissions have skyrocketed.
- Many countries have managed to achieve economic growth while *reducing* emissions, this relationship has been decoupled for developed economies and majority of the developing too.
- To keep the global rise of temperature below 2°C, the current CO_2 emissions need to be reduced to nearly half.

Recommendations

To meet the needs of the future generations, sustainable development goals that have been laid down vary from region to region and are within the context of climate change and other global environmental challenges that have a direct impact. For example - Countries are replacing fossil fuels with low-carbon energy. We can produce more energy, without the emissions that used to come with it.

All countries should proactively take actions to tackle climate change and protect the environment, with the most developed nations bearing the greatest responsibility of implementing and supporting strategies globally.

DATA SOURCE

The basis of this report and the dataset on global emissions used is majorly from Global Carbon Project & The Maddison Project Database. We have also explored World Bank datasets to get accurate data on key economy indicators and to broaden the aforementioned correlation. See Appendix 1.0.0 to view the demographic indicators (fields) in the original dataset and the ones edited. A lot of research and study done to make this report successful is supported by "Our World in Data" website.

SCOPE

The base year taken into consideration for this study and the visualizations is as early as around 1950, soon after the advent of third industrial revolution also known as the digital revolution. Hence, the scope is from 1950 to 2021 and Canada's CO_2 Emissions predictions have been made for 2030, 2040 & 2050.

2 BACKGROUND

The deeper view and a better understanding of the Global CO_2 emissions has steered complete focus of this report on the underlying demographic indicators and the global impact. To combat and try reducing all the impacts, the United Nations (UN) decided to set a target of limiting the average warming to 2°C above the pre-industrial time [2].

As a good standing member of the UN, Canada has committed to decrease their CO_2 emissions over time, and in particular their CO_2 emissions per capita, as the country is ranked one of the top 20 countries in the world in that category [3].Conversely, Canada is also among the top 10 countries to introduce highest amount of carbon-tax to cut down industrial emissions in the country.

Throughout our dataset, Canada's CO_2 emissions per capita data is available and we have utilized it to quantitively forecast the future country's emissions per capita through supervised machine learning model.

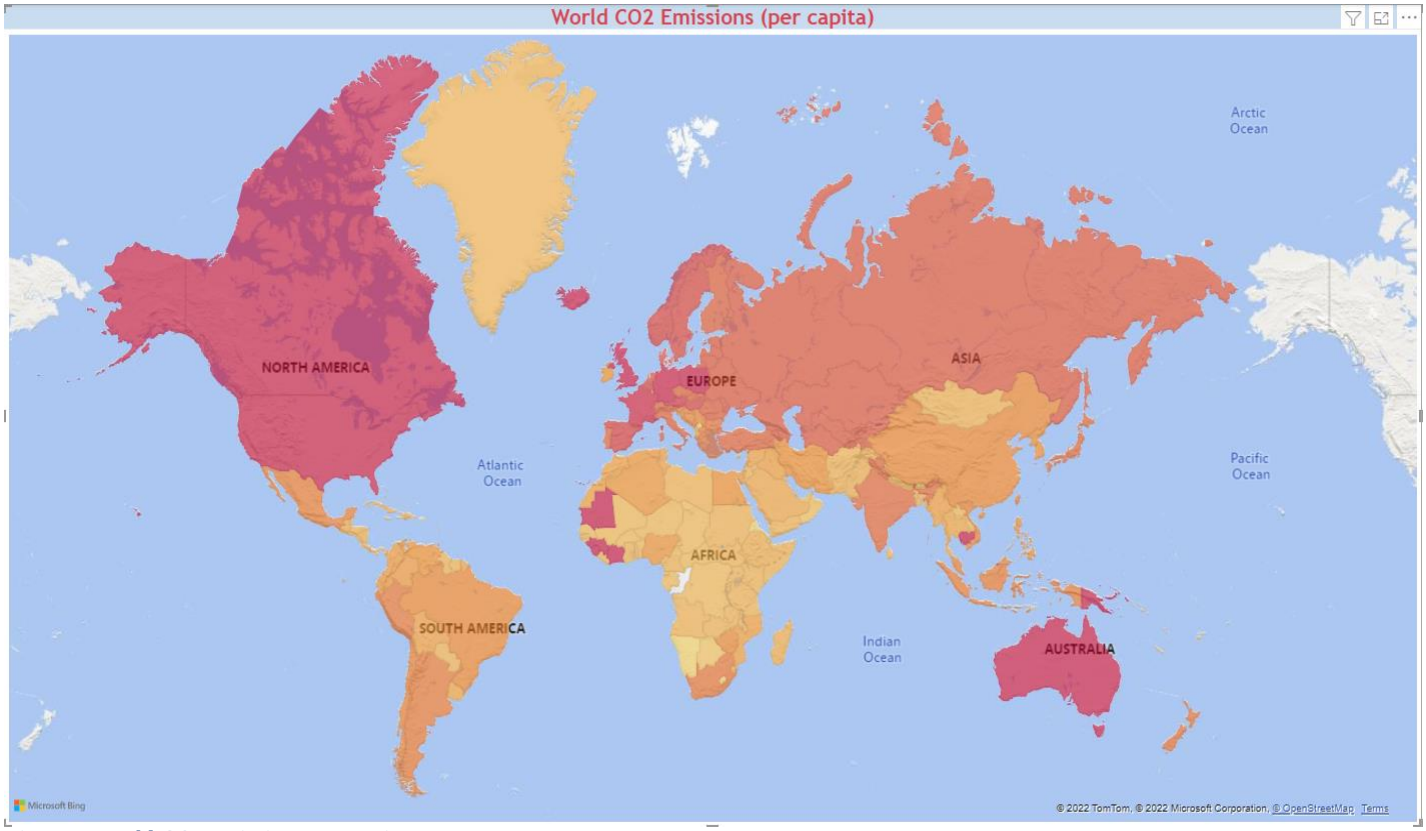


Figure 1 World CO2 emissions per capita

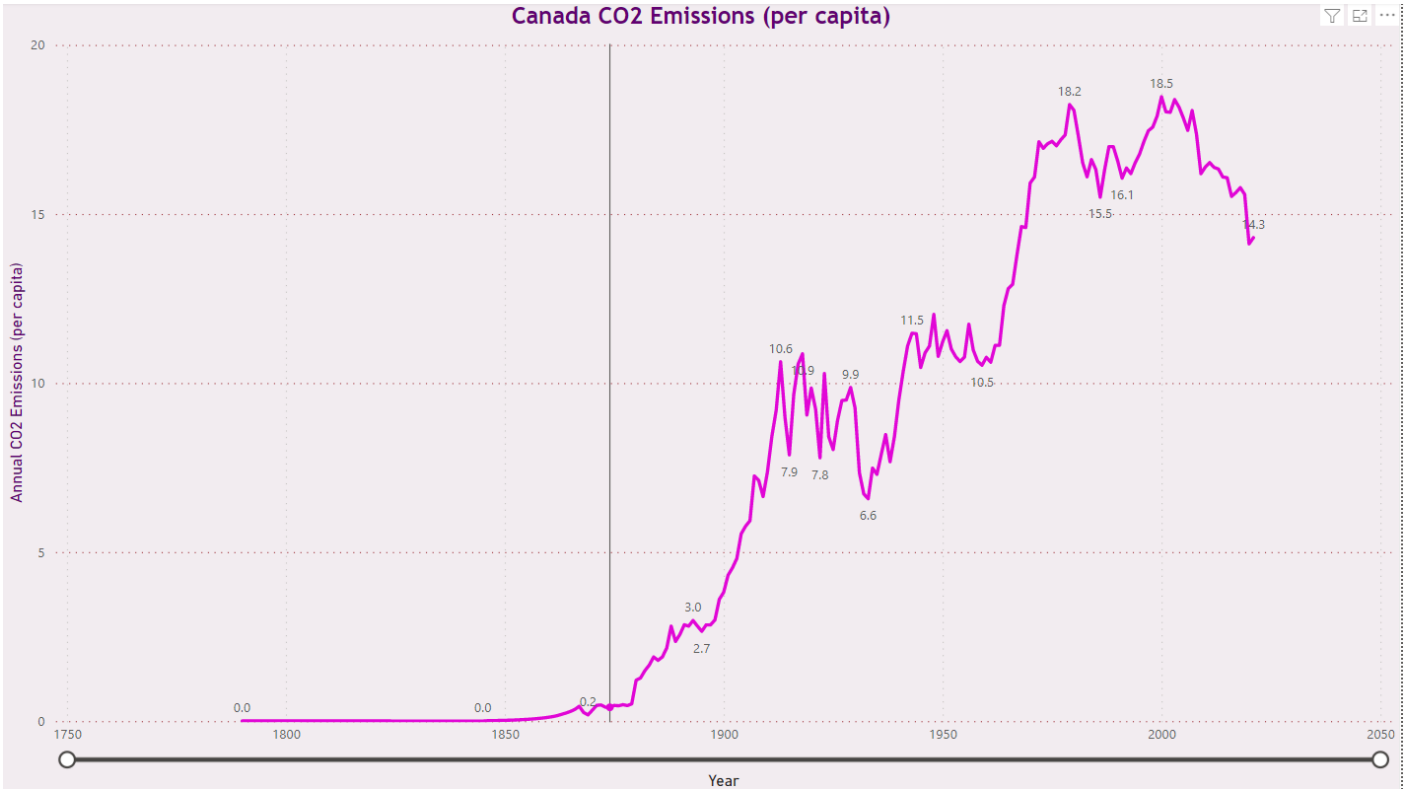


Figure 2 Canada CO2 Emissions (per capita)

Moreover, to cover a major quantity in the world total, the figures of the top 20 countries were taken into further scrutiny and the approach was adopted to further expand on the analysis and visualizations.

Figure 3: 2021 Top 20 in CO2 Emissions per capita

With the Top 20, we could clearly analyze **the key contributors and outliers**. *the* OPEC (Organization of the Petroleum Exporting Countries) being the obvious rank holders here our professor Jim Myronyk was quite interested to know why Trinidad & Tobago (T&T) was among the top 5, with further research we found the small island country contributes a little less than 1% global GHG emissions. In T&T, approximately 80% of total annual emissions are from the power generation and industries. Established Economies and developing countries are also dominating this list. Next, we also visualized the bottom 30 countries.

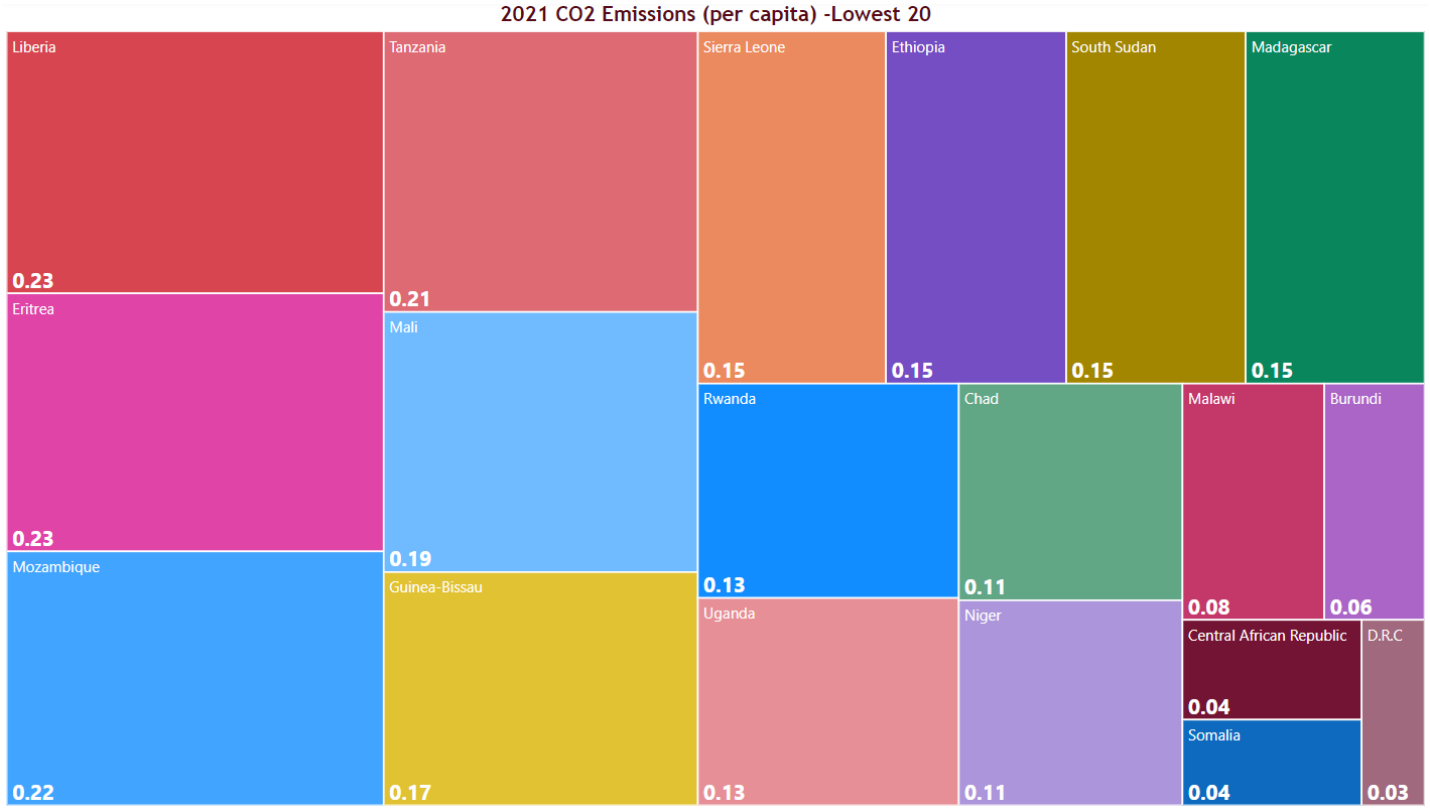


Figure 4 : 2021 CO2 Emissions (per capita) - Lowest 20

Moving forward, the focus of this study was moved towards supervised machine learning through python programming. This was intentional because we wanted to plot the same visualizations in a supervised machine learning model and evaluate the accuracy of the model. The steps are explained in detail in the next section of this document.

3 MACHINE LEARNING USING PYTHON

The analysis of our data will be carried out in line with the problem statements put forward. The following subsections will address the problem statements in detail by providing a supporting analysis from the models and visualizations performed.

PROBLEM STATEMENT 1

"Canada's Per capita CO_2 Emissions and its correlation with the GDP "

APPROACH

I. Importing the required libraries

```
# Importing all required Libraries

import matplotlib.pyplot as plt
import pandas as pd
import pylab as pl
import numpy as np
import seaborn as sns

from sklearn import preprocessing
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import mean_squared_error
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error
from math import sqrt
%matplotlib inline
```

Snippet 1: Libraries imported

II. Loading Dataset

```
#Loading Dataset:

'''This Dataset is loaded from the path specified below, In order to load please use the correct path of downloaded Dataset from your local system.'''

CO2 = pd.read_csv(r'C:\Users\sumit\Desktop\ML Project CO2 emission\Canada_CO2_Emissions_per_Capita.csv')
CO2.head()
```

	Entity	Year	Emissions
0	Canada	1870	0.320623
1	Canada	1871	0.460351
2	Canada	1872	0.471082
3	Canada	1873	0.405382
4	Canada	1874	0.403578

Snippet 2: Dataset loaded

III. Exploratory Data Analysis (EDA)

```
# Exploratory Data Analysis (EDA)

'''The pandas.DataFrame.dropna function removes missing values (e.g. NaN, NaT), if viewed'''

CO2.dropna(how='all', axis='columns')
```

	Entity	Year	Emissions
0	Canada	1870	0.320623
1	Canada	1871	0.460351
2	Canada	1872	0.471082
3	Canada	1873	0.405382
4	Canada	1874	0.403578
...	...	...	...
147	Canada	2017	15.639457
148	Canada	2018	15.778724
149	Canada	2019	15.582994
150	Canada	2020	14.116710
151	Canada	2021	14.300467

152 rows × 3 columns

Snippet 3: Exploring the Data

IV. Change datatype of Emissions column as (float) and setting precision to 5 places

```
# Change datatype of Emissions column as (float) and setting precision to 5 places.

pd.set_option('display.precision', 5)

CO2['Emissions'] = CO2['Emissions'].astype(float)
CO2.head()
```

	Entity	Year	Emissions
0	Canada	1870	0.32062
1	Canada	1871	0.46035
2	Canada	1872	0.47108
3	Canada	1873	0.40538

Snippet 4: Changed Datatype and Precision set

V. Hiding Column

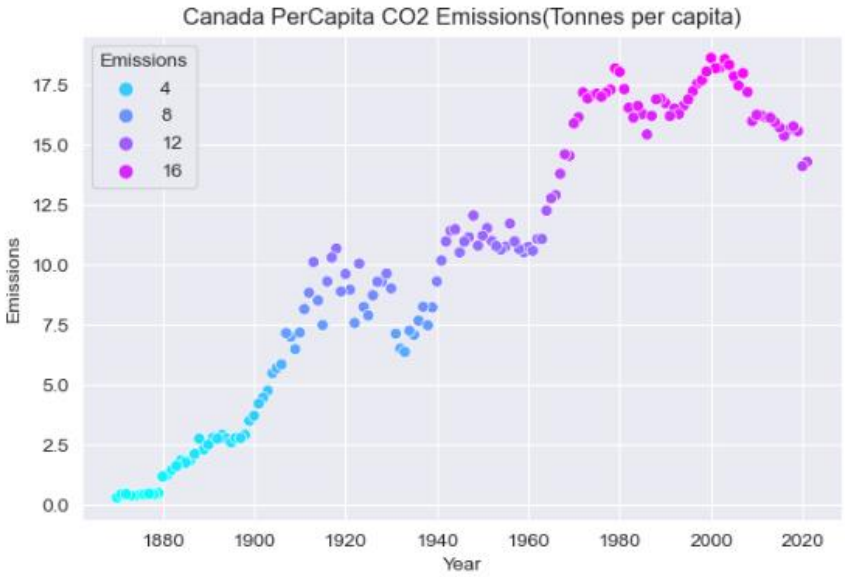
```
# Hiding the Entity column to begin Visualization

cdf = CO2[['Year', 'Emissions']]
cdf.head()
```

	Year	Emissions
0	1870	0.32062
1	1871	0.46035
2	1872	0.47108
3	1873	0.40538
4	1874	0.40358

Snippet 5 : "Entity" column hidden

VI. Defining Samples to create initial visuals (scatterplot)

```
# Defining sample to create intial visuals (Scatter plot)

sample_df = CO2.sample(frac=1.0, random_state=17)

sns.set_style('darkgrid')    # setting grid color

plt.figure(figsize=(7, 4.5))  # define figure size

g1 = sns.scatterplot(data=sample_df, x='Year', y='Emissions', hue='Emissions', palette='cool', legend=True)

g1.set(title='Canada PerCapita CO2 Emissions(Tonnes per capita)') #to add title on graph

plt.show()
```



Snippet 6 :Scatterplot showing Canada Per capita CO2 Emissions (1870-2021)

VII. Splitting the Training and Test dataset

```
# Splitting the dataset in training and Testing - Train : 75% & Test : 25%

'''Due to size of the dataset, we splitted it into default 75-25 for training & testing data respectively'''

X_train, X_test, y_train, y_test = train_test_split(CO2.Year.values.reshape(-1, 1), CO2.Emissions.values, test_size=0.25, random_state=11)
```

```
# Validating Split

X_train.shape
```

```
(114, 1)
```

```
# Validating Split

X_test.shape
```

```
(38, 1)
```

Snippet 7 : Train & Test Spilt

VIII. Training & Testing the Model through Linear Regression

```
# Training and Testing the model through Linear Regression

''' Passing the Train data to the estimator LinearRegression'''

linear_regression = LinearRegression()

linear_regression.fit(X=X_train, y=y_train)

''' Printing the coefficients & intercepts'''

linear_regression.coef_

linear_regression.intercept_

print ('Coefficients: ', linear_regression.coef_ )
print ('Intercept: ', linear_regression.intercept_)
```

```
Coefficients:  [0.12158391]
Intercept:  -226.17194111098672
```

Snippet 8 : Training & Testing using Linear Regression

IX. Predictive Analysis and Results

```
# Predictive Analysis and Results

Predicted = linear_regression.predict(X_test)

Expected = y_test

for p, e in zip(Predicted[::5], Expected[::5]):  # check every 5th element

    print(f'Predicted: {p:.2f}, Expected: {e:.2f}')

'''lambda implements y = mx + b '''

predict = (lambda x: linear_regression.coef_ * x + linear_regression.intercept_)
```

```
Predicted: 15.17, Expected: 16.30
Predicted: 2.28, Expected: 0.51
Predicted: 18.45, Expected: 16.15
Predicted: 6.17, Expected: 8.17
Predicted: 16.27, Expected: 16.64
Predicted: 2.53, Expected: 1.26
Predicted: 1.19, Expected: 0.32
Predicted: 14.56, Expected: 18.05
```

```
# Predicting CO2 emissions in the year 2030 for Canada, The year Canada aims to be Carbon neutral in line with the Paris Agreement
predict(2030)
```

```
array([20.64340537])
```

```
predict(2040)
```

```
array([21.85924452])
```

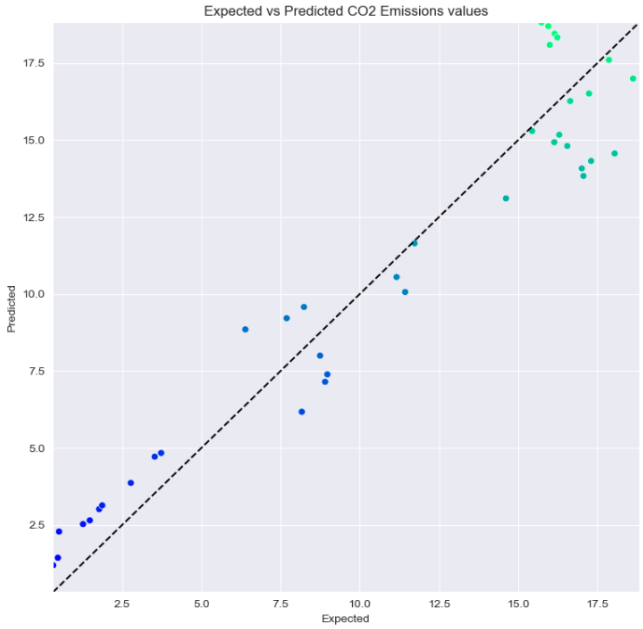
```
predict(2050)
```

```
array([23.07508367])
```

Snippet 9 Expected Vs Predicated & Results for 2030 , 2040 & 2050

X. Visualizing Expected vs Predicted

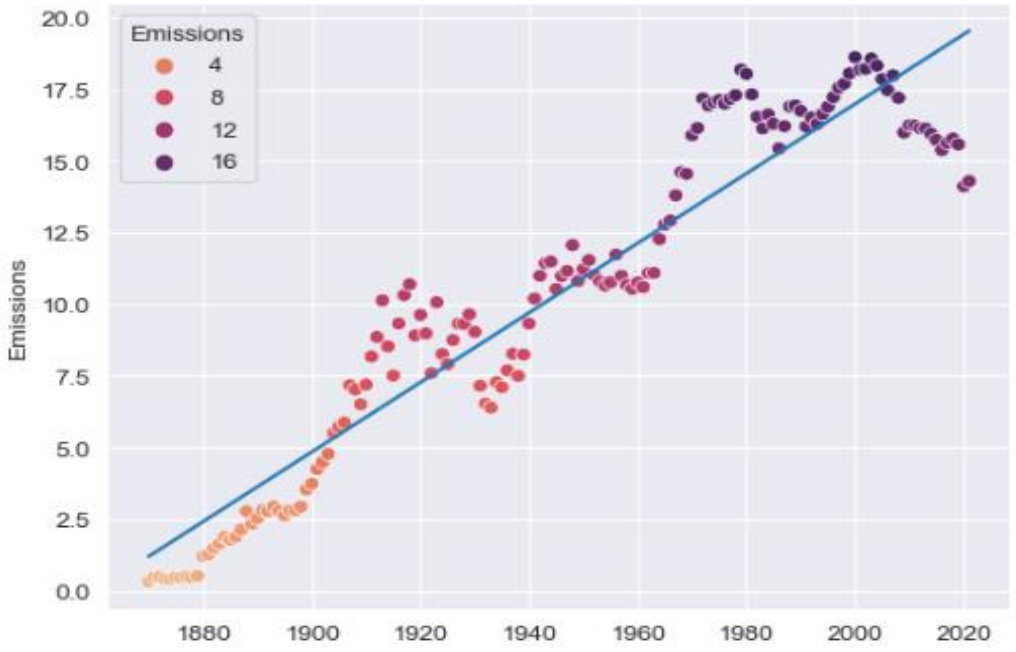
```
# Visualizing the Expected vs. Predicted values for Canada CO2 emissions

CO2_Predict = pd.DataFrame()

CO2_Predict['Expected'] = pd.Series(Expected)
CO2_Predict['Predicted'] = pd.Series(Predicted)

figure = plt.figure(figsize=(9, 9))

axes = sns.scatterplot(data=CO2_Predict, x='Expected', y='Predicted', hue='Predicted', palette='winter', legend=False)

axes.set(title='Expected vs Predicted CO2 Emissions values') #to add title on graph

start = min(Expected.min(), Predicted.min())
end = max(Expected.max(), Predicted.max())

axes.set_xlim(start, end)
axes.set_ylim(start, end)

line = plt.plot([start, end], [start, end], 'k--')
```



Snippet 10 Expected Vs Predicted

XI. Visualization with Linear Regression Line

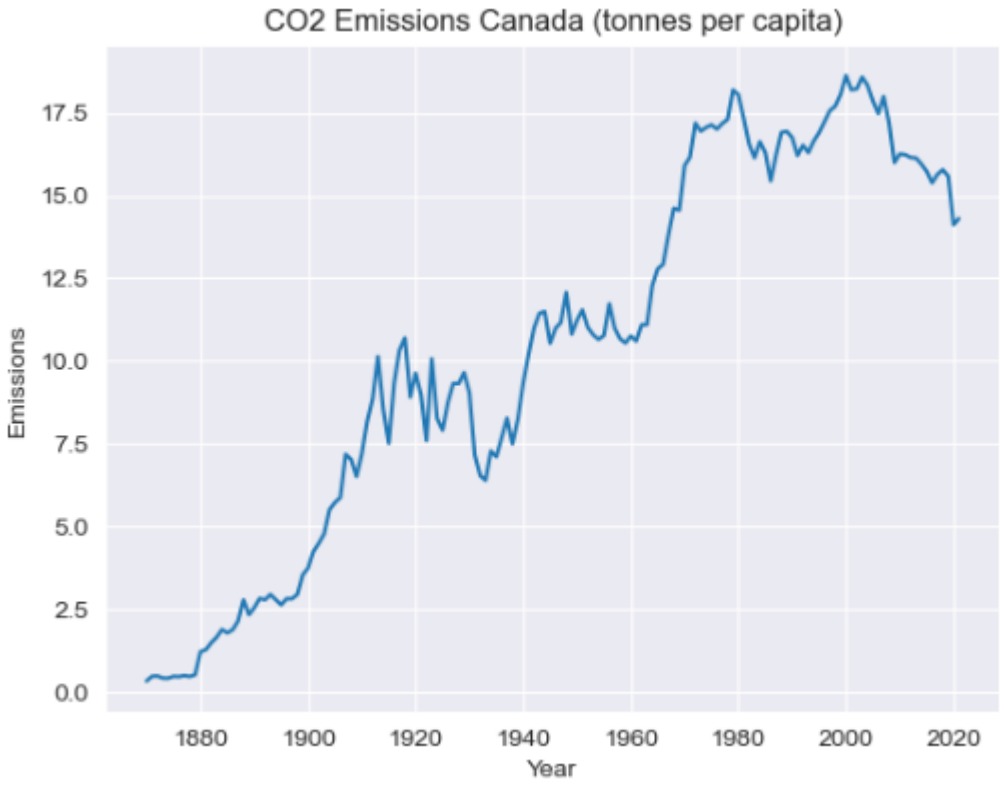
```
# Visualization with Regression Line

axes = sns.scatterplot(data=CO2, x='Year', y='Emissions', hue='Emissions', palette='flare', legend=True)

x = np.array([min(CO2.Year.values), max(CO2.Year.values)])

y = predict(x)

line = plt.plot(x, y)
```



Snippet 11 Linear Regression Line

XII. Time Series

```
# Plotting in Timeseries through Simple Linear Regression

plt.xlabel('Year')
plt.ylabel('Emissions')

plt.plot(CO2['Year'], CO2['Emissions'])

plt.title('CO2 Emissions Canada (tonnes per capita)')

plt.show()
```



Snippet 12 : Timeseries through Simple Linear Regression

XIII. Accuracy Evaluation of the Model

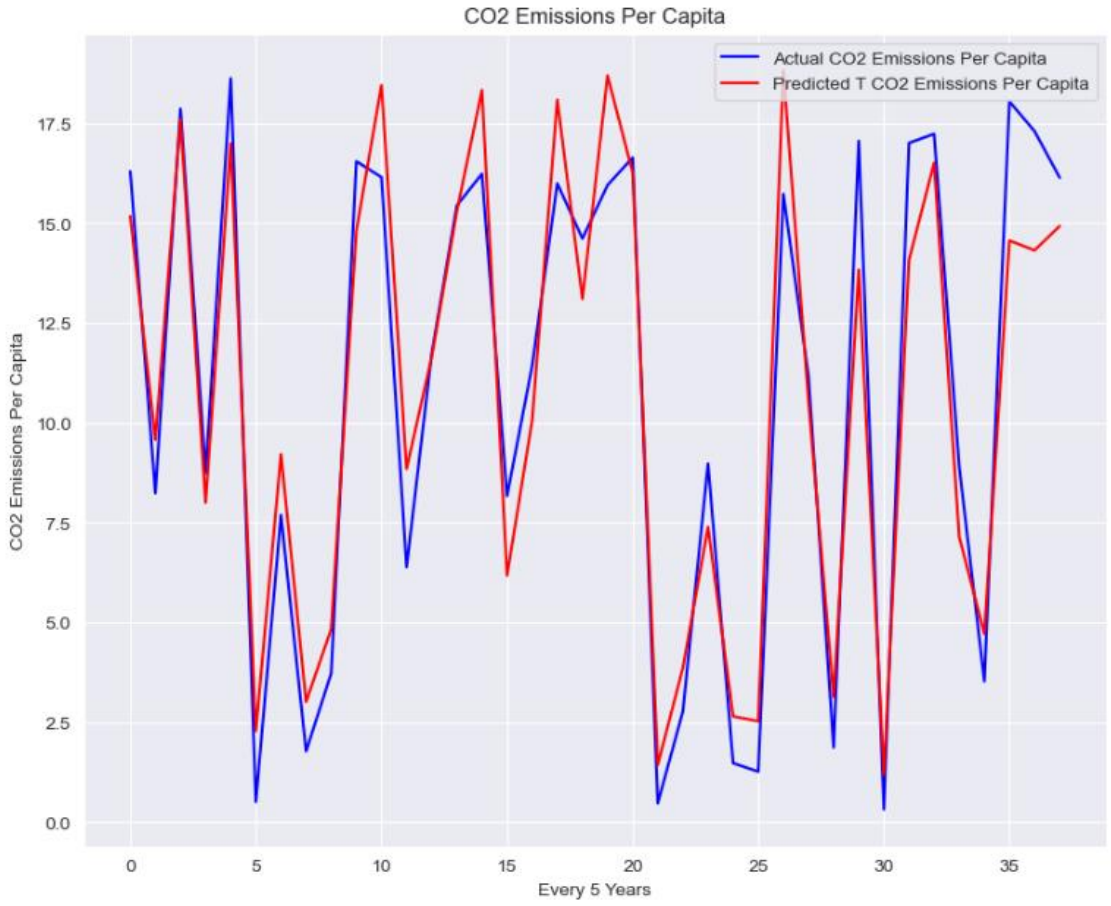
```
# Evaluation of the Model ( Root Mean Squared Error(RMSE), Mean Squared Error(MSE), Mean Absolute Error(MAE))


linear_rmse = np.sqrt(mean_squared_error(Expected, Predicted))

linear_mse = mean_squared_error(Expected, Predicted)

linear_mae = mean_absolute_error(Expected, Predicted)


print ("RMSE: ", linear_rmse) #RMSE
print ("MSE: ", linear_mse) #MSE
print ("MAE: ", linear_mae) #MAE
```

```
RMSE:  1.776750150880522
MSE:  3.156841098653957
MAE:  1.5547228303511558
```

Snippet 13 : RMSE, MSE, MAE Values evaluated for the model

XIV. Accuracy Visualization

```
# Accuracy Visualization

plt.figure(figsize=(10,8))
plt.plot(Expected, color='blue', label='Actual CO2 Emissions Per Capita')
plt.plot(Predicted , color='red', label='Predicted T CO2 Emissions Per Capita')

plt.title('CO2 Emissions Per Capita')
plt.xlabel('Every 5 Years')
plt.ylabel('CO2 Emissions Per Capita')
plt.legend()
plt.show()
```



Snippet : 14 Accuracy Visualization

Correlation with the GDP

To visualize the Correlation for further analysis, we went back to the data democratization tool Power BI. To begin with, the Global GDP (per capita) against the Global CO_2 emissions plot affirms our conclusion that many countries have managed to achieve economic growth while reducing emissions. This relationship has been decoupled for developed economies and majority of the developing too. Contrary to the historic belief that the richer we get, the more CO_2 we emit.

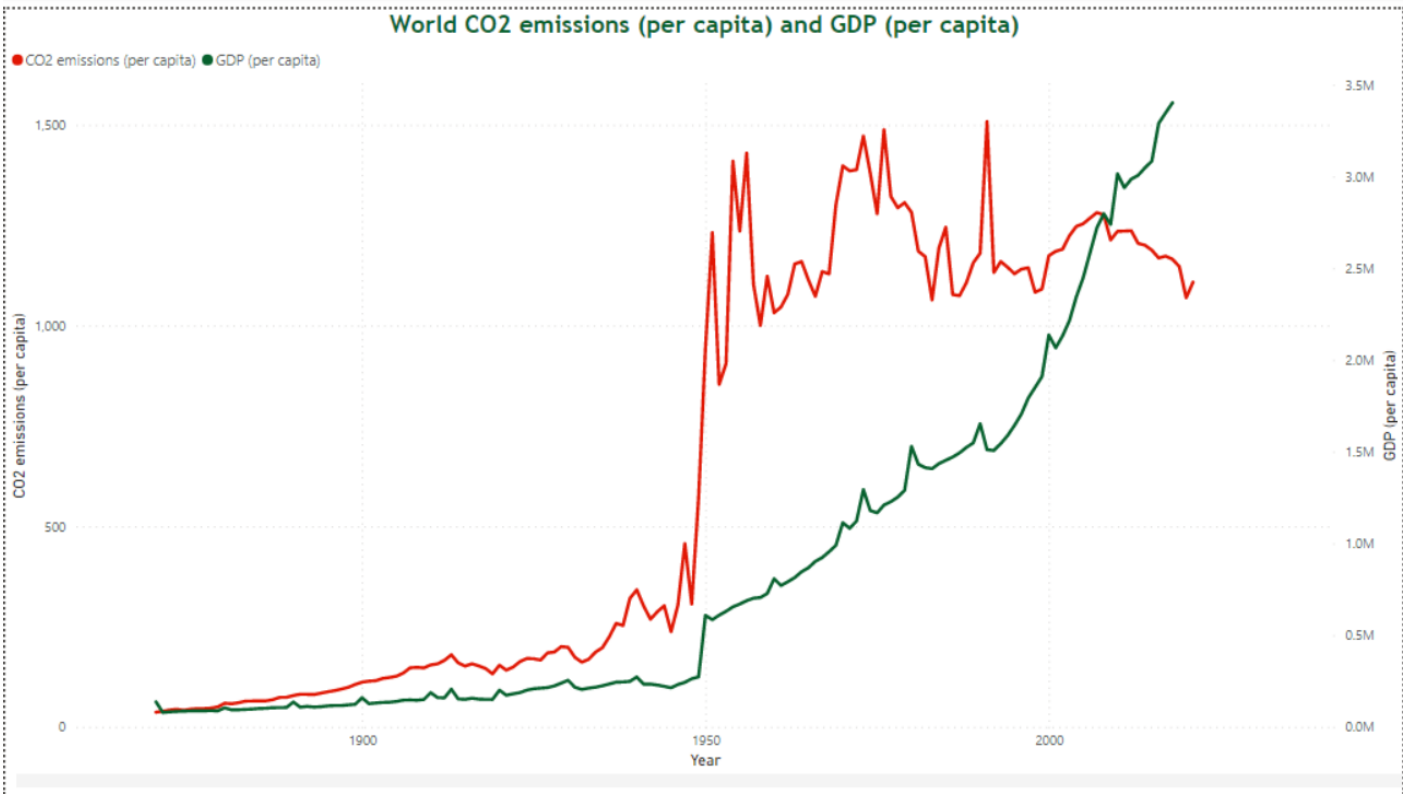


Figure 5 : World CO2 Emissions (per capita) and GDP (per capita)

Now for our country in focus Canada, the correlation plot strongly depicts that even though the CO2 emissions are high, they are on a steady downfall with the growing GDP. This is mainly due to the efforts and policies put in place to reduce the carbon footprint. It's only over the last 20 years that this decoupling has started to happen.

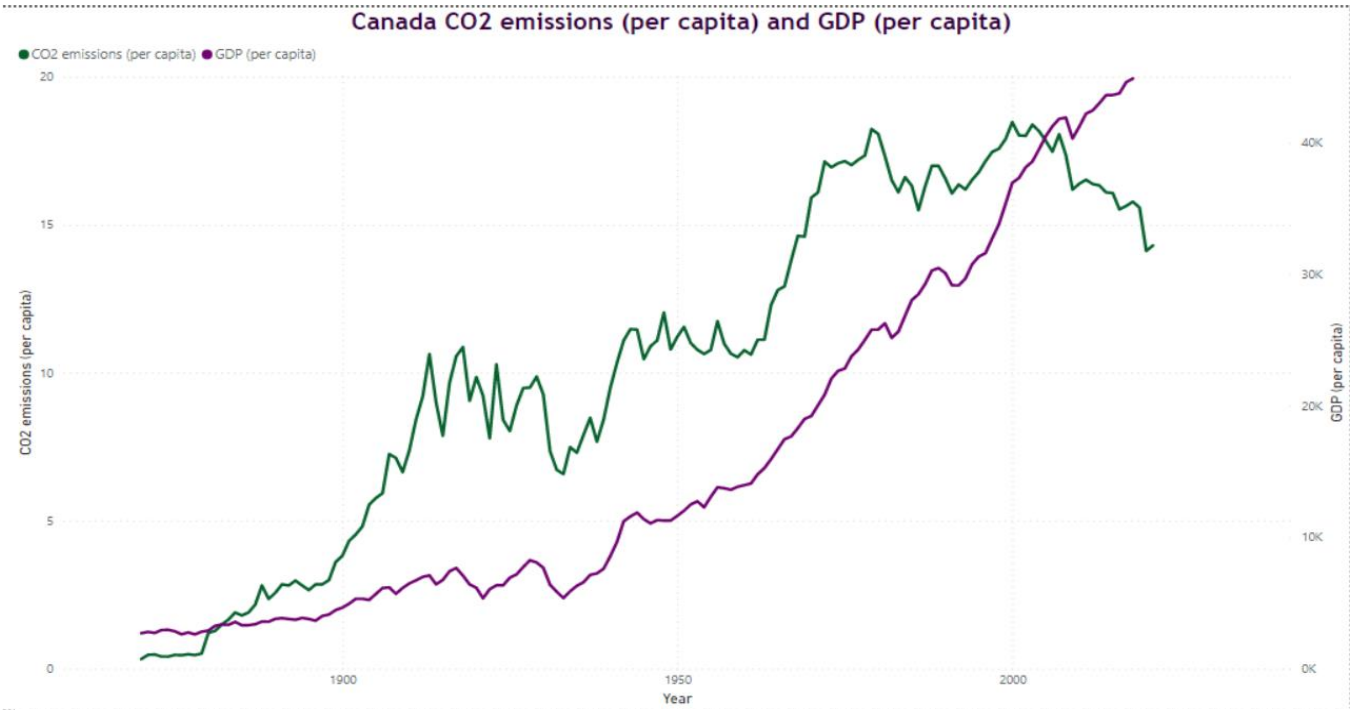


Figure 6 : Canada's CO2 Emissions (per capita) and GDP (per capita)

Furthermore, to see the geographic trend we compared the Nominal GDP for the economic zones with the geographic regions'(continents) Annual CO_2 emissions.

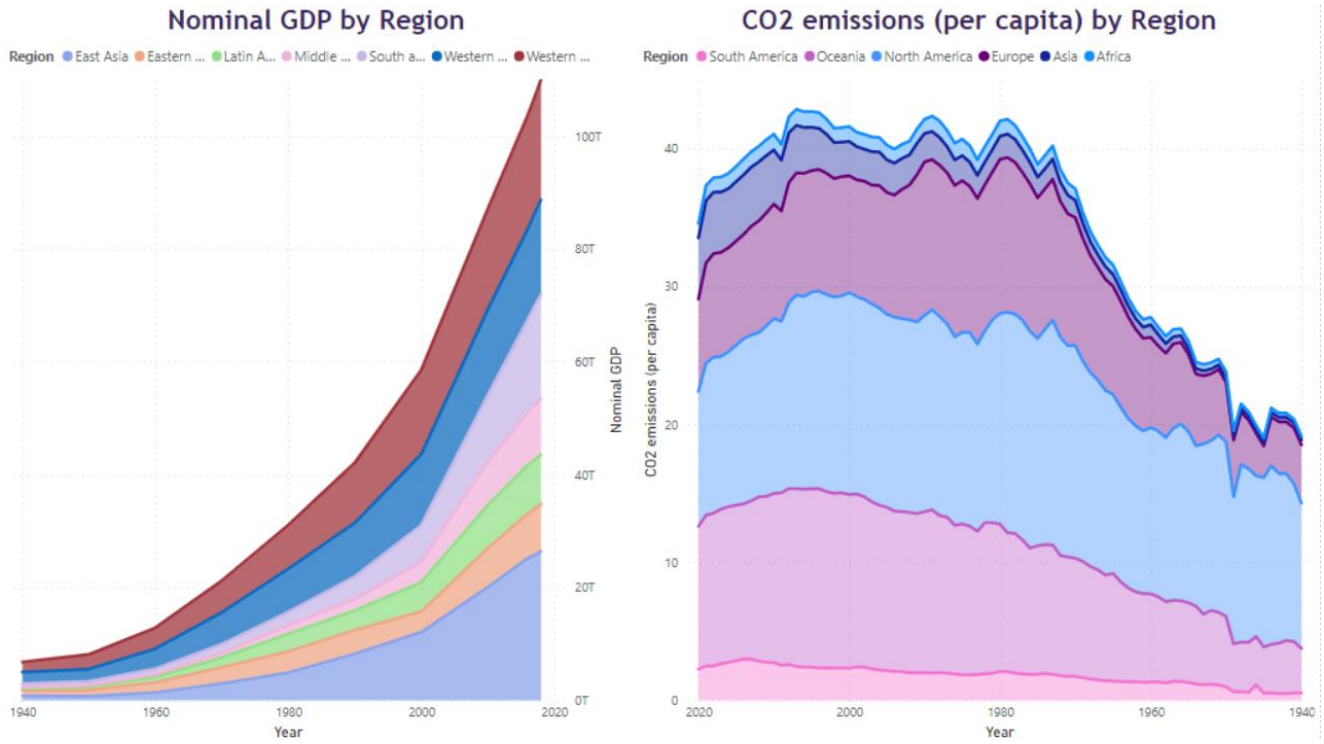


Figure 7 Comparative Study between Nominal GDP & Emissions (per capita) by Region

More and more countries have managed to decouple more recently. Emissions in the North America Region for example, increased substantially in the 1990s. This means that its emissions today are still higher than in 1990, but if we look at the change since 2000, we see a large drop in emissions alongside a rise in GDP.

The two reasons that can be stated for the decline in emissions are that GDP has been increasing almost exponentially while total energy use has plateaued, or even fallen. The most important reason is that countries are replacing fossil fuels and adopting renewable sources with low-carbon energy. They can now produce cleaner and greener energy.

These regions show that economic growth is not incompatible with the reducing emissions.

PROBLEM STATEMENT 2

"Impact of rising CO_2 Emissions on the Climate"

Surface Temperatures globally have increased by more than 1℃ since pre-industrial times. Rising emissions of primarily carbon dioxide and other greenhouse gases are the main driver of climate change - the world's one of the most crucial challenges[3].

To understand better let us present our visualization on how the planet has warmed. The dark blue line represents the median temperature trend through time, with upper and lower values intervals in red and light blue respectively.

We see that over the last few decades, global temperatures have risen sharply — Overall, this would amount to an average temperature rise of 1.1°C leaving us only 0.9°C behind the United Nations' (UN) target of limiting the average warming to 2°C above the pre-industrial time [2] in the Paris Agreement.

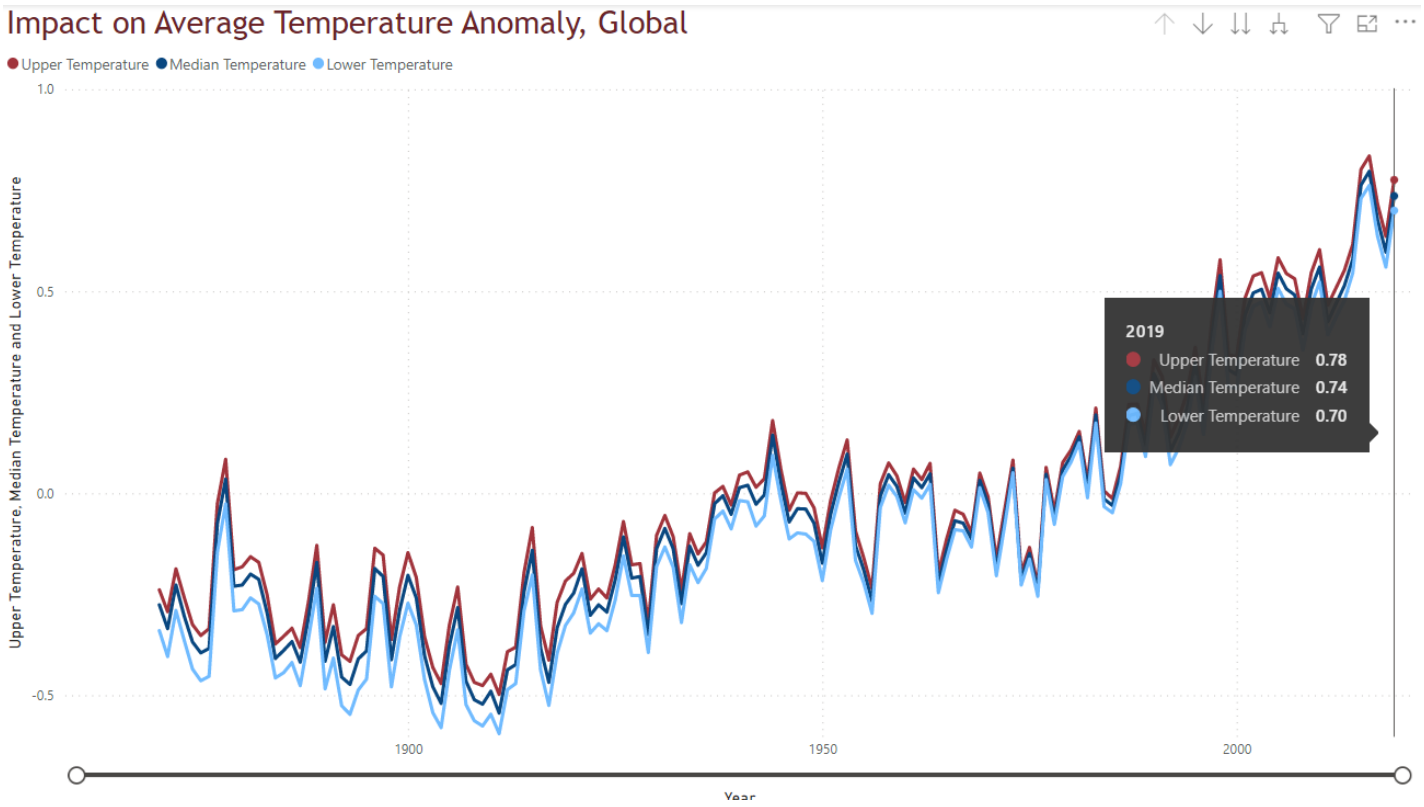


Figure 8 : Impact on Global Temperature

Globally, current policies and measures to reduce or at least slowdown in CO_2 and other greenhouse gas emissions will have some impact on reducing

future warming. Countries have set their respective 'Pledges' and are moving towards achieving them. In this regard, the world is making some progress.

But if our aim is to limit warming to "well below 2°C" – as it is laid out in the Paris Agreement – we are clearly far off track.[4]

Therefore, monitoring the average global temperature change is important and all countries should proactively take actions to tackle climate change to be able to pass on a safer planet to the next generations.

4 TEAM ASSESSMENT

This section of the report describes the ways of working followed by the team to make this report a success. Time management with project management methodologies were implemented to achieve daily targets along with exchange of data, information, and knowledge. This was a completely collaborative effort where in each member fulfilled their part in data gathering, cleaning, coding ,visualizations and final report writing with equal contribution. We worked as a team to deliver on-time a high-quality output.

LESSONS LEARNED

The entire team that worked on this project clearly were very motivated to deliver a detailed analysis on the chosen topic and produce substantial amount of output in terms of quality and quantity. The time was indeed a constraint but that made this even more interesting and challenging. Working on a tight deadline project instilled in the team a sense of great time management and responsibility. However, the only shortcoming the entire team accepted in this report was the absence of seasonality and granularity in the datasets that didn't allow to go deeper in analysis.

5 REFERENCES

[1.] GDP per capita correlation with CO2 emissions available at Our World in Data : https://ourworldindata.org/co2-gdp-decoupling

[2.] Definitions of GHGs and the impact on temperature anomaly available at Wikipedia https://en.wikipedia.org/wiki/Greenhouse_gas

[3.] Climatic Impact of CO2 dataset available at *Our World in Data* and the main source Met Office Hadley Centre at[Retrieved on 04th June 2022] https://www.metoffice.gov.uk/hadobs/hadcrut4/index.html and research study available at our World in data https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions#why-do-greenhouse-gas-emissions-matter

[4.] Global Impact and international targets available at Our world in Data : https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions#current-climate-policies-will-reduce-emissions-but-not-quickly-enough-to-reach-international-targets

6 BIBLIOGRAPHY

[1.] The original data set was available at Our World in Data and the source is from the Global Carbon Project https://www.globalcarbonproject.org/about/index.htm

[2.] CO2 Emissions study and research available at Our World in Data : https://ourworldindata.org/co2-emissions

[3.] GDP per capita correlation with CO2 emissions available at Our World in Data : https://ourworldindata.org/co2-gdp-decoupling

[4.] Climatic Impact of CO2 dataset available at *Our World in Data* and the main source Met Office Hadley Centre at[Retrieved on 04th June 2022] https://www.metoffice.gov.uk/hadobs/hadcrut4/index.html

[5.] World bank dataset for GDP nominal and per capita was available at https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

[6.] The Maddison Project Database provides information on comparative economic growth and the GDP dataset was available at : https://www.rug.nl/ggdc/historicaldevelopment/maddison/releases/maddison-project-database-2020?lang=en

7 APPENDIX (1.0.0)

This section contains the list of all fields or the "parameters" as we call them in the original dataset.

The dataset extracted from the 'Our World in Data' website as mentioned in the previous section includes 4 columns: 'Entity', 'Code', 'Year, and 'Per capita CO2, emissions (tonnes per capita)'. The following table gives a description of each field and its record.

Column	Description	Data example
Entity	Contains the full names of all countries in the world.	'Canada', 'United Kingdom'
Code	Contains the code of the corresponding country in the 'Entity' column	'CAN' , 'UK'
Year	Contains Year	2017
Emissions per capita	Contains Annual CO_2 per capita value	12.08

Omitted columns:

1) Entity column for machine learning model
2) Unnamed Columns seen upon loading the dataset on Jupyter Lab

Renamed columns:

1) "Entity" column renamed to "Country" in PowerBI
2) "Per capita CO2 Emissions" column to "Emissions" in Jupyter Lab

APPENDIX 1.0.1

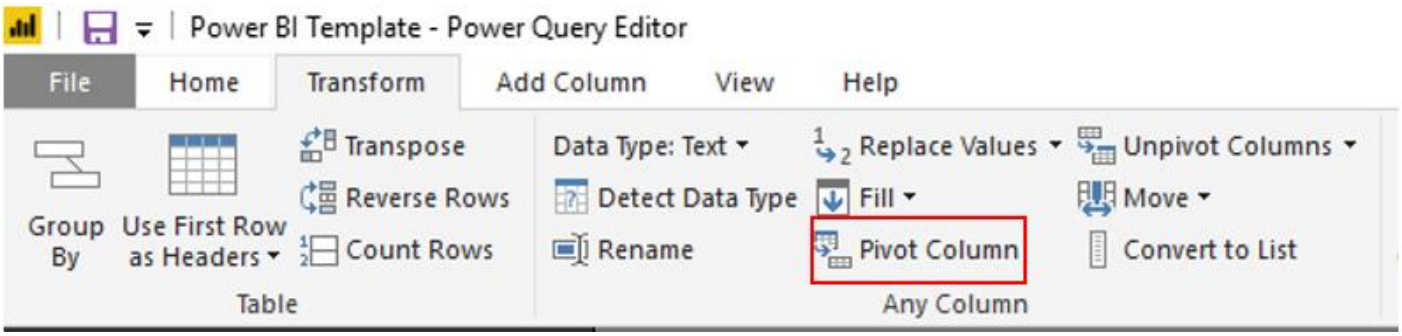
DATA MUNGING & PREPARATION

SESSION 1

The dataset was uploaded onto the PowerBI to initiate the first round of cleansing which included looking for null values, error quantification and missing parameters. Keeping the primary indicators in mind, calculated formulas were added to omit duplicate values and overlapping fields. A cleaned version of the dataset was prepared by the end of the first round.

SESSION 2

The second round was focused on data transformation by pivoting and unpivoting columns to convert them into correct datatypes. This was done with the help of Power Query Editor, by selecting pivot column in the Transform Tab, the datatype was changed by clicking on the left icon on each column header. A clear and valuable view of the dataset with rows being grouped by the years for a particular field was achieved in this round.



SESSION 3

The last session of preparation included renaming of certain fields in a pre-defined naming convention labels for the ease of identification. This round also included segregation of the required fields from dataset, the rest were examined and hidden/omitted if no-correlation was found.