PYTHON FOR DATA ANALYSIS AND DATA SCIENCE

A PROJECT REPORT

SUBMITTED BY:

RAVAL SUGAM JATINBHAI

TRAINNING CUM INTERNSHIP AT TECHNOGEEKS, PUNE

BACHELOR OF ENGINEERING

IN

COMPUTER ENGINEERING DEPARTMENT

SMT.S.R.PATEL ENGINEERING COLLEGE, DABHI

UNJHA , GUJARAT

GUJARAT TECHNOLOGICAL UNIVERSITY

AHMEDABAD 2024

INTRODUCTION OF PROJECT

Topic : Exploring Environmental Sustainability - Data Analysis of Carbon Emission Trends using Python

- Carbon dioxide emissions are the primary driver of global climate change. It’s widely recognised that to avoid the worst impacts of climate change, the world needs to urgently reduce emissions. But, how this responsibility is shared between regions, countries, and individuals has been an endless point of contention in international discussions. This project aims at investigating the distribution of Co2 emission around the world.

Project Objective :

- The project aims to analyze carbon emission trends using Python, exploring factors influencing emissions and identifying regions with high emission levels. By leveraging data analysis techniques, the objective is to gain insights into environmental sustainability and inform strategies for reducing carbon emissions globally.

Project Abstract :

- This project explores global carbon emission trends, employing data analysis techniques with libraries such as Pandas, NumPy, Matplotlib, Seaborn, and Plotly Express.

- By analyzing factors such as population, GDP, and policy interventions, the project aims to understand the drivers of carbon emissions and their impact on environmental sustainability.

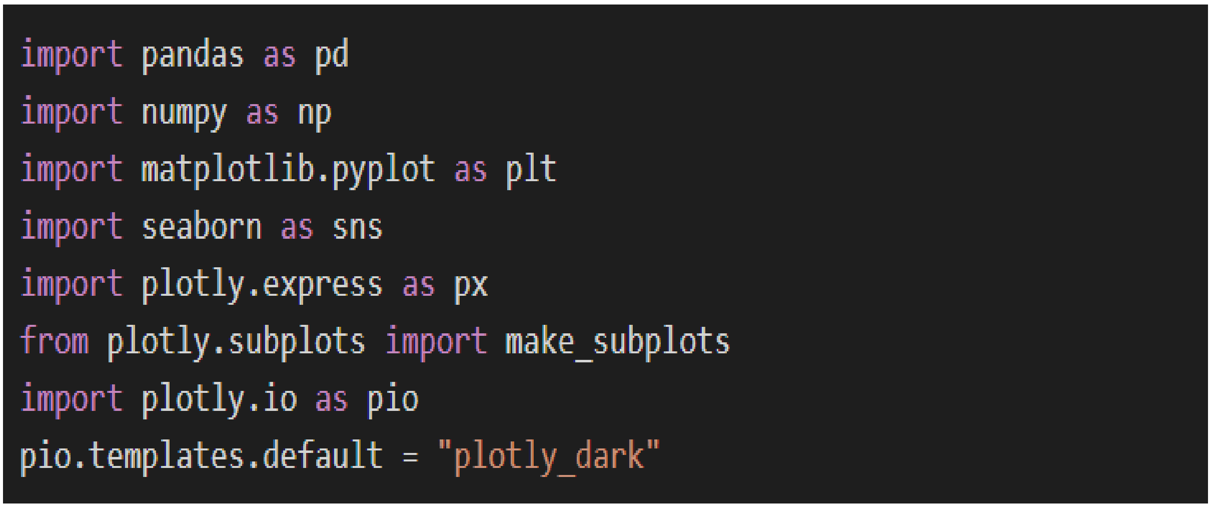
- The utilization of diverse libraries enables a thorough examination of carbon emission data, facilitating informed decision-making towards achieving environmental sustainability goals.

IMPLEMENTATION OF PROJECT

Data Analysis Process :

1) **Asking Questions**
2) **Data Wrangling**
    - **a) Gathering Data**
        - i. CSV files
        - ii. APIs
        - iii. Web Scraping
        - iv. Databases
    - **b) Assessing Data**
    - **c) Cleaning Data**
3) **Exploratory Data Analysis**
4) **Drawing Conclusion**
5) **Communicating Results**

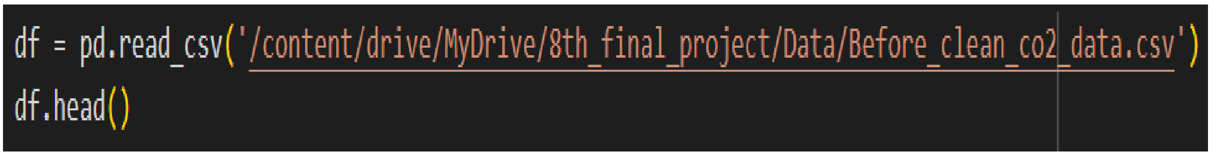
Importing Libraries :



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
from plotly.subplots import make_subplots
import plotly.io as pio
pio.templates.default = "plotly_dark"
```

DATA WRANGLING

➢ Data Gathering :



```
df = pd.read_csv('/content/drive/MyDrive/8th_final_project/Data/Before_clean_co2_data.csv')
df.head()
```

➢ Data Assessing and Cleaning :

- **Shape of dataset** : The dataset contains 46523 rows and 74 columns.
- **About the dataset :**
    - The data consist of various emission indicators around the world from 1750 to 2021.
    - It contains 46000 rows and 74 columns with different parameters. I need only Important columns from those columns.
    - By Observation I found Null values into the dataset.
    - Country columns contains country name and also continents name.
    - This analysis focuses on 2000 to 2021. This analysis will focus on Co2 emission.
- **Important columns description :**
    - **year**: This column represents the year in which the data was recorded. It allows for tracking changes in carbon emissions over time, providing insight into trends and patterns in emissions levels.
    - **iso_code:** This column typically contains the ISO 3166-1 alpha-3 country codes, which are three-letter codes assigned to countries and dependent territories by the International Organization for Standardization (ISO). These codes uniquely identify countries and are useful for aggregating or filtering data by country.

- **country**: This column contains the names of the countries or regions for which carbon emission data is recorded. It provides the context for understanding which geographical areas are contributing to carbon emissions and allows for comparisons between different countries.
- **population:** This column represents the population of each country or region. Population size can influence the level of carbon emissions, as more populous countries tend to have higher overall emissions due to increased energy consumption, transportation needs, and industrial activities.
- **gdp:** GDP stands for Gross Domestic Product, which is a measure of the economic performance of a country or region. It represents the total value of all goods and services produced within the country's borders over a specific period, usually a year or a quarter. GDP is often correlated with carbon emissions, as higher levels of economic activity generally lead to increased energy consumption and carbon emissions.
- **co2:** This column represents the total carbon dioxide emissions (in metric tons) attributed to each country or region. Carbon dioxide is a greenhouse gas that contributes to global warming and climate change. Tracking CO2 emissions is essential for understanding a country's environmental impact and assessing progress toward emission reduction goals.
- **co2_per_capita:** This column represents the carbon dioxide emissions per capita, calculated by dividing the total CO2 emissions by the population of each country or region. It provides insight into the average emissions produced by individuals within a given population, allowing for comparisons of emission levels between countries with different population sizes.

Manual Assessment :

1) Select the necessary columns
2) Deal with missing values
3) Create related columns from existing

> **Numerical (Through Manual Assessment)**

1) gdp_per_capita

2) decade

3) co2_growth_rate

> **Categorical (Through Automatic Assessment)**

1) co2_emission_level

2) development_status

3) policy_status_need

4) Change the data types of columns

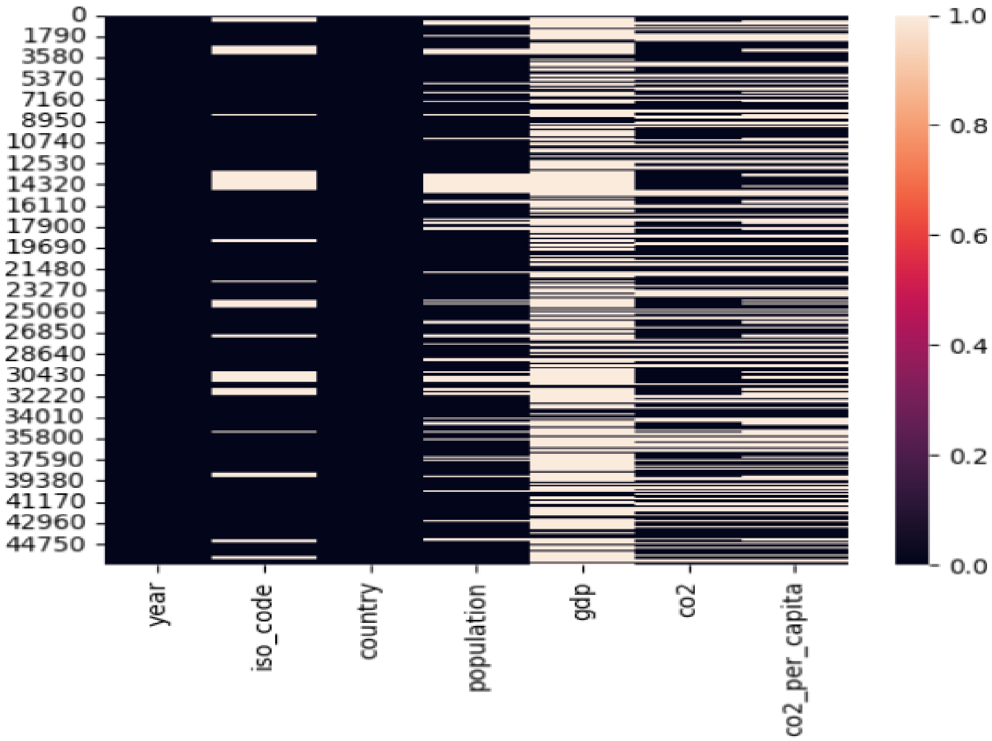
5) Creating the new Data Frame.

1) Select the important time frame (2000-2021)

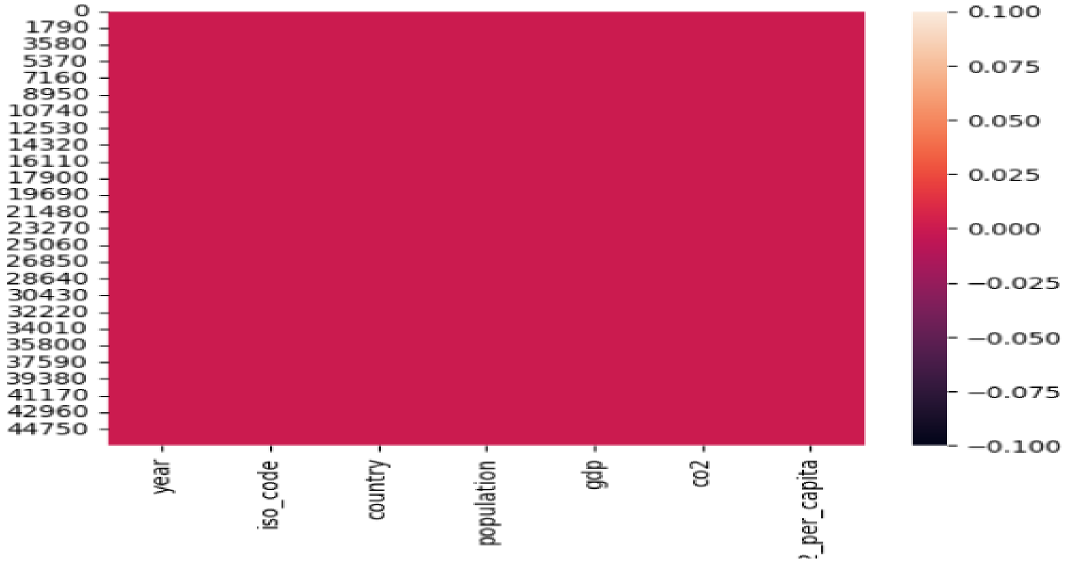
2) Separate dataset for the continents in the dataset (2000-2021)

➢ Dealing with missing values :

○ Before cleaning :



- **After cleaning :**



- Numerical data is clean and fill with interpolate() method and categorical values iso_code is fill with No Data.

➢ Create related columns from existing : (Numerical)

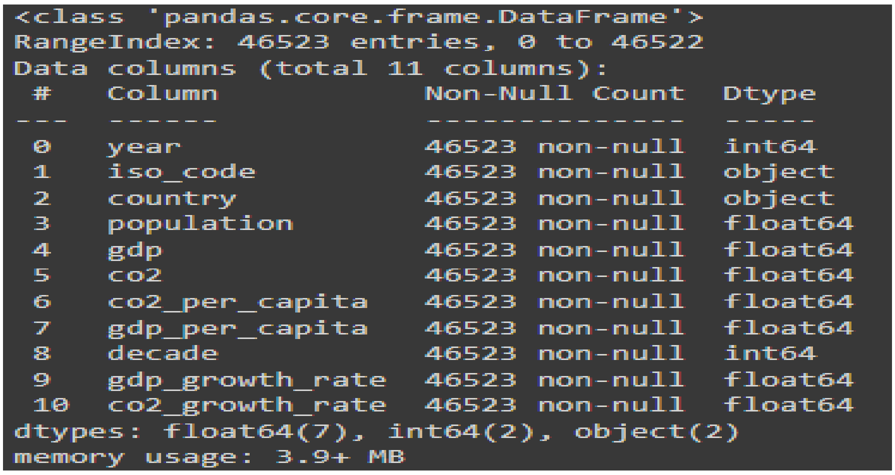
- **Gdp_per_capita :** Depends on gdp and population column
- **Decade :** Depends on year column
- **Gdp_growth_rate :** Depends on gdp columns
- **Co2_growth_rate** Depends on co2 columns

➢ Automatic Assessment :

- **Methods :**

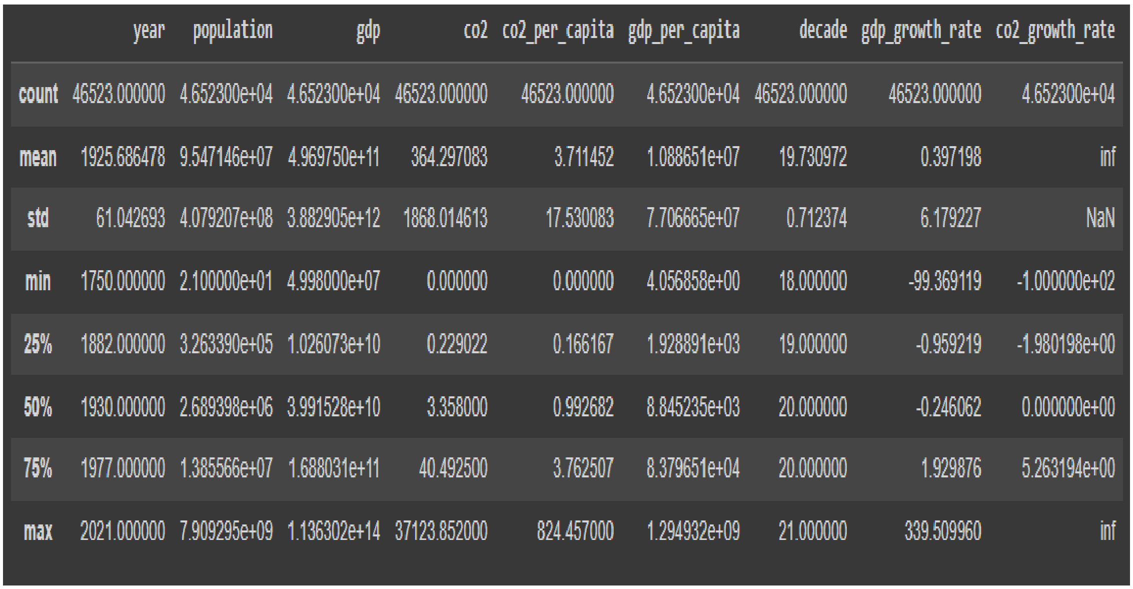
  - head and tail
  - sample
  - info
  - isnull
  - duplicated
  - describe
  - outliers removal

○ Observation through info



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 46523 entries, 0 to 46522
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   year             46523 non-null  int64
 1   iso_code         46523 non-null  object
 2   country          46523 non-null  object
 3   population       46523 non-null  float64
 4   gdp              46523 non-null  float64
 5   co2              46523 non-null  float64
 6   co2_per_capita   46523 non-null  float64
 7   gdp_per_capita   46523 non-null  float64
 8   decade           46523 non-null  int64
 9   gdp_growth_rate  46523 non-null  float64
 10  co2_growth_rate  46523 non-null  float64
dtypes: float64(7), int64(2), object(2)
memory usage: 3.9+ MB
```

○ Describe - Mathematic Summary



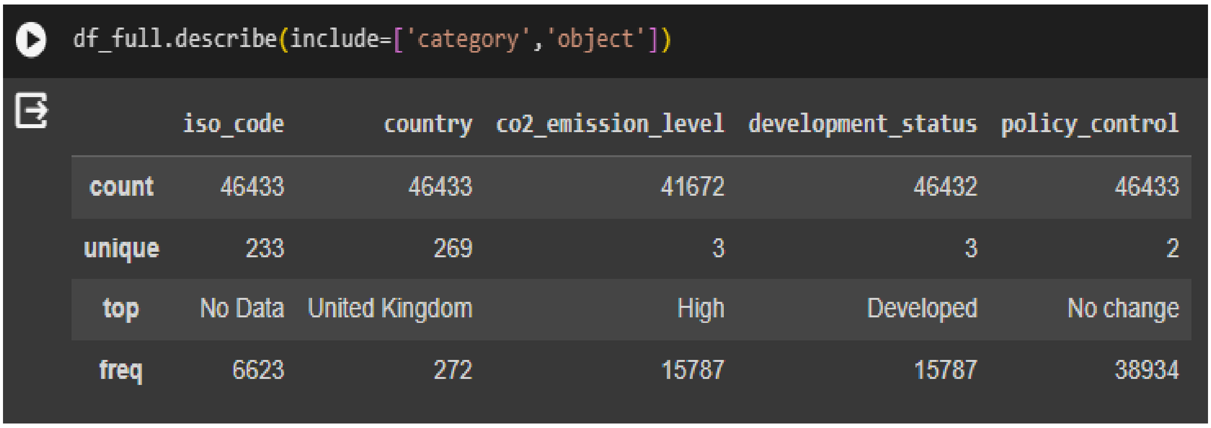
	year	population	gdp	co2	co2_per_capita	gdp_per_capita	decade	gdp_growth_rate	co2_growth_rate
count	46523.000000	4.652300e+04	4.652300e+04	46523.000000	46523.000000	4.652300e+04	46523.000000	46523.000000	4.652300e+04
mean	1925.686478	9.547146e+07	4.969750e+11	364.297083	3.711452	1.088651e+07	19.730972	0.397198	inf
std	61.042693	4.079207e+08	3.882905e+12	1868.014613	17.530083	7.706665e+07	0.712374	6.179227	NaN
min	1750.000000	2.100000e+01	4.998000e+07	0.000000	0.000000	4.056858e+00	18.000000	-99.369119	-1.000000e+02
25%	1882.000000	3.263390e+05	1.026073e+10	0.229022	0.166167	1.928891e+03	19.000000	-0.959219	-1.980198e+00
50%	1930.000000	2.689398e+06	3.991528e+10	3.358000	0.992682	8.845235e+03	20.000000	-0.246062	0.000000e+00
75%	1977.000000	1.385566e+07	1.688031e+11	40.492500	3.762507	8.379651e+04	20.000000	1.929876	5.263194e+00
max	2021.000000	7.909295e+09	1.136302e+14	37123.852000	824.457000	1.294932e+09	21.000000	339.509960	inf

- **population** - continuous increase
- **gdp** - continuous increase
- **co2** - sudden max (need to observe) - through tail
- **co2_per_capita** - sudden max (need to observe) - through tail
- **gdp_per_capita** - sudden max (need to observe) - through tail
- **gdp_growth_rate** - sudden max (need to observe) - through tail
- **co2_growth_rate** - sudden max (need to observe) - through tail

➢ Create related columns from existing : (Categorical)

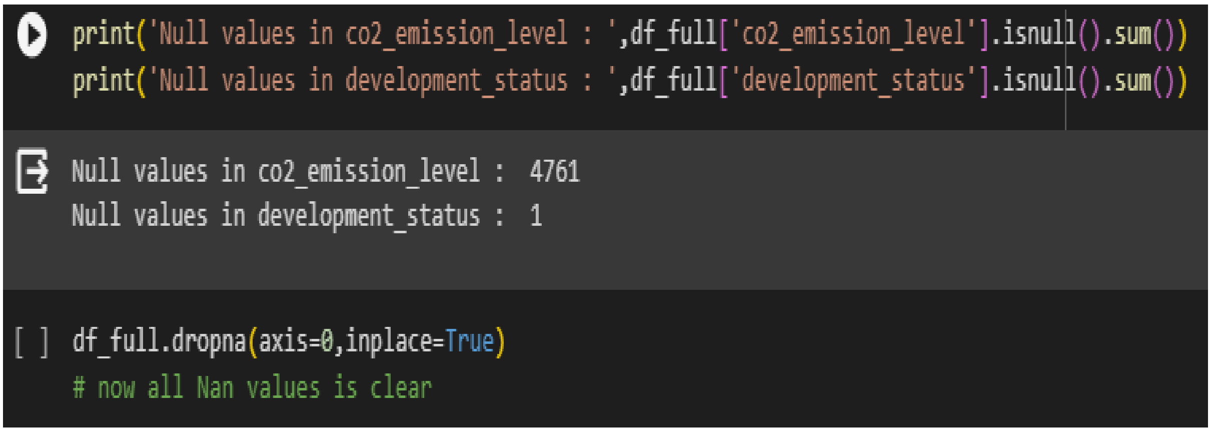
- **Co2_emission_level** : Depends on co2 quantile()
  - **Low** : min < x < 0.33
  - **Medium** : 0.33 < x < 0.66
  - **High** : 0.66 < x < max
- **Development_status**: Depends on gdp_per_capita quantile()
  - **Underdeveloped** : min < x < 0.33
  - **Developing** : 0.33 < x < 0.66
  - **Developed** : 0.66 < x < max
- **Policy_control** : Depends on gdp and co2 quantile()
  - **No Change**: [~ (co2 > thersh_co2) and (gdp < thresh_gdp)]
  - **Improvement** : (co2 > thersh_co2) and (gdp < thresh_gdp)

➢ Describe - Categorical Summary



```
df_full.describe(include=['category','object'])
```

	iso_code	country	co2_emission_level	development_status	policy_control
count	46433	46433	41672	46432	46433
unique	233	269	3	3	2
top	No Data	United Kingdom	High	Developed	No change
freq	6623	272	15787	15787	38934

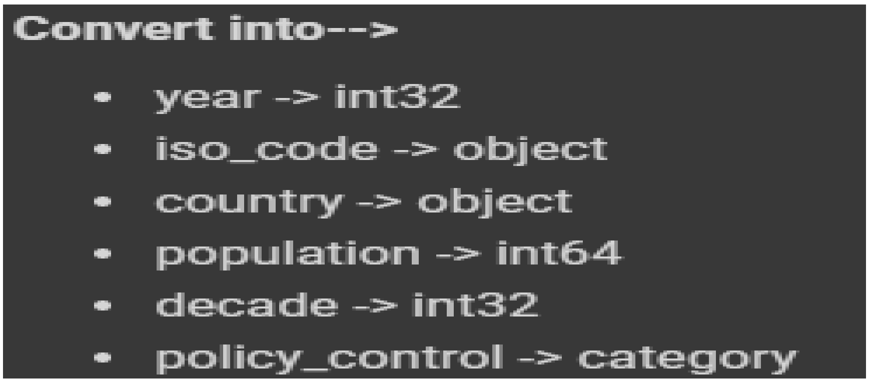


```
print('Null values in co2_emission_level : ',df_full['co2_emission_level'].isnull().sum())
print('Null values in development_status : ',df_full['development_status'].isnull().sum())
```

```
Null values in co2_emission_level :  4761
Null values in development_status :  1
```

```
[ ] df_full.dropna(axis=0,inplace=True)
    # now all Nan values is clear
```

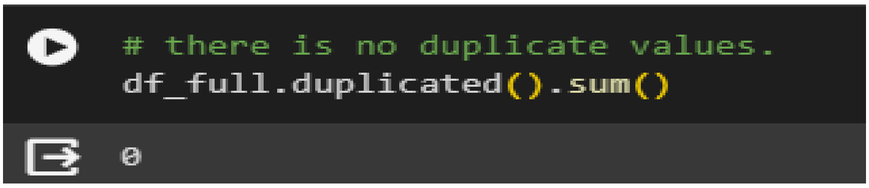
➢ Change the data types of the columns:



Convert into-->

- year -> int32
- iso_code -> object
- country -> object
- population -> int64
- decade -> int32
- policy_control -> category

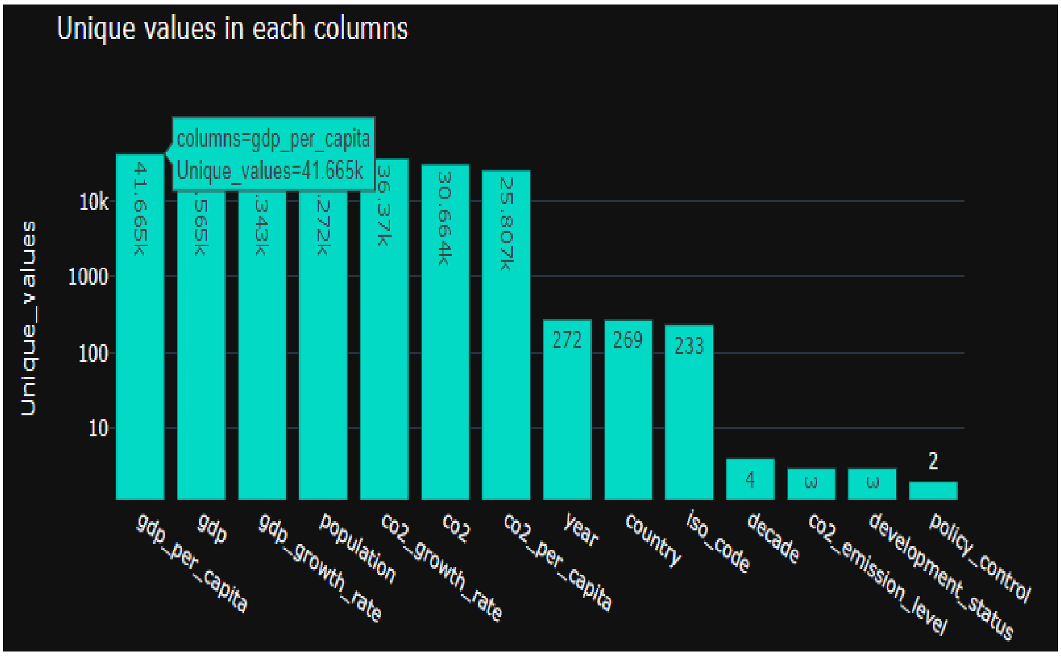
➢ Check duplicates values:



```
# there is no duplicate values.
df_full.duplicated().sum()
```

```
0
```

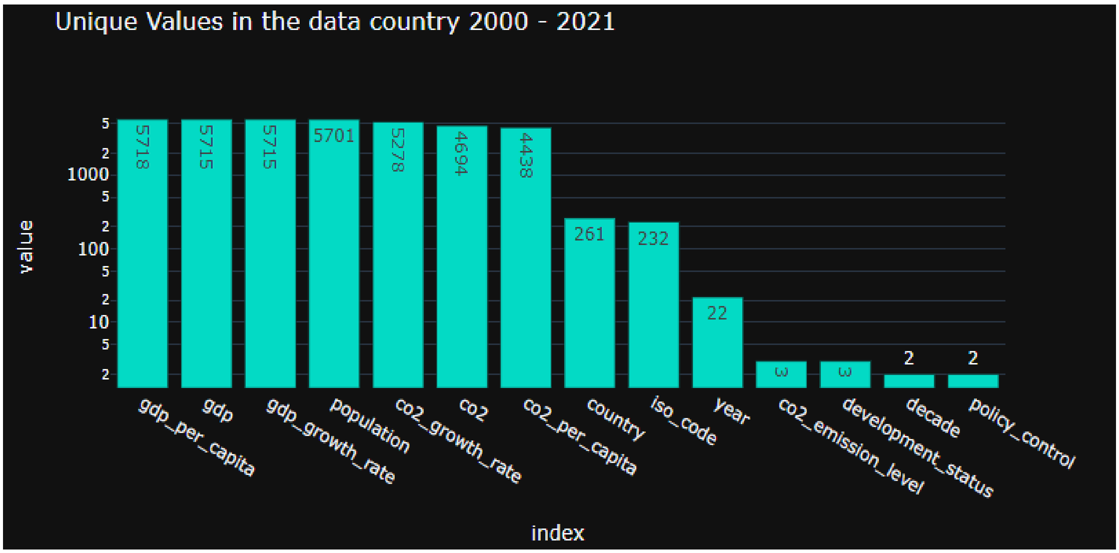
➢ Unique values in each columns :



- ○ Here, **country- 233** and **iso_code- 269** so create discrepancy between **country** and **iso_code**, it means there is **No Data** valuefor country in **iso_code**.

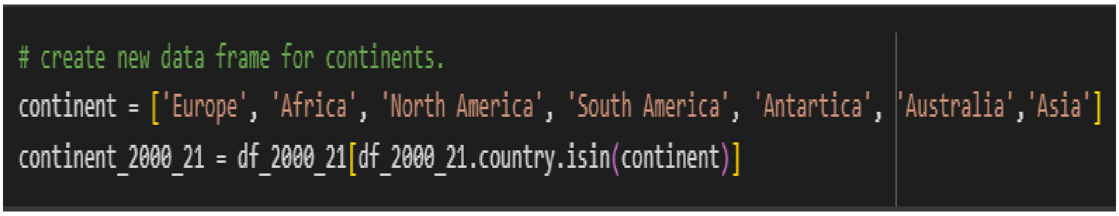
➢ Create separate Data frame :

- **Select the important time frame(2000-2021) – df_2000_21**
- **Unique values :**



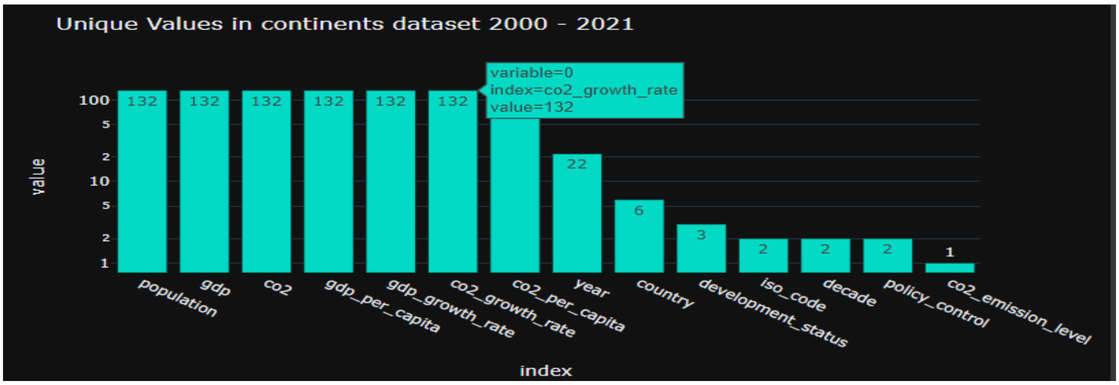
- The original data frame can now be adjusted to include only individual countries (countries with iso_code) means take only those data which must have iso_code as **country_2000_21**.

- **Separate dataset for the continents from the dataset(2000-2021) – continent_2000_21 from df_2000_21**



```
# create new data frame for continents.
continent = ['Europe', 'Africa', 'North America', 'South America', 'Antartica', 'Australia','Asia']
continent_2000_21 = df_2000_21[df_2000_21.country.isin(continent)]
```

- **Unique values :**



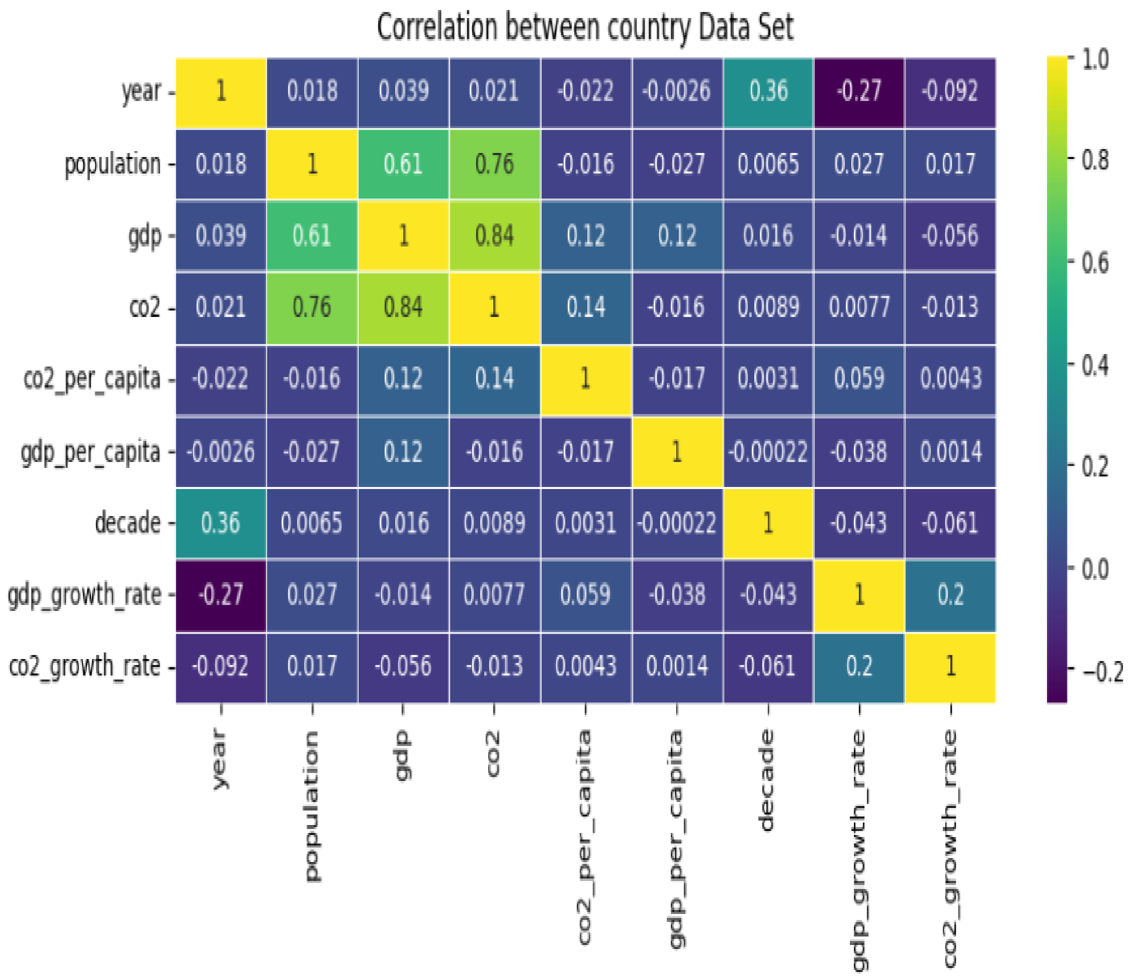
➢ Descriptive statistics on the data :

▪ Country_2000_21 :

- There's too much controversy about the territory of Western Sahara (ESH) country.
- Create data frame which have no country with Western Sahara as **country_2000_21_wo_ESH** data frame.

▪ Correlation between numerical columns :

(country_2000_21_wo_ESH)

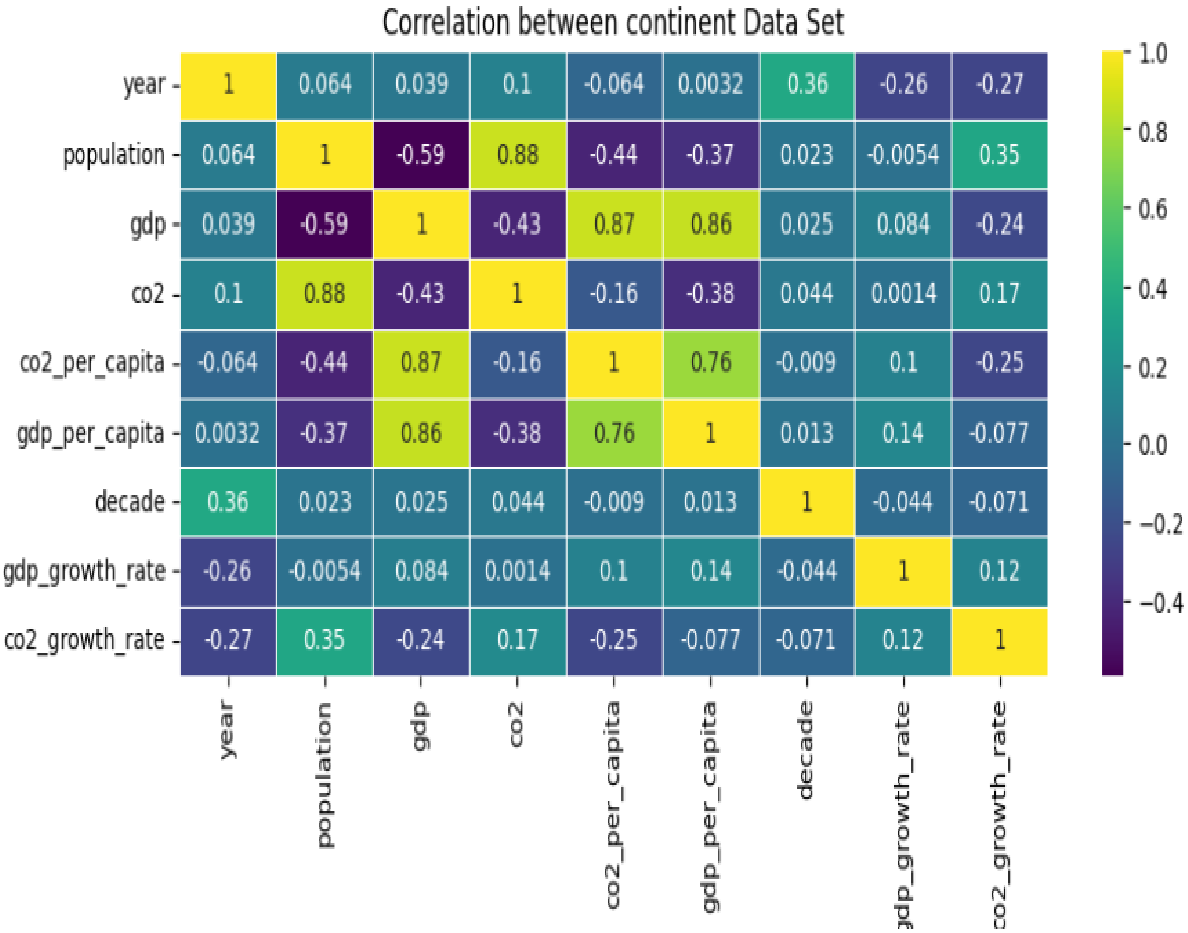


Correlation between country Data Set

	year	population	gdp	co2	co2_per_capita	gdp_per_capita	decade	gdp_growth_rate	co2_growth_rate
year	1	0.018	0.039	0.021	-0.022	-0.0026	0.36	-0.27	-0.092
population	0.018	1	0.61	0.76	-0.016	-0.027	0.0065	0.027	0.017
gdp	0.039	0.61	1	0.84	0.12	0.12	0.016	-0.014	-0.056
co2	0.021	0.76	0.84	1	0.14	-0.016	0.0089	0.0077	-0.013
co2_per_capita	-0.022	-0.016	0.12	0.14	1	-0.017	0.0031	0.059	0.0043
gdp_per_capita	-0.0026	-0.027	0.12	-0.016	-0.017	1	-0.00022	-0.038	0.0014
decade	0.36	0.0065	0.016	0.0089	0.0031	-0.00022	1	-0.043	-0.061
gdp_growth_rate	-0.27	0.027	-0.014	0.0077	0.059	-0.038	-0.043	1	0.2
co2_growth_rate	-0.092	0.017	-0.056	-0.013	0.0043	0.0014	-0.061	0.2	1

- **Correlation between numerical columns :**

(continent_2000_21_wo_ESH)

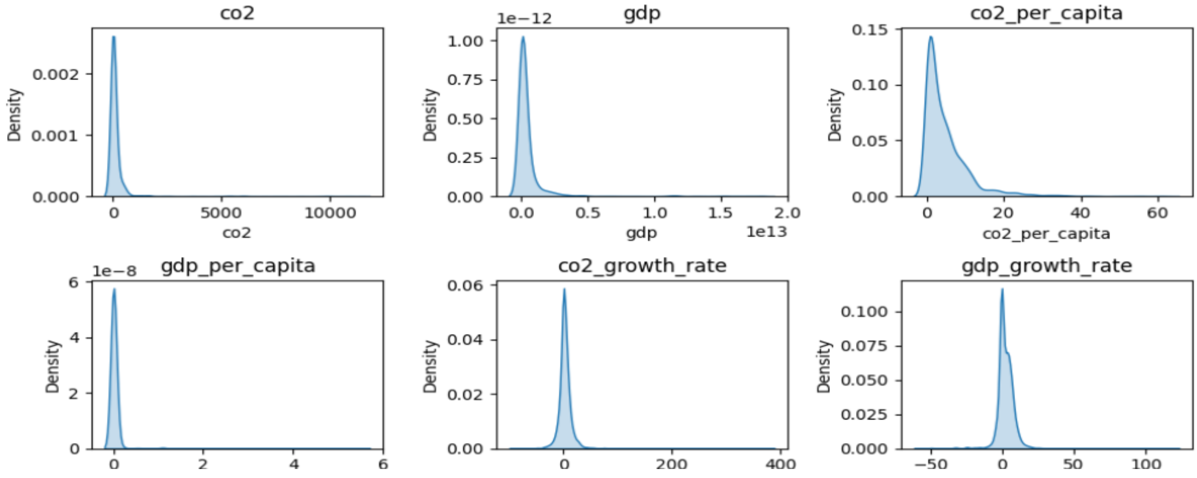


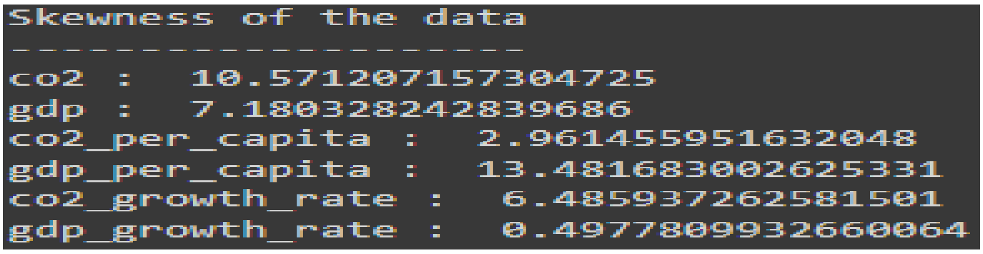
EXPLORATORY DATA ANALYSIS

- **Numeric data :** year, population, gdp, co2, co2_per_capita, gdp_per_capita, decade, gdp_growth_rate, co2_growth_rate
- **Categorical data :** iso_code, country, co2_emission_level, development_status, policy_control
- **Mixed data :** No data available

➢ Univariate Analysis

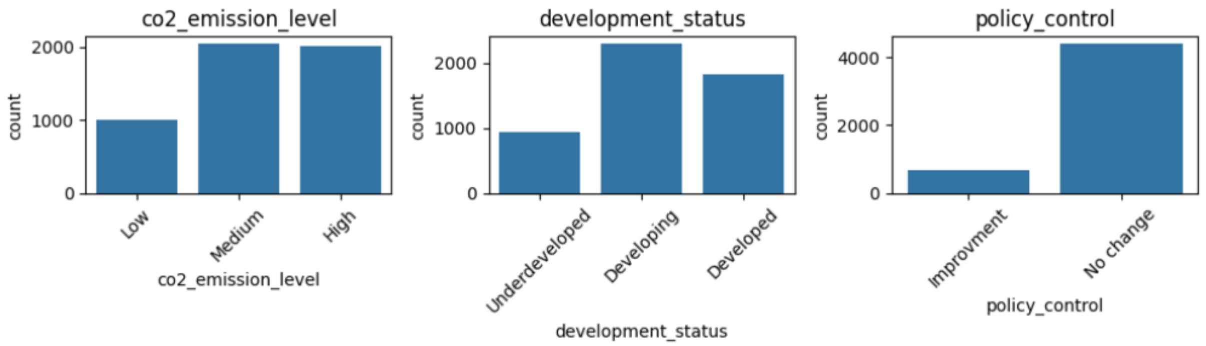
- ❖ **Numerical : country_2000_21_wo_ESH**
- ❖ **Distribution analysis:** The distribution of each feature is examined to identify its shape, central tendency, and dispersion.
  - **Shape :**
    - Normal , Skewed, Bimodel, Uniform Distribution
  - **Central tendency:**
    - Mean, Median, Mode
  - **Dispersion:**
    - Range, Var, Std, IQR

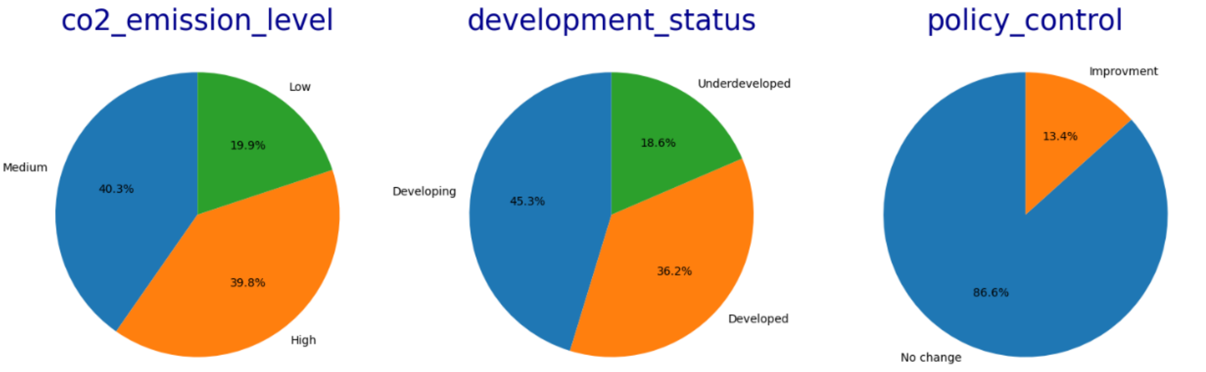




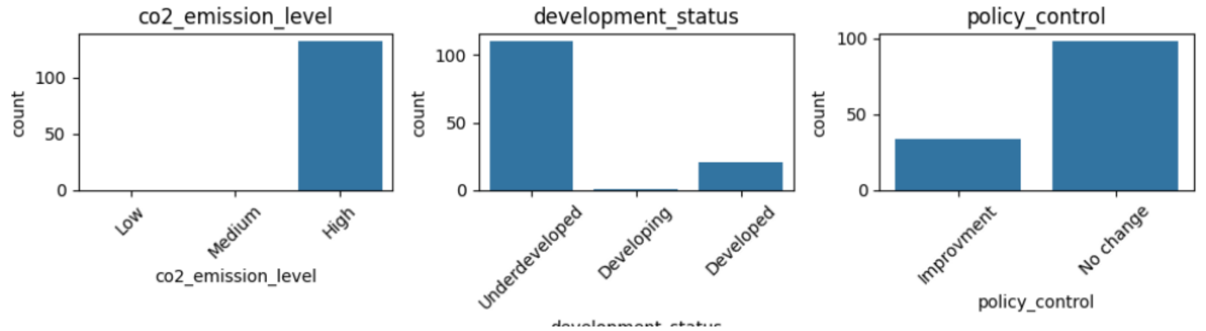
```
Skewness of the data
---------------------
co2  :   10.571207157304725
gdp  :   7.180328242839686
co2_per_capita  :   2.961455951632048
gdp_per_capita  :   13.481683002625331
co2_growth_rate  :   6.485937262581501
gdp_growth_rate  :   0.4977809932660064
```

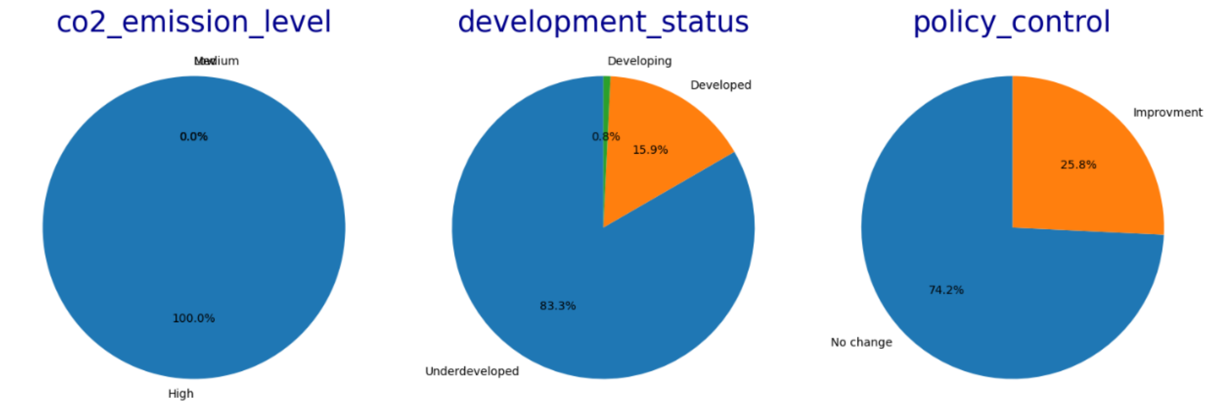
❖ Categorical : country_2000_21_wo_ESH





❖ Categorical: continent_2000_21

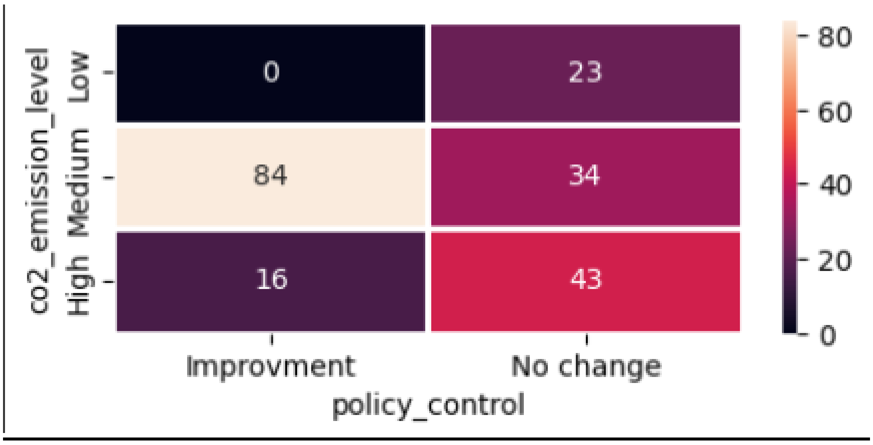




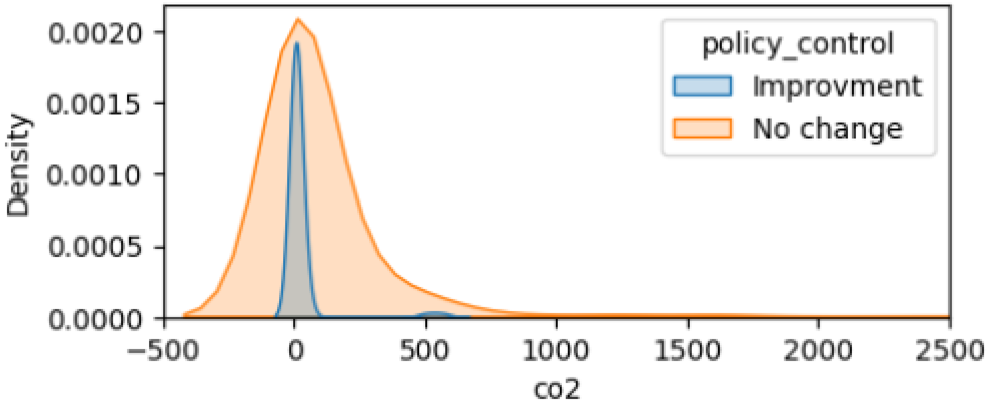
➢ Bivariate Analysis

❖ Country_2000_21_wo_ESH

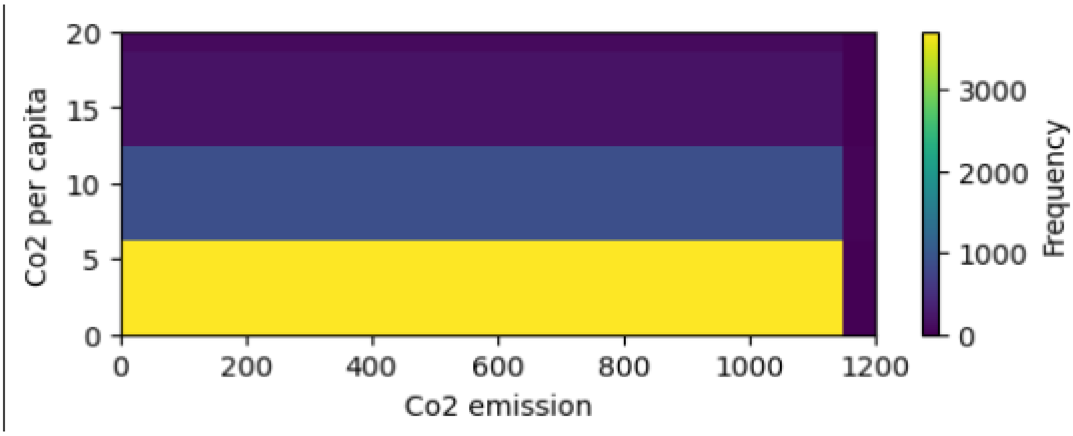
- **CATEGORY – CATEGORY :**



- **NUMERIC – CATEGORY :**

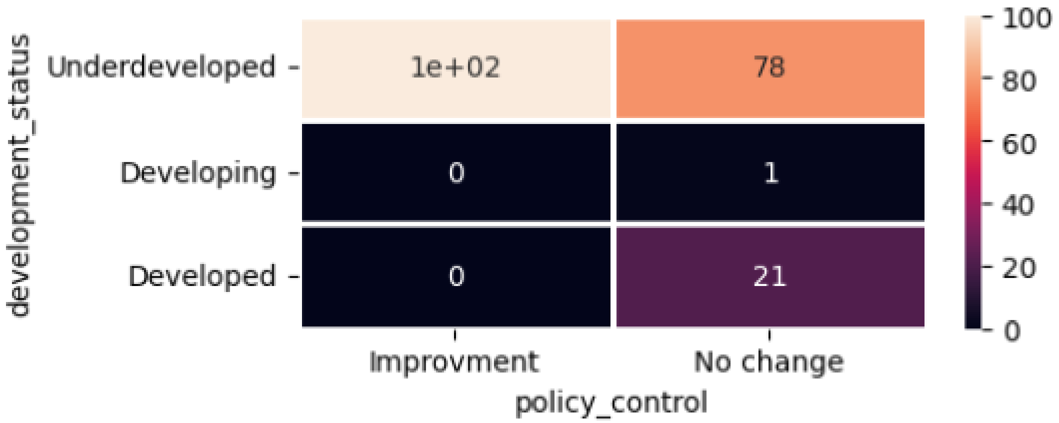


- **NUMERIC – NUMERIC :**

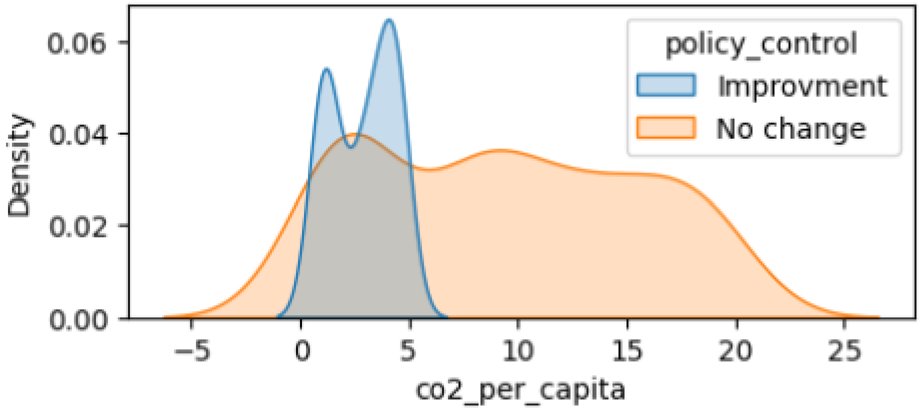


❖ CONTINENT_2000_21

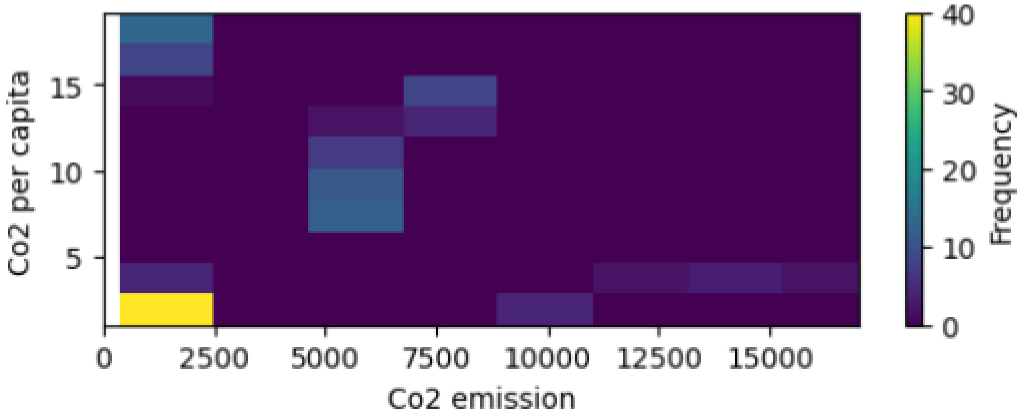
○ CATEGORY – CATEGORY :



○ NUMERIC – CATEGORY :

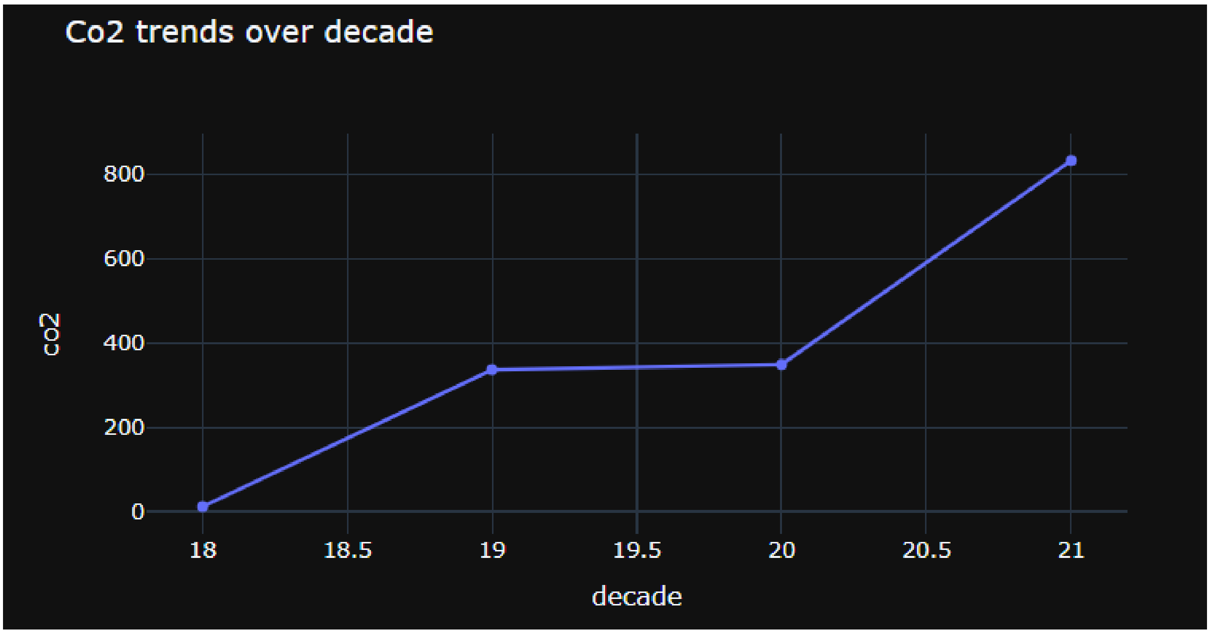


○ NUMERIC – NUMERIC:



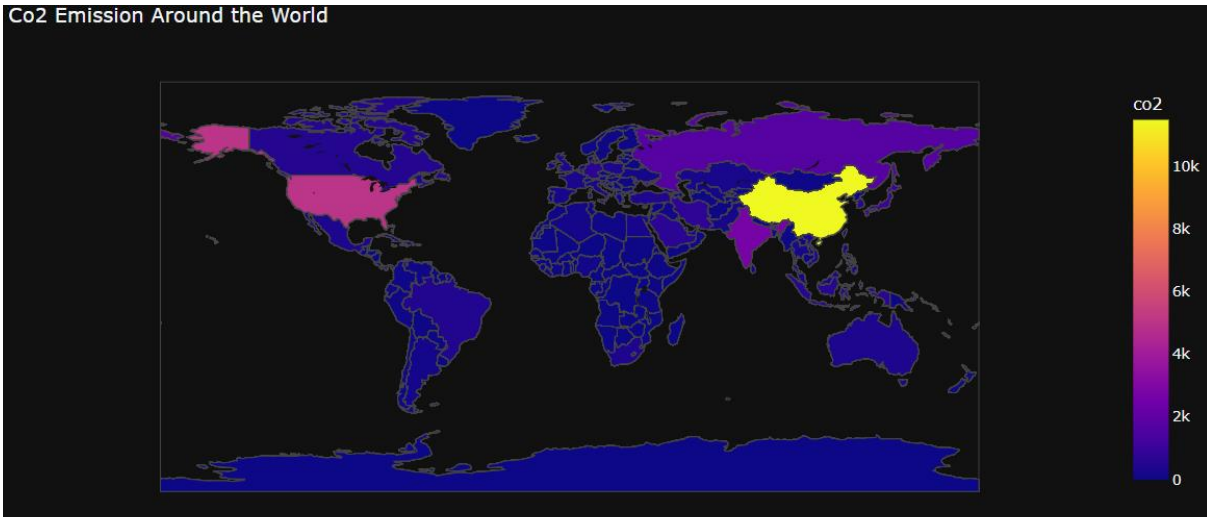
➢ Use Cases :

1) Trend of Co2 Emission over decades



Insight 1: Co2 emissions have increased over the decade due to the rise in population and corresponding increase in consumption and industrial activities.

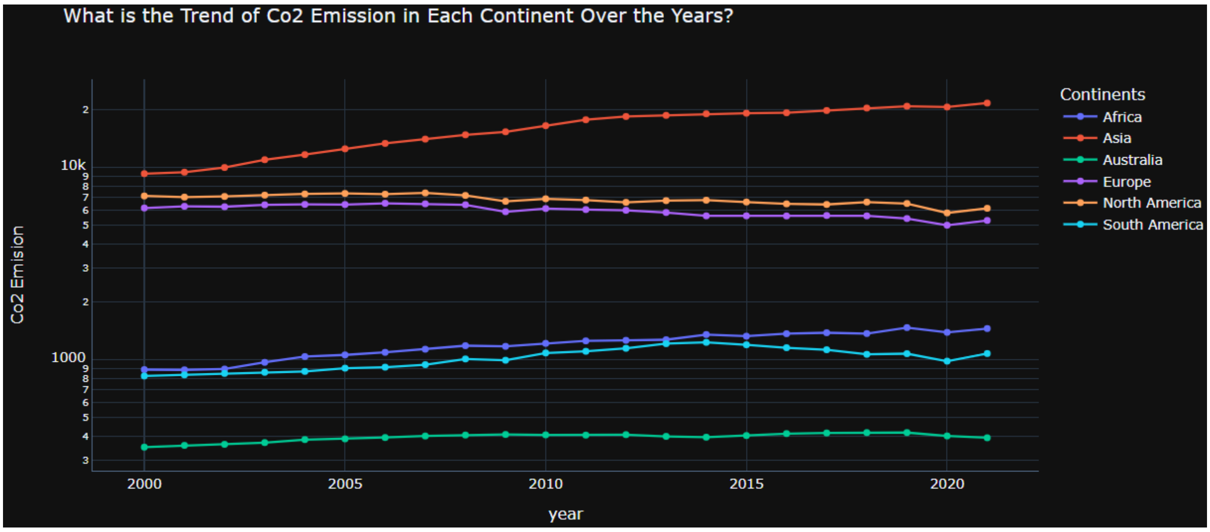
2) What is the Distribution and Trend of Co2 Emission Around the World?



Insight 2: - China - 1st co2 emitter

- United State - 2nd co2 emitter
- India & Russia - 3rd co2 emitter
- Africa & South America - Mid level emitter

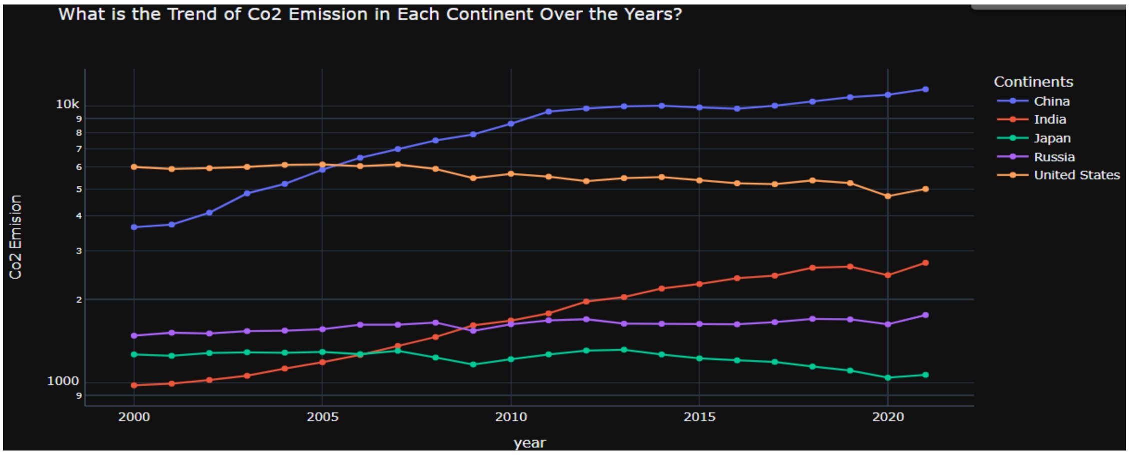
3) What is the Trend of Co2 Emission in Each Continent Over the Years?



Insight 3: Trend over co2 over 21th century.

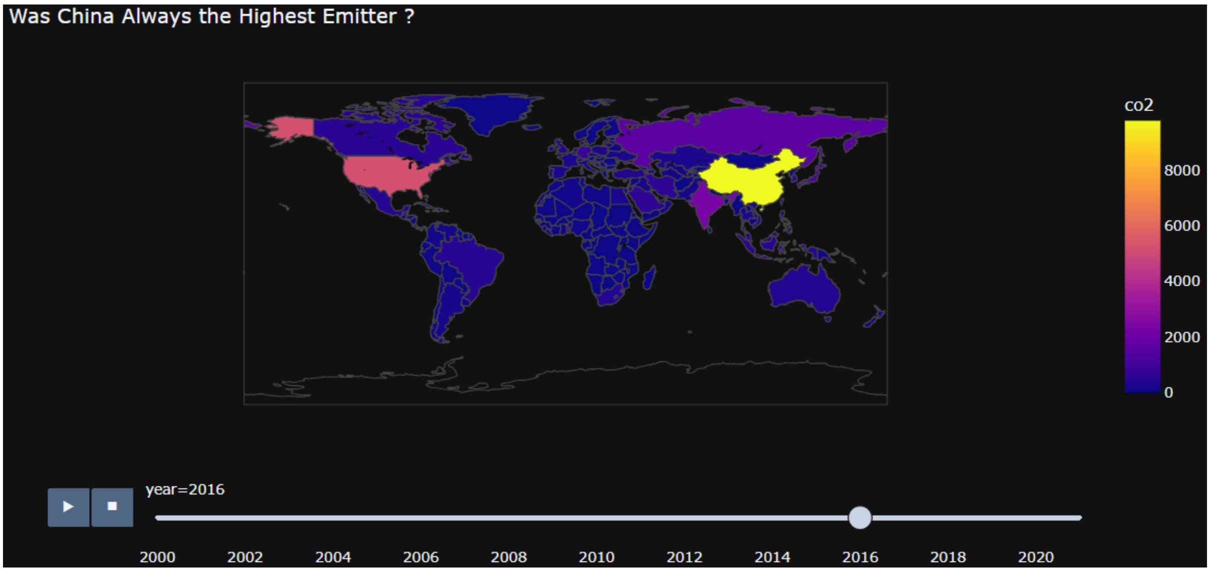
- **Slightly Upward** : Africa & South America
- **Highly Upward** : Asia
- **Downward** : Europe & North America
- **Constant** - Australia

4) co2 trend over year for top 5 co2 emission country.



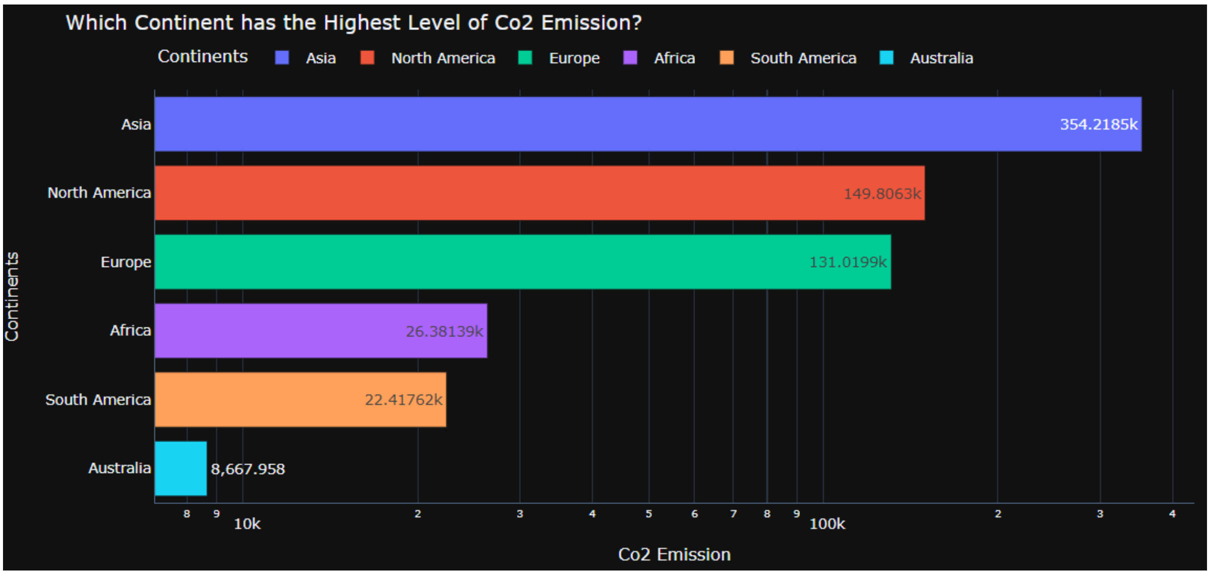
Insight 4: China and India have seen exponential increases in CO2 emissions over the centuries, reflecting their rapid industrialization and economic growth.

5) Was China Always the Highest Emitter?



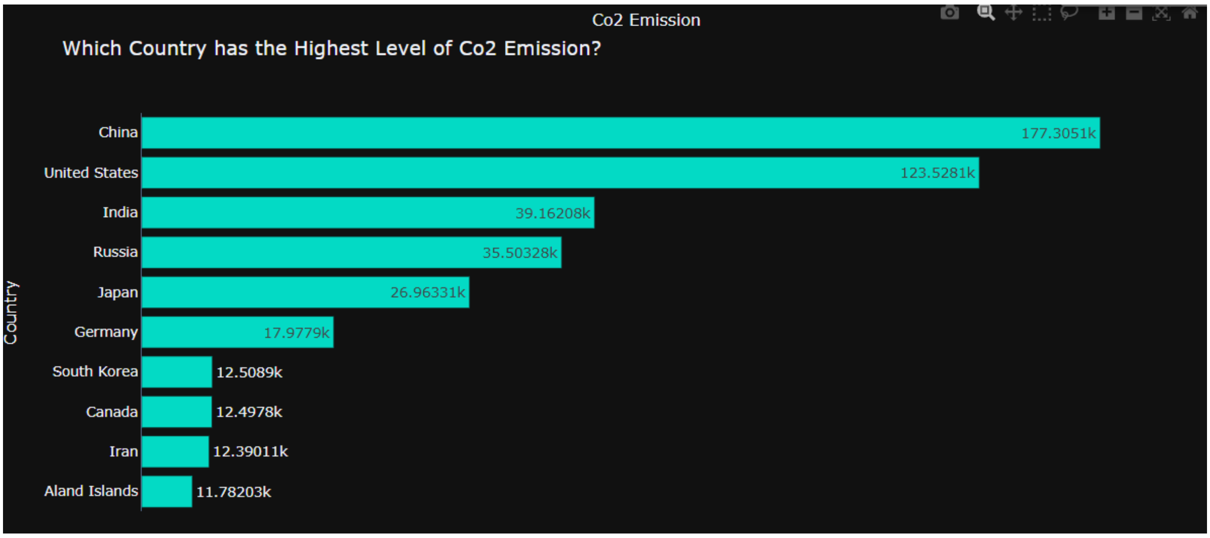
Insight 5: China consistently ranks as the highest CO2 emitter throughout the 21st century, indicating its significant contribution to global emissions.

6) Which Continent has the Highest Level of Co2 Emission?



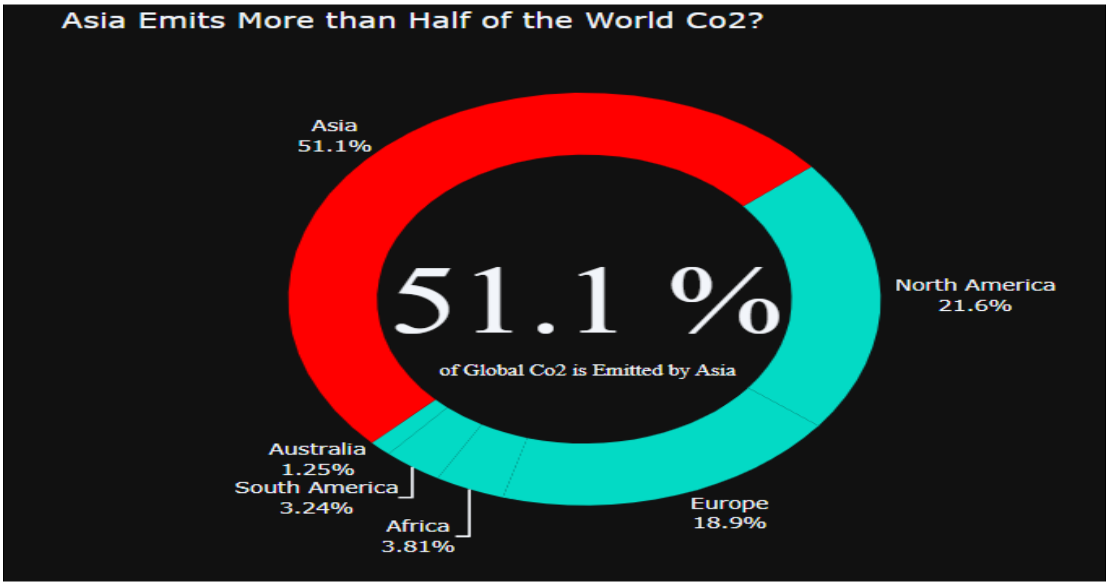
Insight 6: Asia has the highest Carbon Emission.

7) Which country has the Highest Level of Co2 Emission?



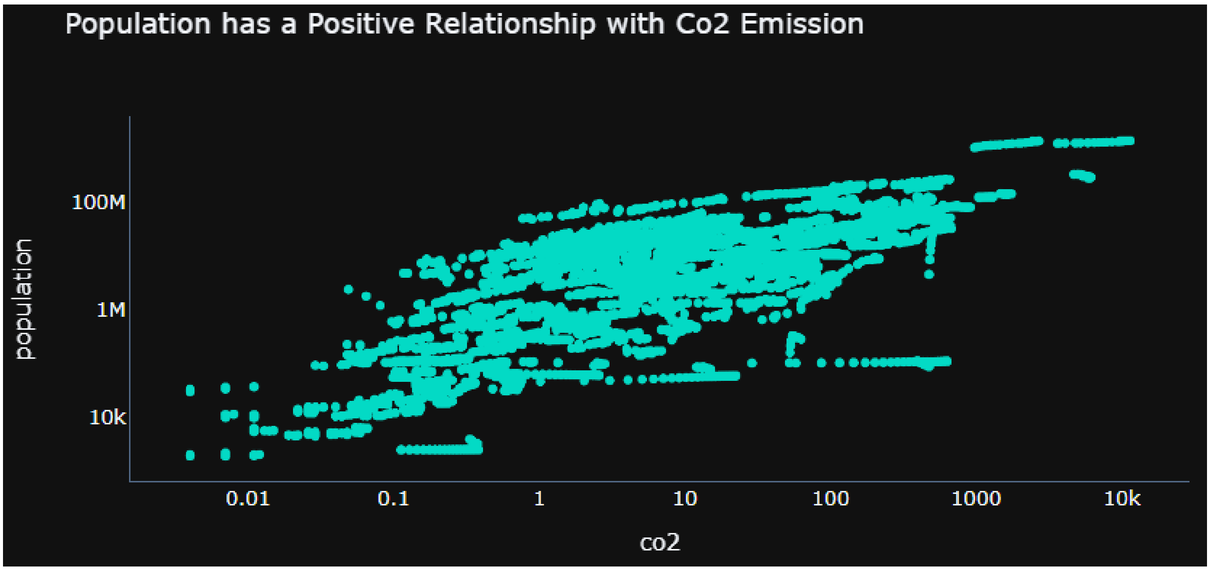
Insight 7: China ha the highest level of Co2 emission. This is in line with what we discovered in the first choropleth map. It is also in line with what we saw at the continental level as six of the top ten countries are Asian.

8) Asia Emits More than Half of the World Co2 ?



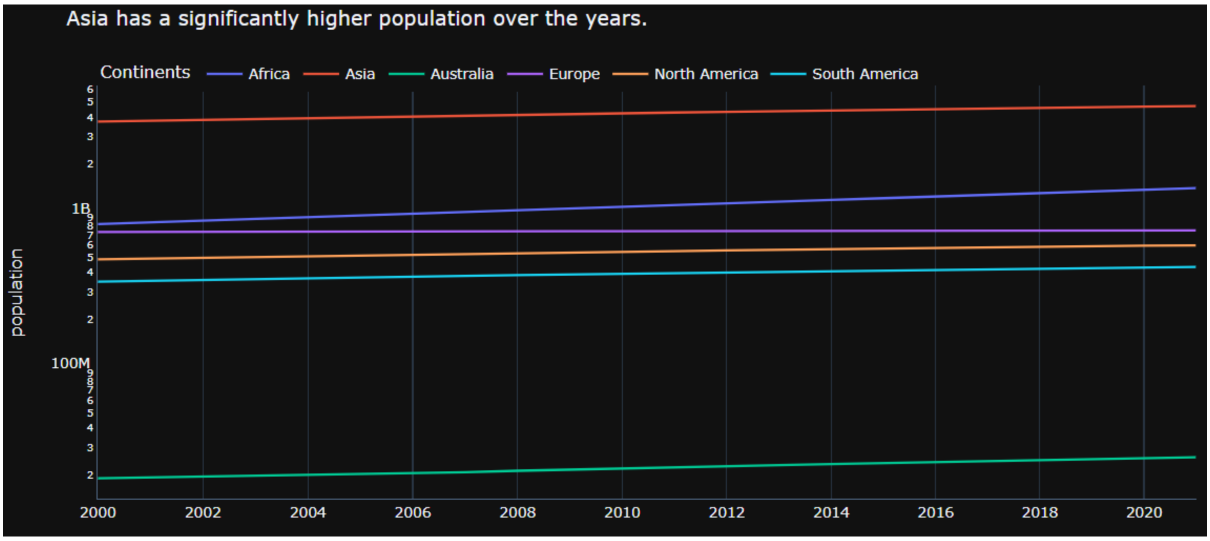
<u>Insight 8:</u> More than half of the world's CO2 emissions originate from Asia, highlighting its critical role in global emissions. That is 51.1% of global co2 emission is comes from Asia.

9) Population has a Positive Relationship with Co2 Emission or Not ?



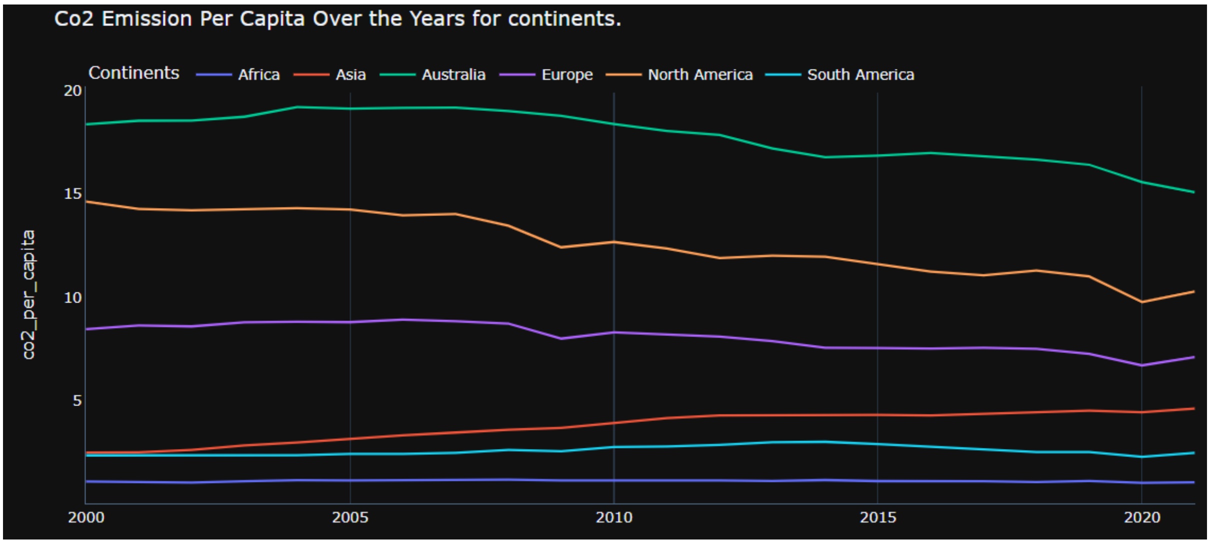
<u>Insight 9:</u> Population has also been seen as a determinant of Co2 emission as the more people mean more demand for pollution product.

10) Asia has a significantly higher population over the years.



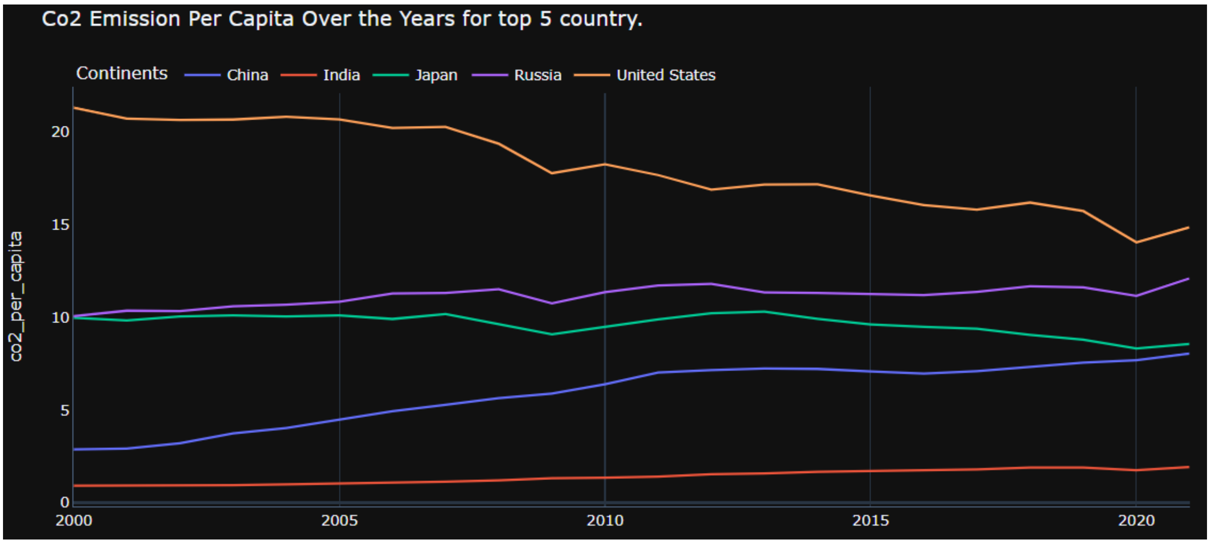
<u>Insight 10:</u> Let's explore if there truly is any relationship between population and Co2 emission. The map above shows that the Asian countries have a large population compared to the other continents. Therefore, we have to account for population.

11) Co2 Emission Per Capita Over the Years for continents.



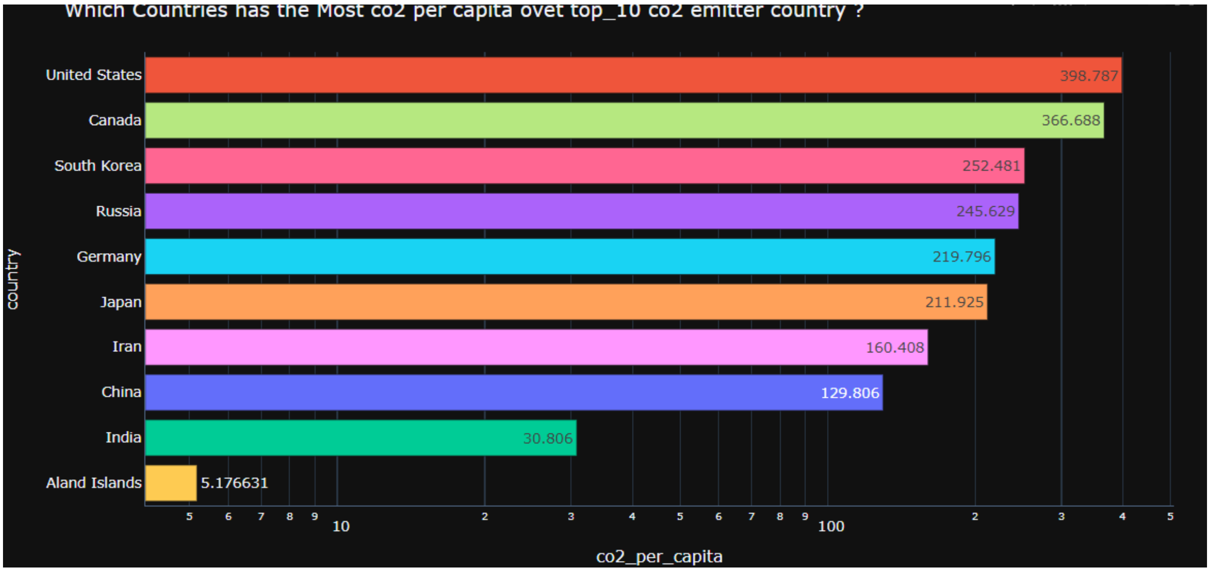
<u>Insight 11:</u> Australia exhibits the highest CO2 emissions per capita, while Asia has the third-lowest emission level per capita.

12) Co2 Emission Per Capita Over the Years for top 5 country with co2 emission.



Insight 12: US with highest Co2_per capita but over the time it is decrease and china is increase fast.

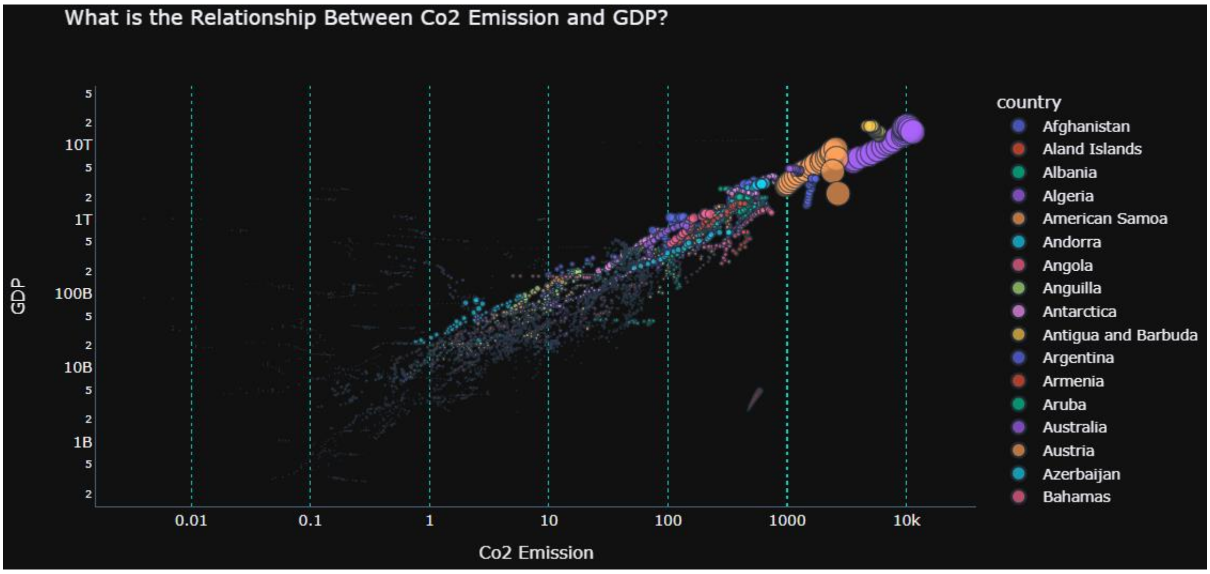
13) Which Countries has the Most co2 per capita over top_10 co2 emitter country?



Insight 13: - **United States** - 1st co2 per capita

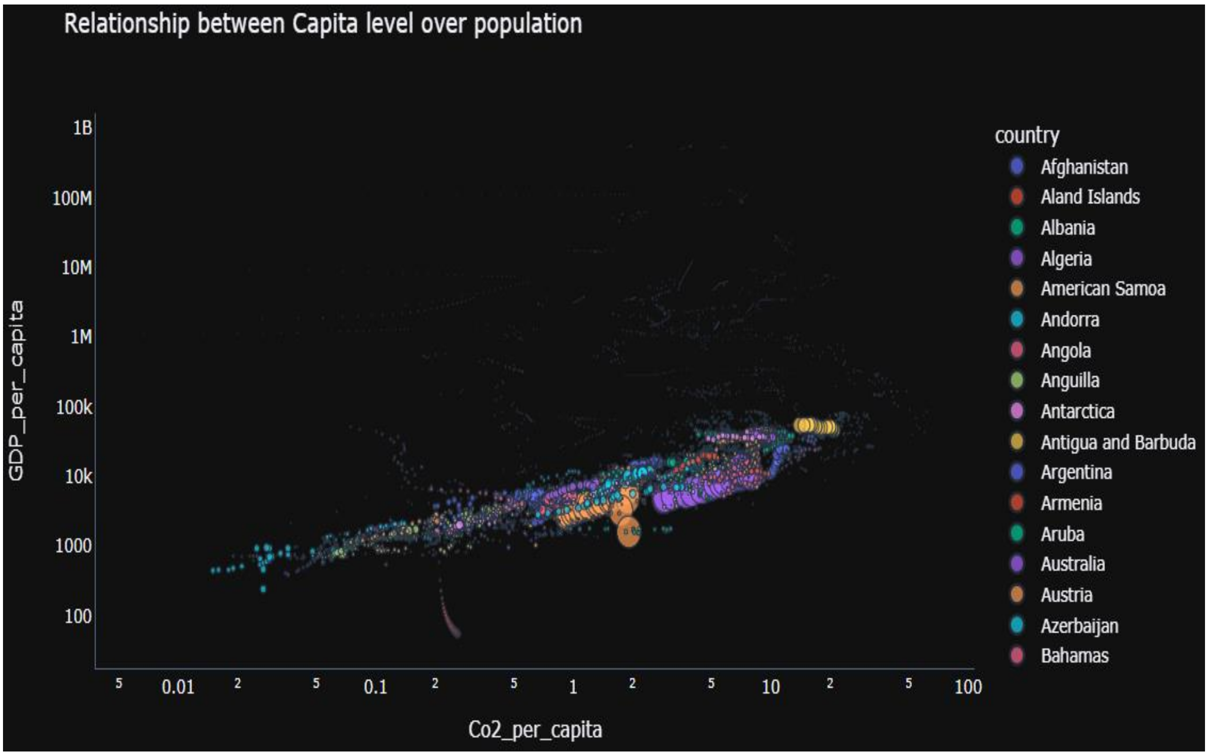
- **Canada** - 2nd co2 per capita
- **South Korea** - 3rd co2 per capita

14) What is the Relationship Between Co2 Emission and GDP over population ?



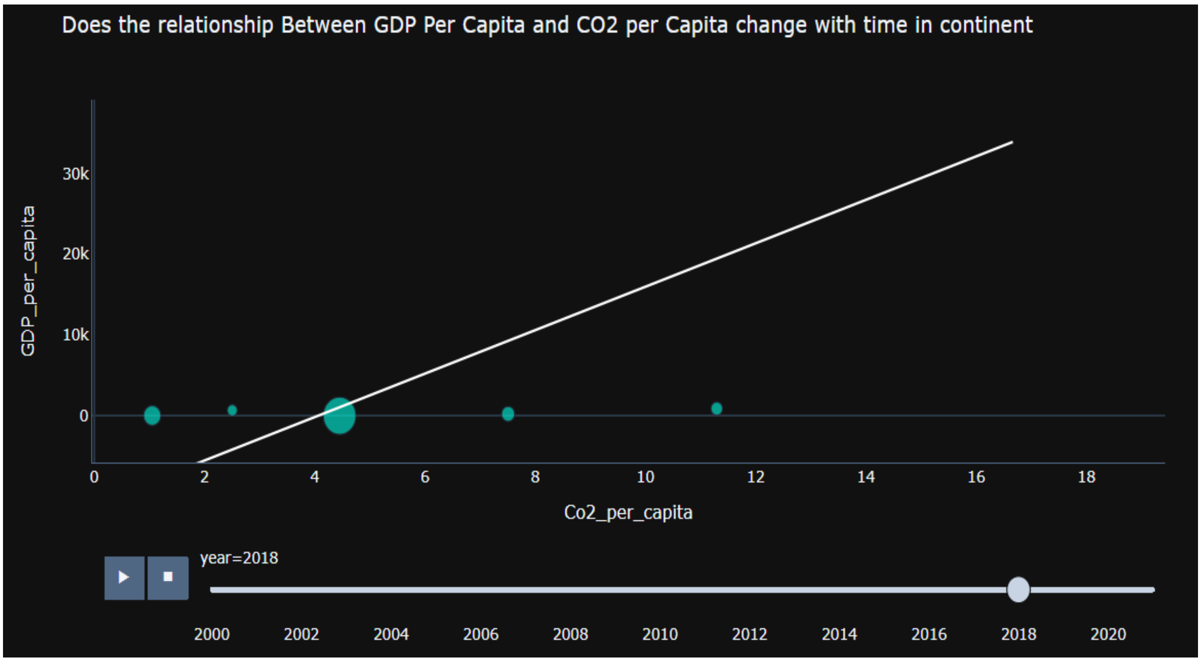
Insight 14: There is a positive relationship between GDP and Co2 Emission per Capita. Looking at the graph, Co2 and GDP increases as population increases. The more populated countries are produces high level of Co2 Emission and have a high GDP level.

15) Relationship between Capita level.



Insight 15: Wealthier countries tend to produce more carbon dioxide (CO2) per person. In the past, it was believed that richer countries polluted less, but recent findings show the opposite. On the other hand, countries with moderate levels of wealth emit CO2 levels closer to the average, even though they may have larger populations.

16) Does the relationship Between GDP Per Capita and CO2 per Capita change with time in continents?



Insight 16: The relationship between GDP per capita and Co2 per capita has remained positive for the whole of the 21st century.

CONCLUSION

- In summary, the analysis of carbon emission trends reveals critical insights for addressing climate change and promoting environmental sustainability. The observed increase in CO2 emissions underscores the urgency of implementing proactive measures to curb greenhouse gas emissions globally.

- China's status as the top global CO2 emitter highlights the need for targeted interventions in rapidly industrializing economies. Varying emission trends across continents emphasize the necessity of tailored strategies for each region. The exponential growth of emissions in China and India underscores the imperative for transitioning to renewable energy sources. Asia's prominence as the leading emitter underscores the region's pivotal role in global climate action.

- The positive correlation between GDP and CO2 emissions challenges conventional assumptions about pollution levels in wealthier nations. Population size emerges as a significant determinant of emissions, necessitating sustainable urban planning strategies. Overall, urgent and collaborative efforts are essential to mitigate climate change, transition to low-carbon economies, and secure a sustainable future.

REFERENCES

1) pandas
2) Numpy
3) Matplotlib
4) Seaborn
5) Plotly
6) Kaggle