# EFFECTS OF AVIATION ON ENVIRONMENT

BITS ZG628T: Dissertation

by

**Ankitha Yamini Bathinozu**

**2022MT12088**

Dissertation work carried out at

**Honeywell Technology Solutions, Hyderabad**

Submitted in partial fulfilment of **M. Tech Software Systems** degree programme

Under the Supervision of

**Chiruvolu, Karthikeya**

**Honeywell Technology Solutions, Hyderabad**

**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE**

**PILANI (RAJASTHAN)**

**April 2024**

## *CERTIFICATE*

This is to certify that the Dissertation entitled **'EFFECTS OF AVIATION ON ENVIRONMENT'** and submitted by **Bathinozu Ankitha Yamini** having ID-No. **2022MT12088** for the fulfilment of the requirements of **M. Tech Software Systems** degree of BITS, embodies the bonafide work done by him under my supervision.

Ch Karthikeya

Signature of the Supervisor

Place : Hyderabad

Date : 29.04.2024

Ch Karthikeya, Software Eng Manager, Honeywell, Hyderabad

Name, Designation & Organization &Location

## ABSTRACT

The exponential growth of the aviation industry has undeniably facilitated global connectivity and economic prosperity. However, this expansion has brought forth environmental challenges, primarily in the form of air pollution, which poses significant threats to both human health and climate stability. This dissertation explores the impacts of aviation on the environment, focusing on carbon dioxide (CO2) emissions as a significant contributor to air pollution. Despite the broad range of pollutants emitted into the atmosphere, aviation stands out due to its unique emissions profile and operational characteristics.

Understanding aviation's environmental impact requires distinguishing between general air pollution and aviation-specific emissions. While air pollution encompasses a wide array of pollutants from various sources, aviation-specific pollution stems specifically from aircraft operations, maintenance, and infrastructure. This differentiation is crucial as aviation emissions, majorly CO2, are released directly into the upper atmosphere, where they interact uniquely with atmospheric chemistry and climate dynamics.

Carbon dioxide, a prominent greenhouse gas emitted by the aviation industry, exacerbates climate change by trapping heat in the atmosphere for extended periods. Commercial aircraft contribute significantly to global CO2 emissions, with their high-altitude releases amplifying their warming effect compared to ground-level emissions from other sources.

This dissertation explores the potential of machine learning for sustainable aviation practices by utilizing daily CO2 emission data from various countries. Predictive ML models are developed to forecast future CO2 emissions from aviation, enabling informed policy decisions, such as setting emission reduction targets and evaluating regulations' effectiveness. Furthermore, airlines and aviation authorities can optimize flight operations and infrastructure development based on these forecasts, potentially leading to more fuel-efficient routes and airspace management strategies.

| | |
|---|---|
| B. Ankitha Yamini | Ch Karthikeya |
| **Signature of the Student** | **Signature of the Supervisor** |
| **Name:** Ankitha Yamini B | **Name:** Chiruvolu Karthikeya |
| **Date:** 29.04.2024 | **Date:** 29.04.2024 |
| **Place:** Nalgonda | **Place:** Hyderabad |

# ACKNOWLEDGEMENT

I extend my sincere thanks to my Supervisor, Chiruvolu Karthikeya, and Additional Examiner, Rangu Bhargavi, for their invaluable guidance throughout this project. Their support has been instrumental in my academic journey, and I'm truly grateful for their contributions.

## Contents

## 1. DATA SOURCE OVERVIEW:

The dataset, titled "CO2_emissions_by_aviation.csv," utilized in this study, originates from Carbon Monitor, a platform dedicated to tracking and analyzing carbon emissions globally

**Description:** This dataset provides comprehensive insights into carbon dioxide (CO2) emissions attributed to aviation activities across various countries and sectors. It includes information on the daily CO2 emissions recorded for each country, categorized by different aviation sectors such as International Aviation.

- The dataset comprises columns, "country," indicating the country where the CO2 emissions data was recorded, "date" specifying the date of the data record, "sector" denoting the domestic aviation and international aviation sectors, "value" representing the CO2 emissions value, and "timestamp" indicating the time of data recording.
- The dataset serves the purpose of facilitating in-depth analysis of temporal and spatial trends in aviation-related CO2 emissions. This data is utilized to understand the environmental impact of aviation and explore strategies for mitigating emissions.

## 2. DATA PRE-PROCESSING:

Data pre-processing is essential for guaranteeing the integrity and dependability of data intended for further analysis. This phase encompasses several crucial steps aimed at refining the datasets, such as data loading, cleaning, transformation, and feature engineering.

### Data transformation and cleaning

- Load the global CO2 emissions dataset from the provided CSV file.
- The main label 'value', representing CO2 emissions, is log-transformed using the base 10 logarithm to address potential skewness in the distribution of CO2 emissions.
- This Log-transformation is applied to skewed data distributions to stabilize variance and improve the performance of machine learning models, especially when the data exhibits a long-tailed distribution.
- The 'date' column in the dataset is converted to a datetime format to facilitate temporal analysis.
- Dropped unused columns ('value', 'timestamp', 'date') from the dataset.

### Feature Engineering

- The month-year information is extracted from the 'date' column and stored as a separate column (**'month_year'**).
- This feature allows for the aggregation of CO2 emissions data at the monthly level, enabling the exploration of temporal patterns and trends.
- It aims to preprocess the dataset and extract meaningful information from the raw data, making it suitable for training machine learning models. By transforming and creating features appropriately, the feature engineering process enhances the

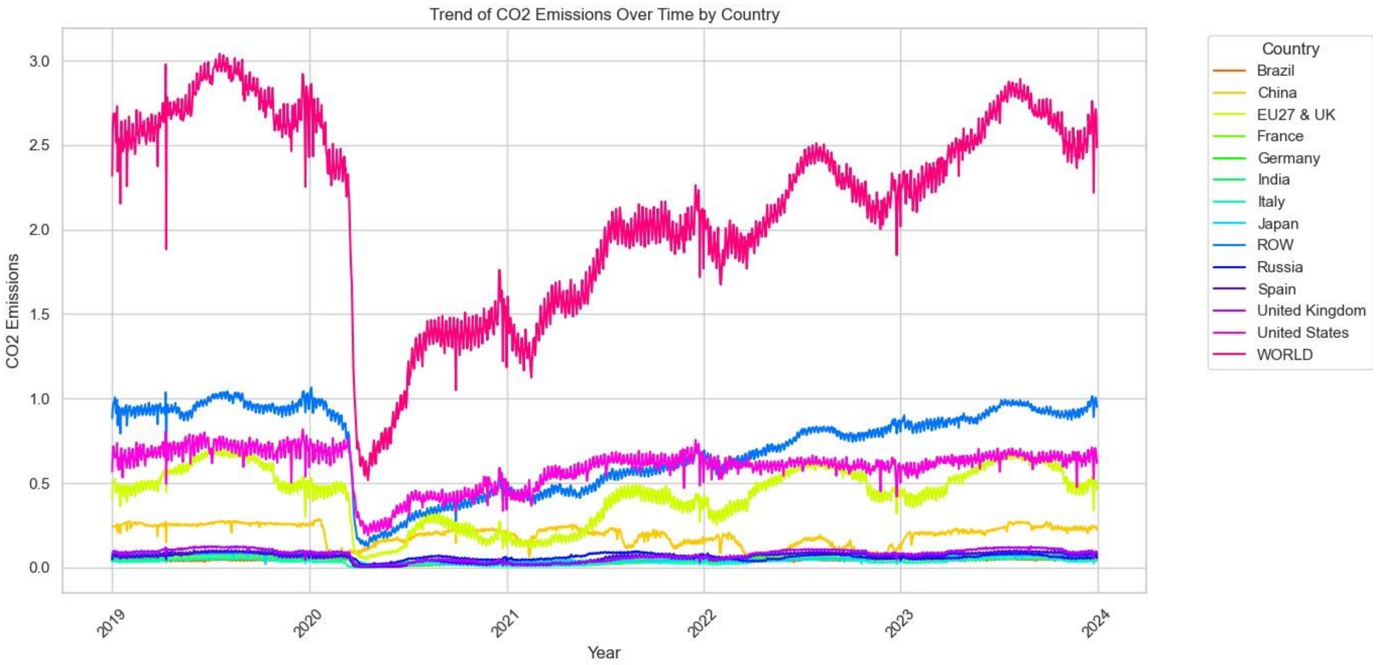
predictive performance of the models and facilitates the interpretation of model outputs.

## 3. EXPLORATORY DATA ANALYSIS

During the exploratory data analysis (EDA), several significant insights have found:

- Distinct disparities in CO2 emissions across countries and aviation were apparent, making significant contributions to the total emissions.
- A significant decrease in CO2 emissions was noted during the COVID-19 lockdown period.
- Seasonal fluctuations in CO2 emissions were observable annually
- Various sources of CO2 emissions underwent scrutiny to examine their distribution and trends over time.

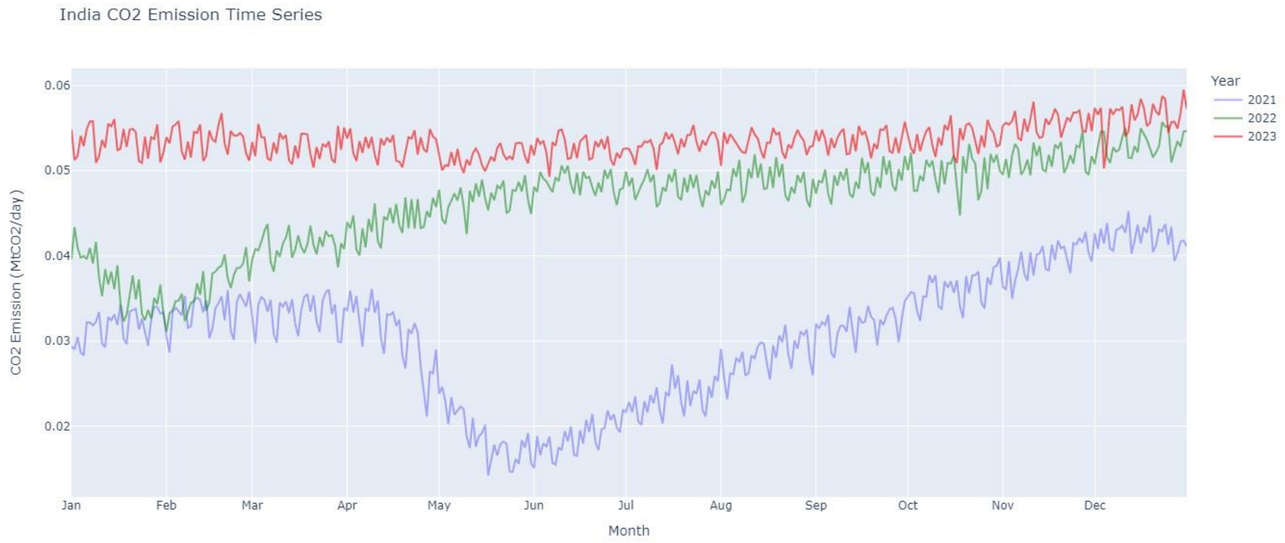
**Trend of CO2 Emissions from Aviation by Country:**



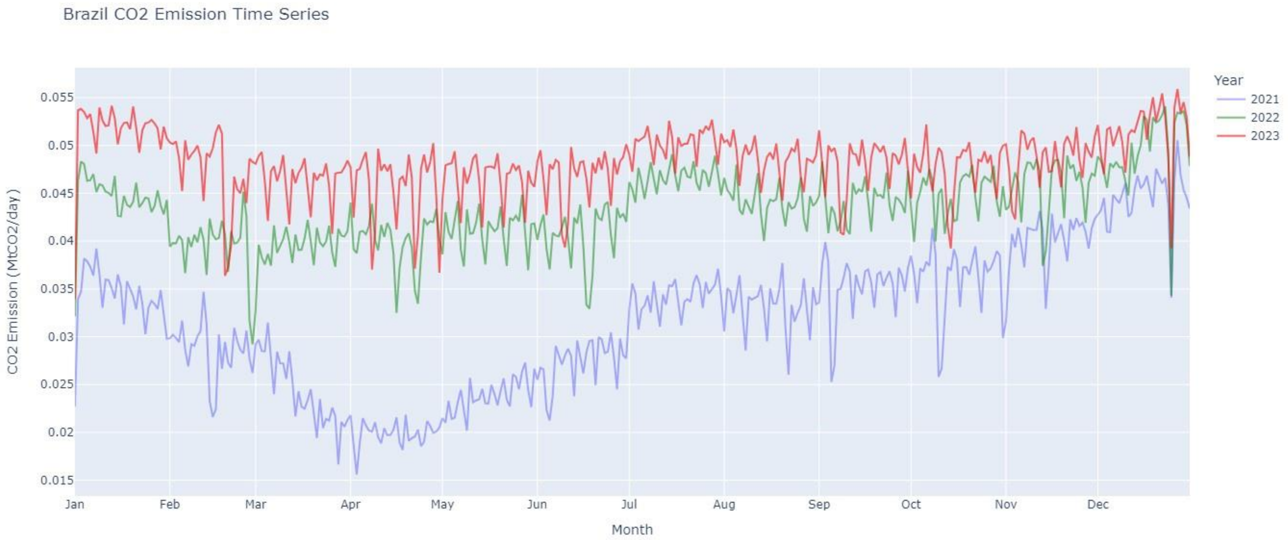
**Time Series Analysis of Country-wise CO2 Emissions by Aviation over a period of last 3 years**:

**Insight:** There is a seasonal trend in the data, with emissions appearing to be higher in the winter months and lower in the summer months. This is likely because there is more heating energy use in the winter.

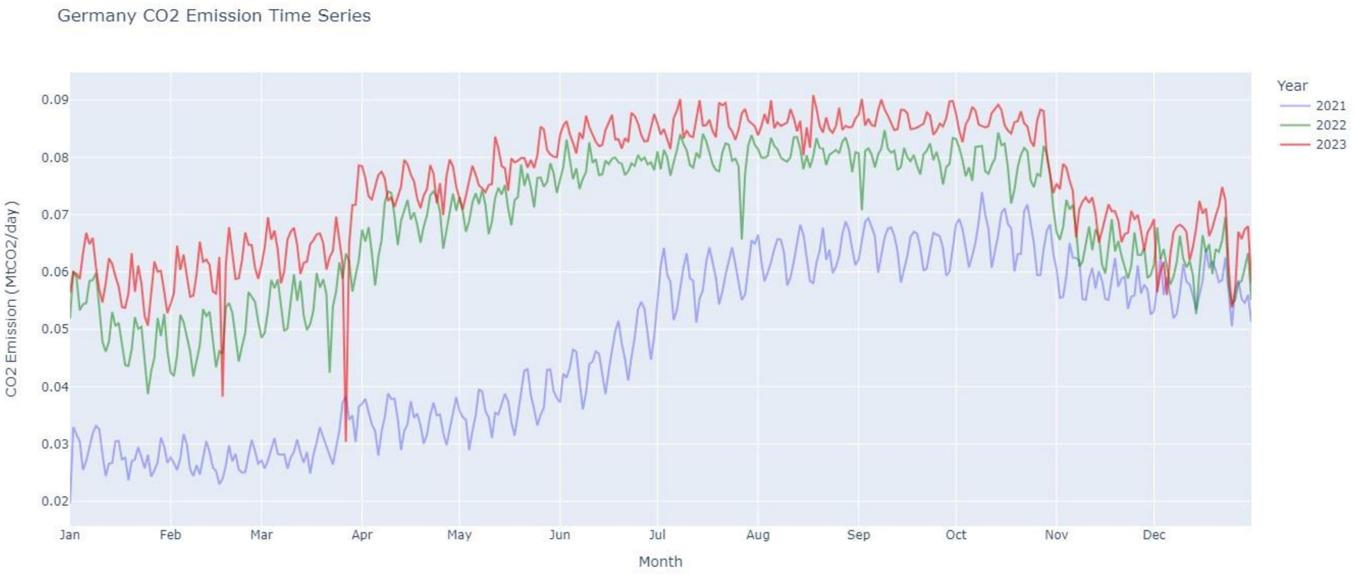
## India’s CO2 emissions:



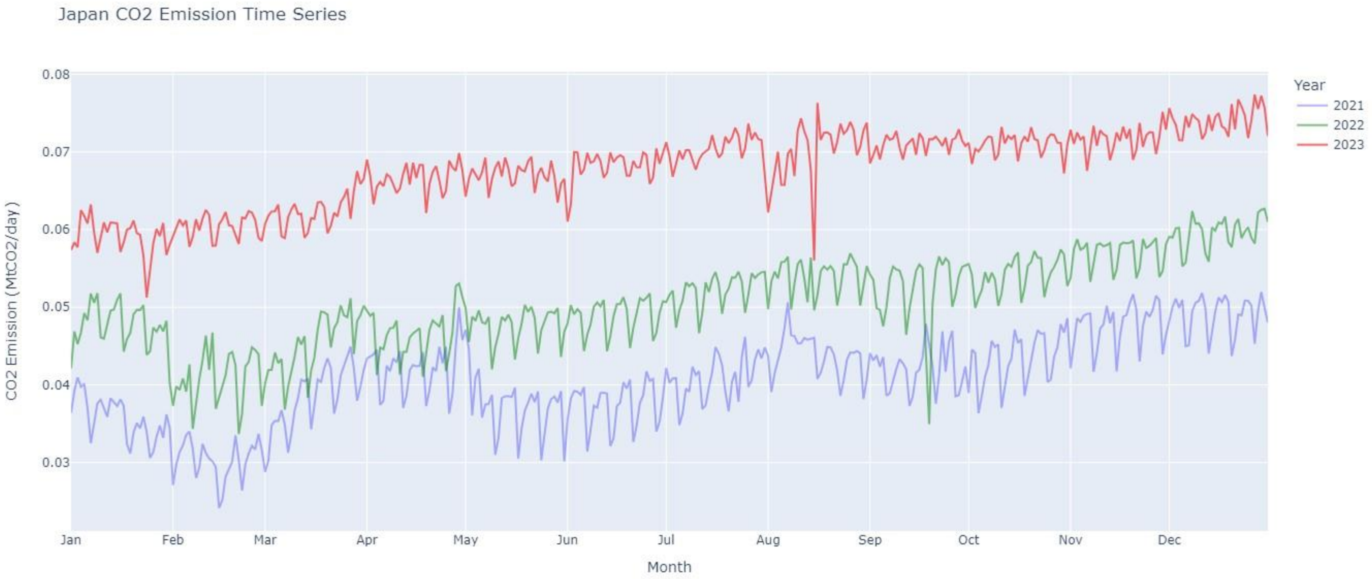
## Brazil’s CO2 emissions:



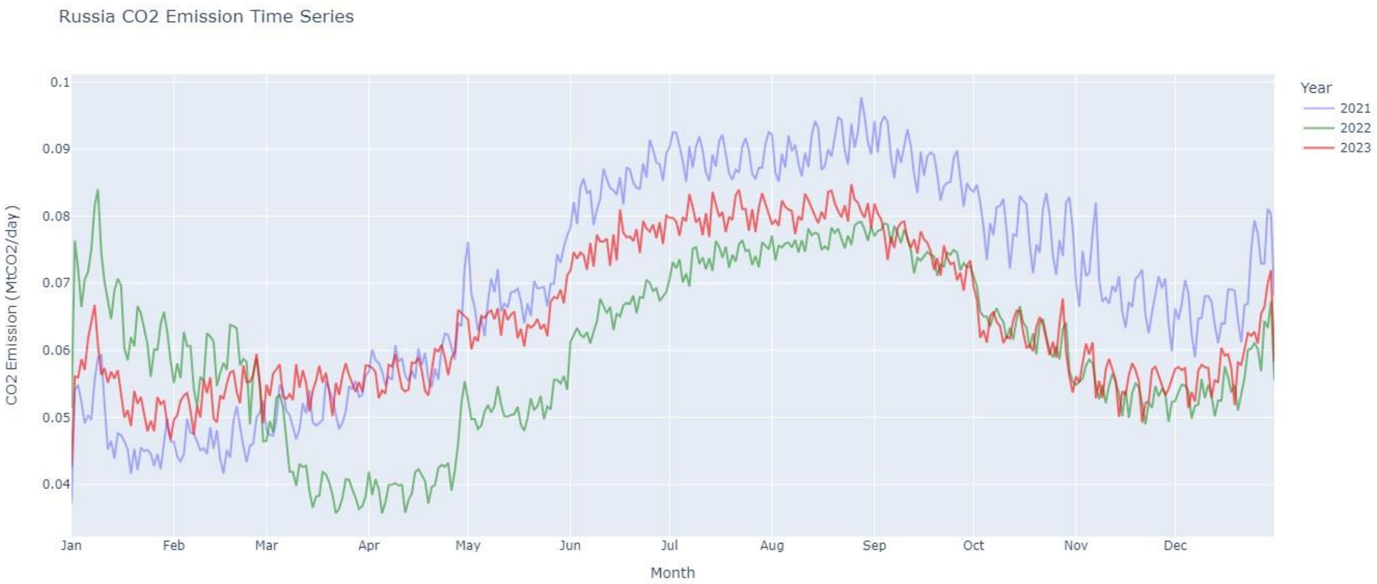
## Germany's CO2 emissions:



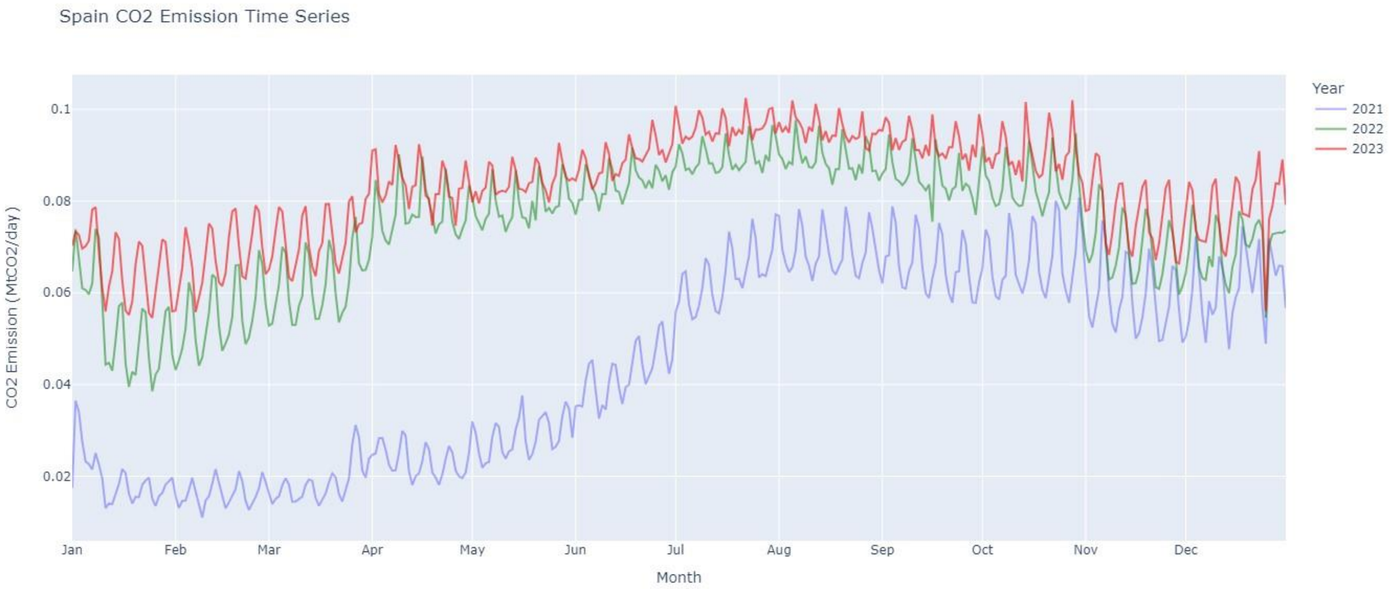
## Japan's CO2 emissions:



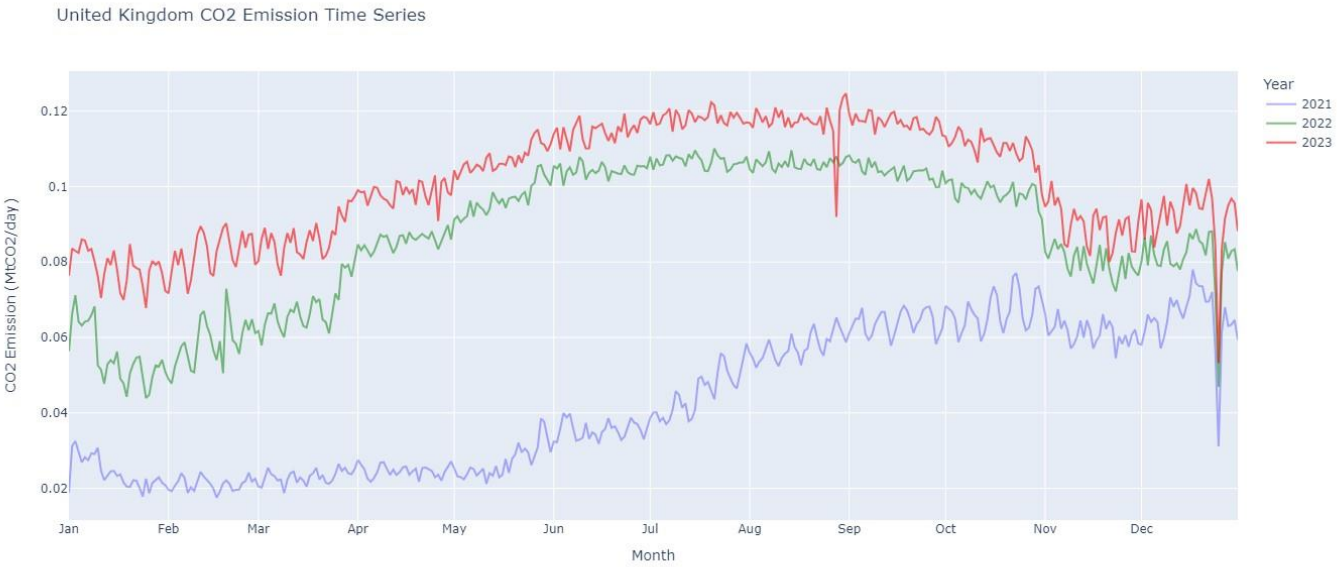
## Russia’s CO2 emissions:



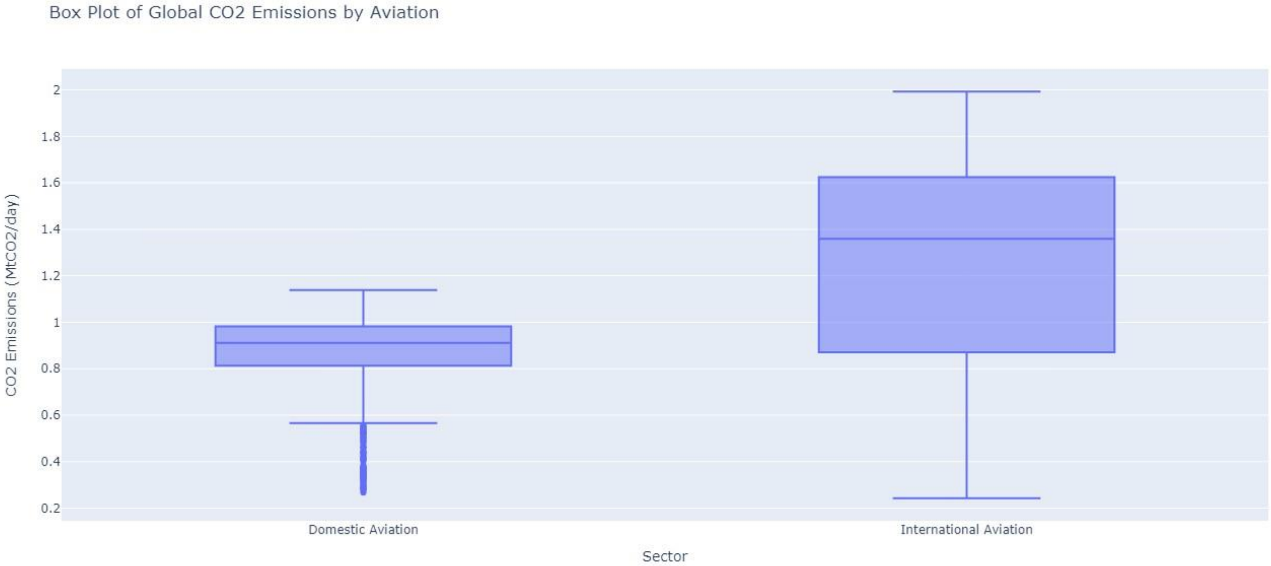
## Spain’s CO2 emissions:



### United Kingdom's CO2 emissions:



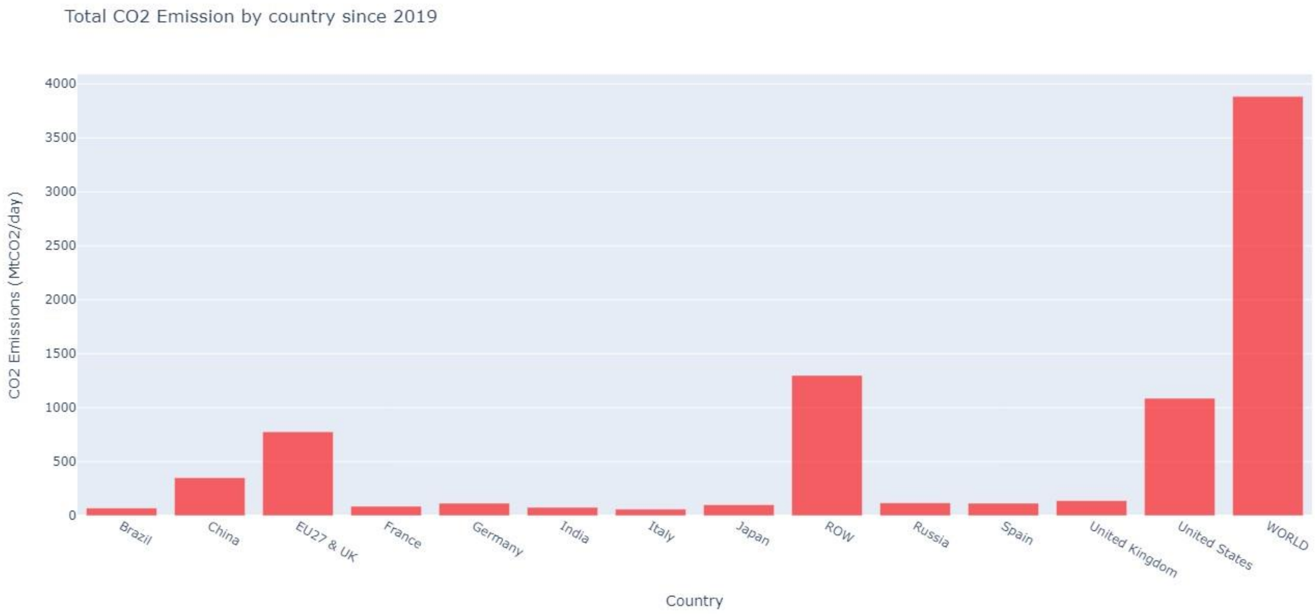
### Box Plot of Global CO2 Emissions by Aviation:



### Insights:

- The box shows the spread of CO2 emissions for domestic aviation and international aviation.
- Based on the position of the line in the middle of the box, it appears that CO2 emissions from domestic aviation tend to be lower than CO2 emissions from international aviation.
- The box for domestic aviation is smaller than the box for international aviation. This suggests that CO2 emissions from domestic flights are more tightly clustered around the median than CO2 emissions from international flights. There is more variability in emissions from international flights.

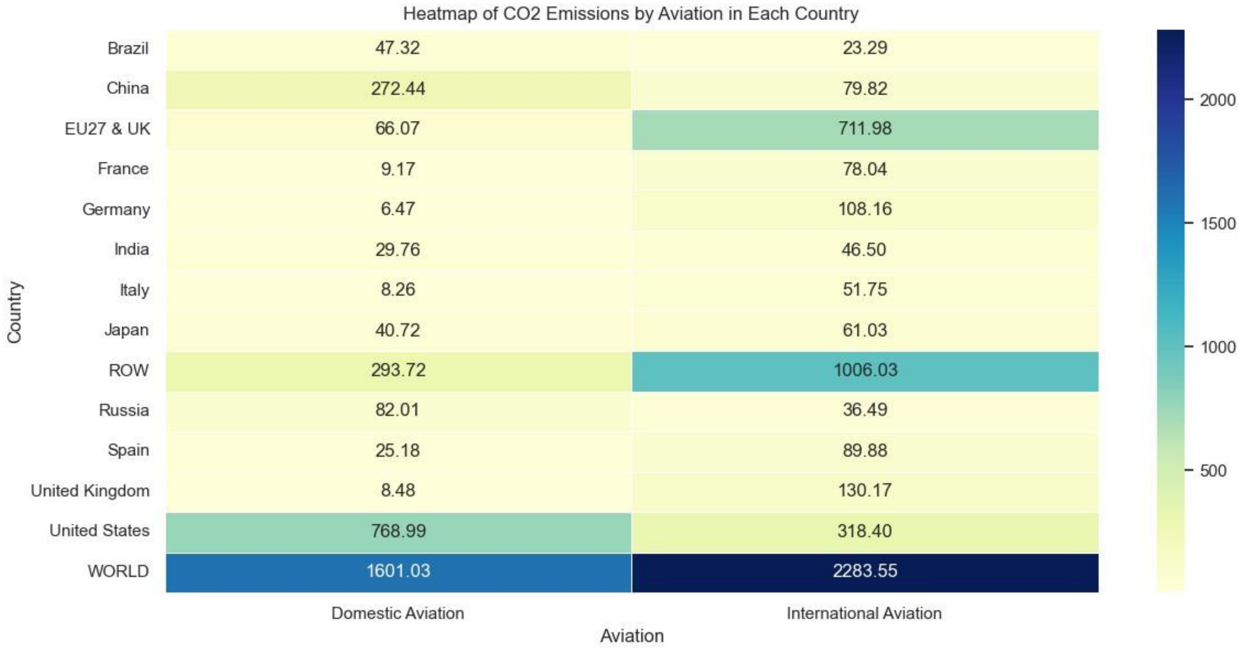
**Bar graph of CO2 emissions from aviation across countries:**



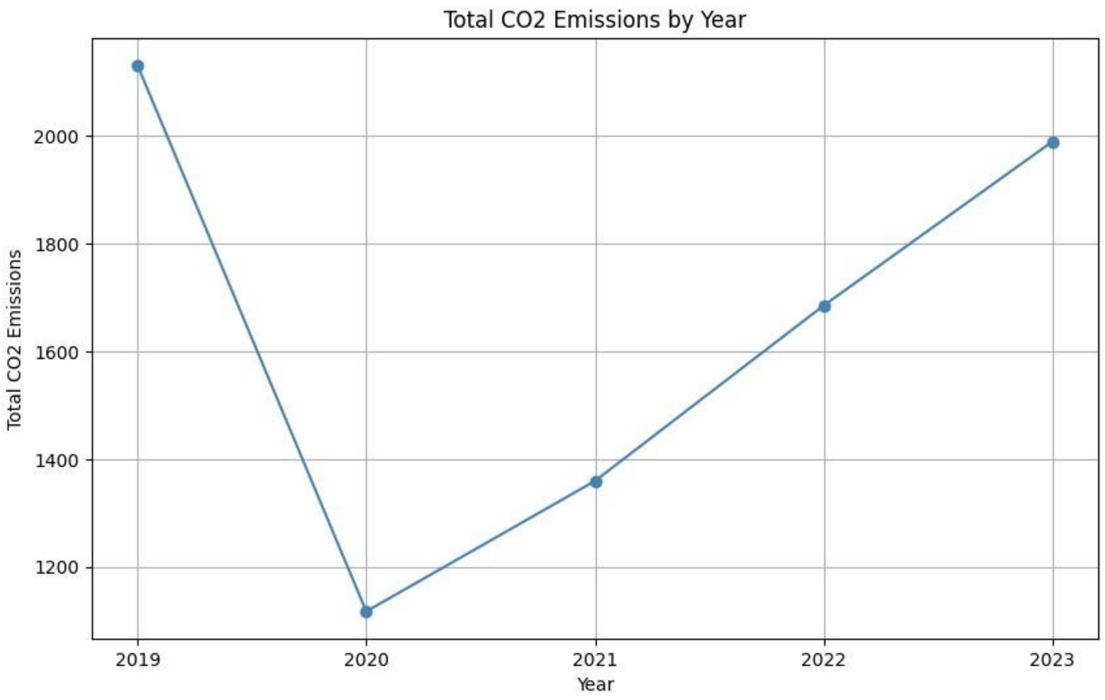
## Insights:

- The histogram shows the distribution of total CO2 emissions from aviation across different counties. The bars on the x-axis represent the range of CO2 emissions in MtCO2, while the height of each bar represents the number of countries that fall within that emissions range.
- We can now say that some countries emit lower amounts of CO2 (left side of the histogram), while others emit higher amounts (right side).
- By looking at the far right side of the histogram, we can see if there are any countries with a significant number of emissions compared to other countries. These countries could be potential high emitters of CO2 from aviation.

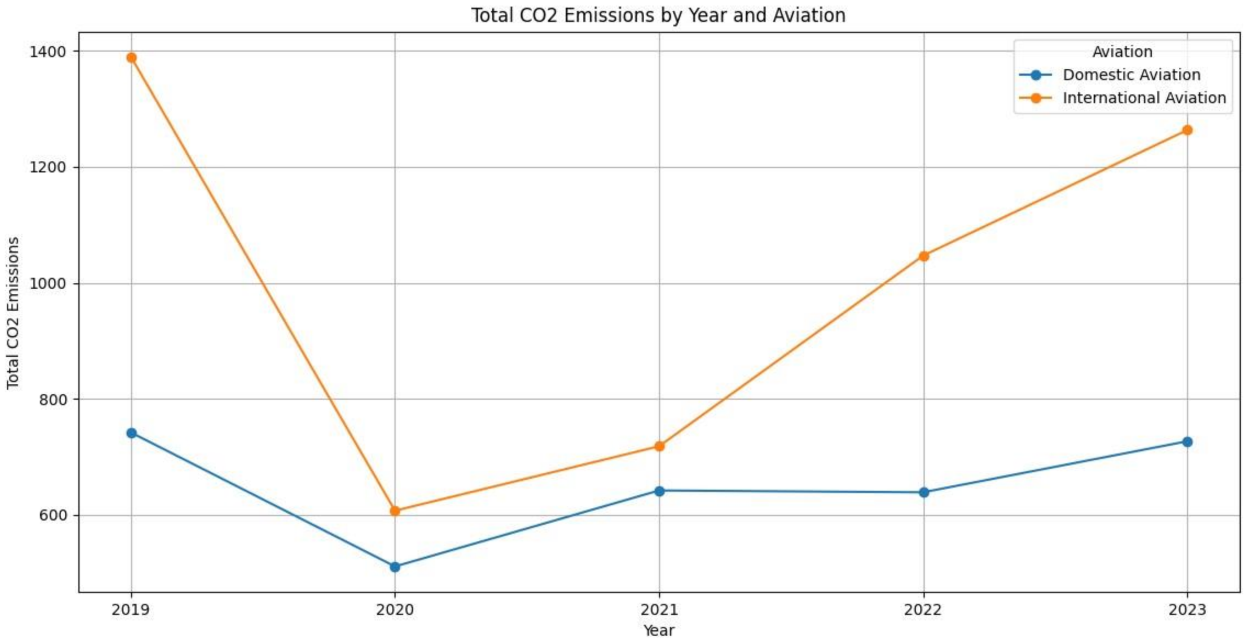
## Heatmap of CO2 Emissions by Aviation in Each Country:



## Line Plot of Total CO2 emissions from Aviation (Domestic & International) by Year:



**Line plot of Total CO2 emissions by Year and Aviation:**



## 4. MACHINE LEARNING MODEL

- CatBoostRegressor model is used for predicting CO2 emissions by aviation in a country . It is a gradient boosting algorithm specifically designed for regression tasks. It belongs to the family of ensemble learning methods and is known for its high performance, robustness to overfitting, and ability to handle categorical features efficiently.
- It utilizes a gradient boosting framework, which combines multiple weak learners (decision trees) to create a strong predictive model. It optimizes the learning process by iteratively fitting new trees to the residuals of the previous ones, gradually improving the model's predictive accuracy.
- This model is chosen due to its scalability and flexibility to introduce more categorical features (like flight type, Aircraft type, Airport, Operational conditions, Regulatory Compliance) for future developments.
- Additionally, if the dataset grows in complexity or if more nuanced relationships between features emerge, Its robustness to overfitting and ability to handle non-linear relationships become more beneficial.
- Therefore, CatBoostRegressor is well-suited for regression tasks like predicting CO2 emissions, as it can effectively capture the complex relationships between input features and the target variable while minimizing the risk of overfitting.

**Splitting the Dataset:**

The dataset is split into training and testing sets using the **`train_test_split()`** function from scikit-learn. The training set contains a specified percentage of the data (in this case, 50%), which is used to train the machine learning model, while the remaining data is reserved for testing the model's performance.

**Initializing and Training a CatBoostRegressor Model:**

- The model is trained using the training data to learn the relationship between the input features such as country, date, sector (domestic aviation and international aviation), month_year  and the target variable is log-transformed CO2 emissions (`log10_CO2`)
- Predictions are made on both the training and testing sets.

**SHAP:**

SHAP (SHapley Additive exPlanations) is employed in the context of the prediction model to provide insights into the underlying factors driving individual predictions. Here's how SHAP is used in relation to the model:

- After training the CatBoostRegressor model to predict CO2 emissions by aviation, SHAP values are calculated for the test dataset. These SHAP values quantify the impact of each feature on the model's predictions for individual data points.
- By analyzing SHAP values, we can determine which features have the most significant influence on the predicted CO2 emissions. This analysis helps prioritize efforts for mitigating environmental impacts by identifying the key factors driving emissions.
- SHAP values offer a transparent way to explain the model's predictions. They provide a clear breakdown of how each feature contributes to individual predictions, enhancing the interpretability of the model's decision-making process.
- The transparent explanations provided by SHAP values foster trust and confidence in the model's predictions. This trust is essential for the adoption of the model in decision-making processes related to environmental policies and interventions.

In summary, SHAP usage in the model facilitates interpretation, validation, and improvement of the prediction model, enhancing its transparency, trustworthiness, and suitability for informing environmental decision-making.

**Evaluating the Trained Model:**

- The trained model is evaluated on both the training and testing sets using various evaluation metrics:
  - **RMSE (Root Mean Squared Error):** Measures the average deviation of the predicted values from the actual values, with lower values indicating better model performance.
  - **R-squared (Coefficient of Determination):** Represents the proportion of the variance in the target variable that is explained by the model. A value closer to 1 indicates a better fit of the model to the data.
  - **MAE (Mean Absolute Error):** Measures the average absolute difference between the predicted and actual values, providing insight into the model's prediction accuracy.

**Implementing a Baseline Model:**

- A simple baseline model is implemented to serve as a point of comparison for the performance of the trained CatBoostRegressor model.
- The baseline model simply predicts the mean value of the target variable (log-transformed CO2 emissions) for all data points regardless of their features.

- This simplistic approach serves as a reference point to assess the effectiveness of the trained model in making accurate predictions.

**Performance metrics:**

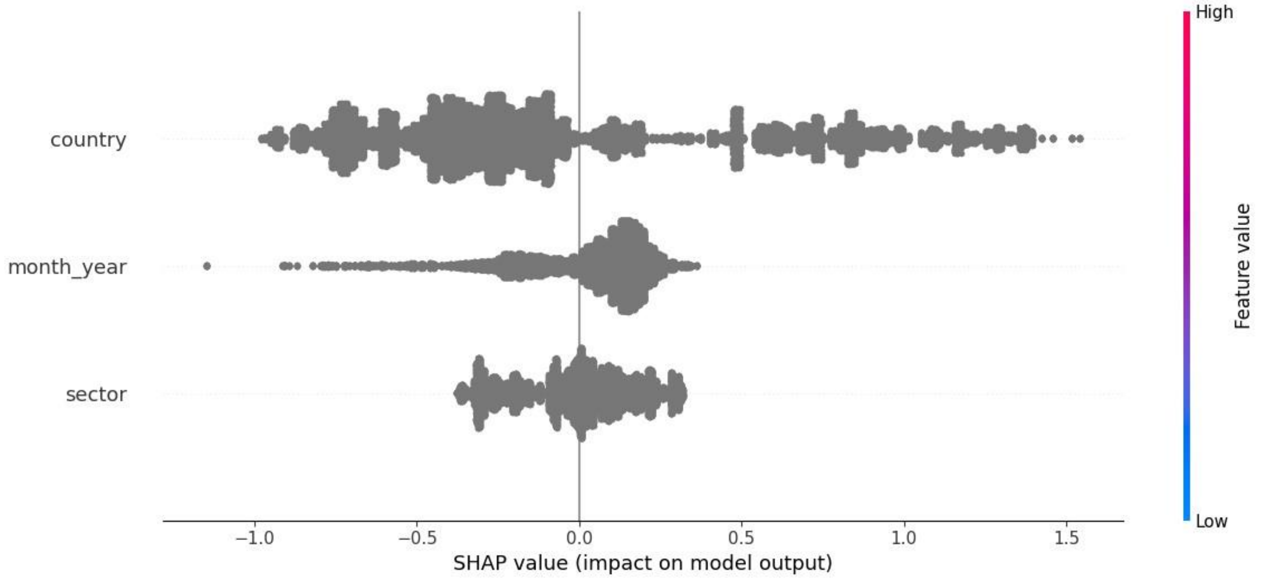
**Trained Model:**

| | **Train data** | **Test data** |
|---|---|---|
| RMSE | 0.0756 | 0.0823 |
| R-squared | 0.9891 | 0.9872 |
| MAE | 0.0453 | 0.0468 |

Average predicted CO2 emissions is 0.043
Average actual CO2 emissions is 0.043

**Baseline Model:**

| | **Train data** | **Test data** |
|---|---|---|
| RMSE | 0.5290 | 0.5307 |
| R-squared | 0.0 | 0.0 |
| MAE | 0.5767 | 0.5765 |

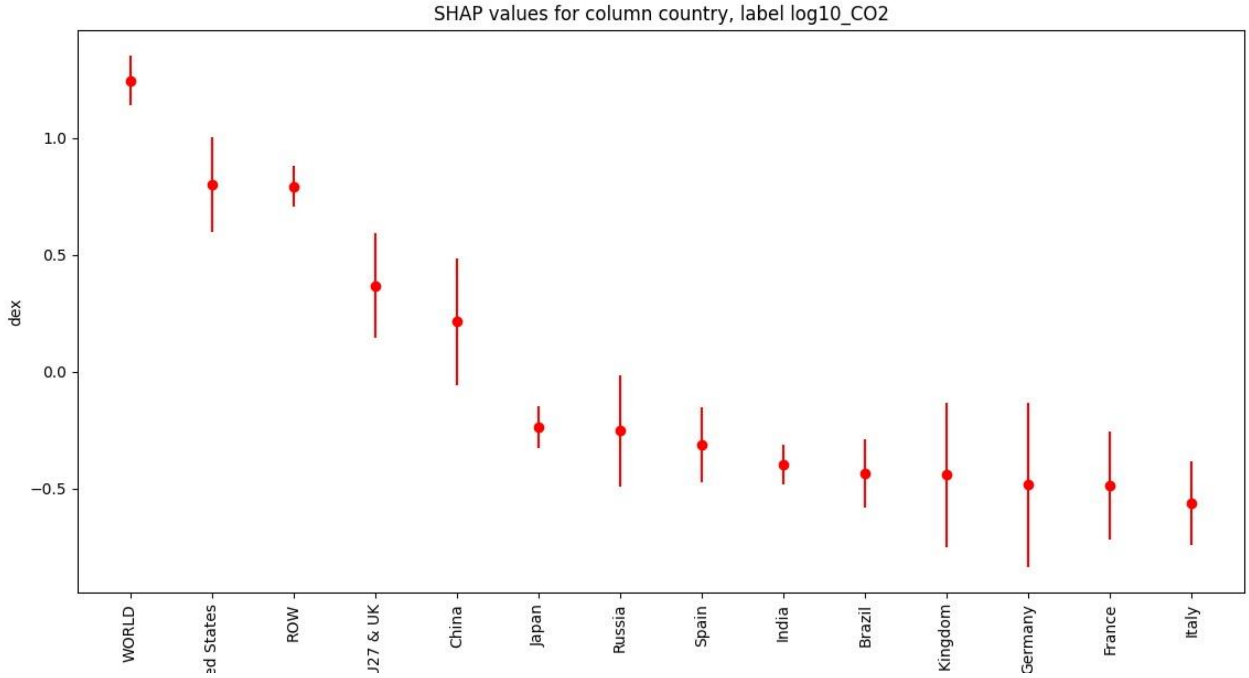


This scatter plot visualizes SHAP feature importance for a model predicting log10(CO2) emissions. Each point on the plot represents a data instance (country, month_year, sector). The model considers various features, including country, sector, and month/year, to predict CO2 emissions.

- **X-axis:** SHAP value (impact on model output)
- **Y-axis:** Feature value (country, sector, month_year)
- **Color:** The color of the point represents the value of the feature. Red indicates high values, and blue indicates low values.

**Interpretations:**

- **Country:** Countries with higher SHAP values (more red) tend to have a greater positive influence on the model's prediction of CO2 emissions. Conversely, countries with lower SHAP values (more blue) tend to have a negative influence on predicted CO2 emissions.
- **Sector:** Sector here is Domestic Aviation and International Aviation. Similar to countries, sectors with higher SHAP values are likely to be associated with higher predicted CO2 emissions.
- **month_year:** The month_year encoded points show the model's learned seasonality in CO2 emissions. For example, a red point for a specific month_year might indicate that the model predicts higher CO2 emissions in that month_year on average.



**Insights from the Graph:**

This SHAP plot has some insights into the relative impact of different countries on the model's predictions for CO2 emissions:

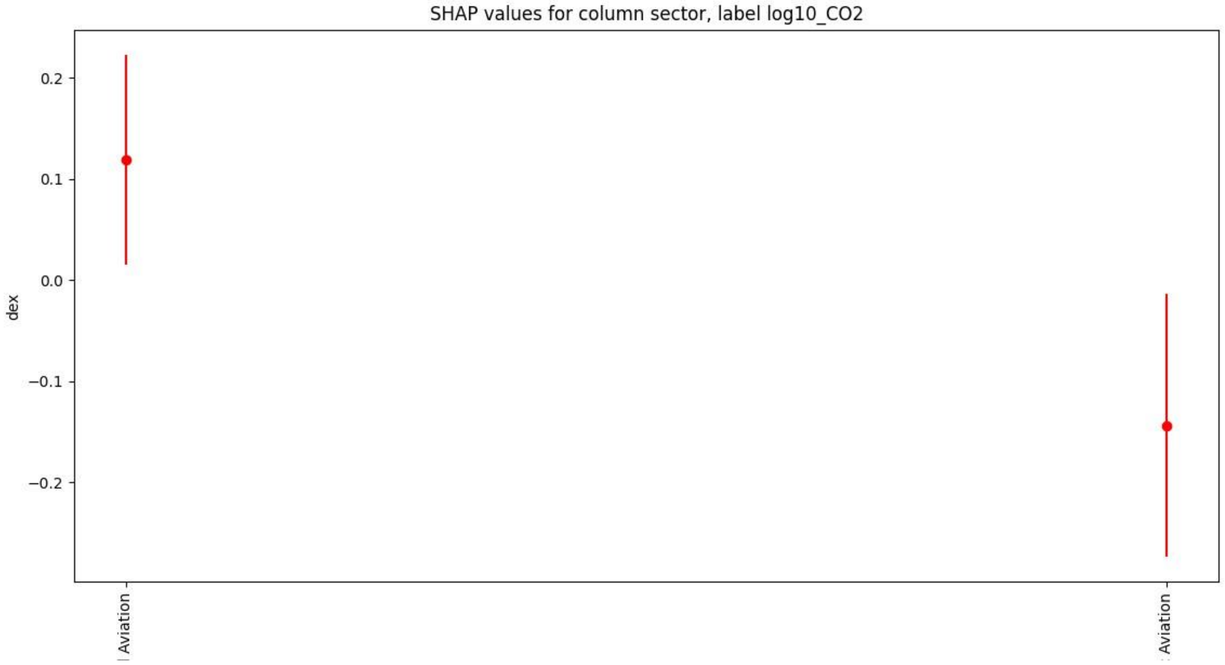
- Countries with the highest positive SHAP values (i.e., those that contribute most to higher predicted CO2 emissions) are:
  - WORLD
  - United States
  - ROW (Rest of the World)
  - EU27 & UK
- Countries with the highest negative SHAP values (i.e., those that contribute most to lower predicted CO2 emissions) are:
  - China
  - Japan
  - Russia

- Spain
- India
- Brazil
- Germany
- France
- Italy

### Interpretation:

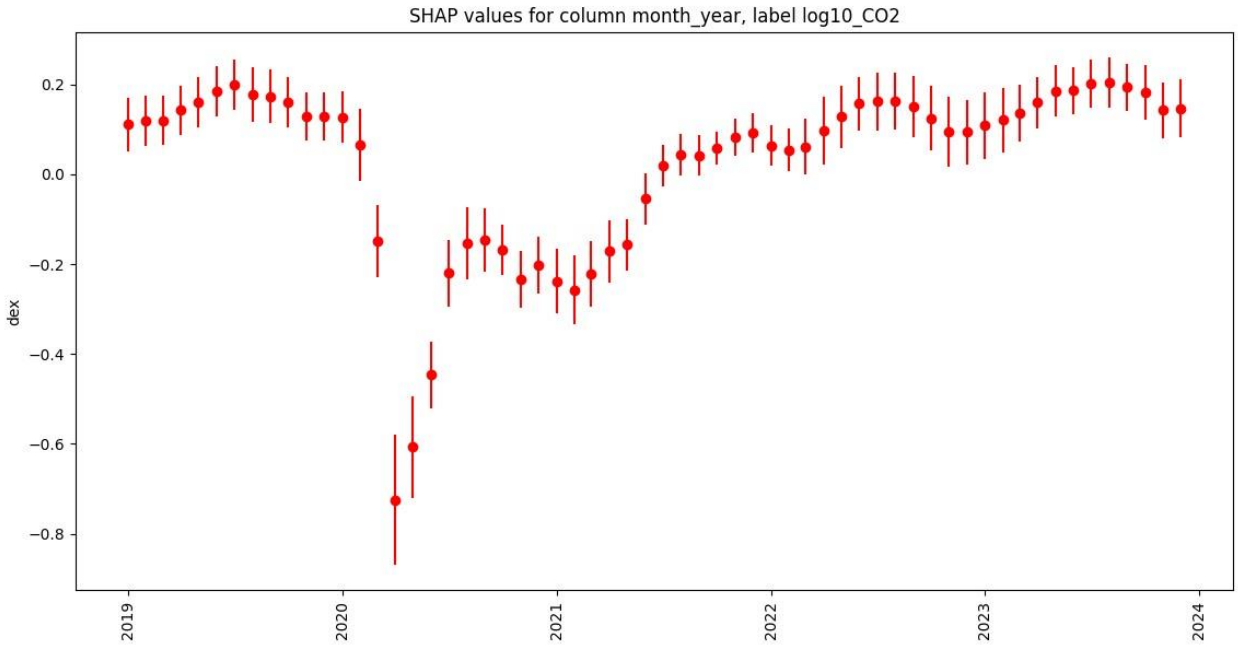
The SHAP plot suggests that the model predicts higher CO2 emissions for countries like the World (presumably referring to developed countries as a whole), the United States, and ROW (Rest of the World). This could be due to factors such as higher levels of economic activity and energy consumption in these countries.

On the other hand, countries like China, Japan, and Russia are predicted to have lower CO2 emissions. This could be due to a number of reasons, such as differences in industrial structure, energy sources, and emissions regulations.



### Insight:

This SHAP plot insights into the relative impact of the sector on the model's predictions for CO2 emissions. It suggests that the model strongly associates aviation with higher CO2 emissions. This is likely due to the fact that airplanes emit a significant amount of CO2 into the atmosphere during flight.



**Insights**:
Below are some insights into the relative impact of different years on the model's predictions for CO2 emissions:

- Years with the highest positive SHAP values (i.e., those that contribute most to higher predicted CO2 emissions) are:
    - 2024 (most recent year in the data)
    - 2023
    - 2022
- Years with the lowest SHAP values (i.e., those that contribute most to lower predicted CO2 emissions) are:
    - 2019
    - 2020
    - 2021

**Interpretation:**

The SHAP plot suggests that the model predicts higher CO2 emissions for more recent years (2022-2024) compared to earlier years (2019-2021). This could be due to a number of factors, such as:

- An increase in air traffic activity in recent years
- Changes in fuel efficiency of airplanes over time
- Other factors that may affect CO2 emissions, such as economic activity or weather patterns

## 5. SUMMARY

The dissertation analyses the environmental impact of aviation, focusing on carbon dioxide (CO2) emissions and employing data analysis and machine learning techniques to derive insights and predictions.

The study reveals that aviation significantly contributes to CO2 emissions, exacerbating air pollution and climate change. Through exploratory data analysis, distinct disparities in CO2 emissions across countries and aviation sectors are identified, with notable seasonal fluctuations and observable trends over time.

CatBoostRegressor model is used to predict CO2 emissions from aviation activities, and evaluated its performance against a baseline model. Leveraging features such as country, sector (domestic vs. international aviation), and temporal information, the model demonstrates robust performance in forecasting emissions, with high accuracy and efficiency.

The trained model exhibits superior performance compared to a baseline model, as evidenced by evaluation metrics such as RMSE, R-squared, and MAE. This indicates the model's effectiveness in capturing the complex relationships between input features and CO2 emissions, thereby enabling accurate predictions.

- **RMSE** of 0.0756 on the training data and 0.0823 on the testing data. These low RMSE values indicate that the model's predictions are close to the actual values, demonstrating its accuracy in estimating CO2 emissions.
- **R-squared** values of 0.9891 for the training data and 0.9872 for the testing data indicate that the model explains a high proportion of the variance in CO2 emissions. This implies a strong fit of the model to the data.
- **MAE** values of 0.0453 on the training data and 0.0468 on the testing data. These low MAE values indicate minimal average deviation between the predicted and actual CO2 emissions.

**Baseline Model Comparison:** It exhibits significantly higher RMSE, R-squared, and MAE values compared to the trained model, indicating inferior predictive performance.

## 6. RECOMMENDATIONS

**Adopting Sustainable Practices:** Encourage the aviation industry to adopt sustainable practices, such as investing in fuel-efficient aircraft, optimizing flight routes, and implementing alternative propulsion technologies like electric or biofuel-powered planes.

**Regulatory Measures:** Implement stringent regulations and emission reduction targets for the aviation sector to curb CO2 emissions. This could include emissions trading schemes, carbon pricing mechanisms, and incentives for investing in green technologies.

**Infrastructure Development:** Invest in modernizing aviation infrastructure to enhance efficiency and reduce emissions. This includes upgrading air traffic management systems, developing eco-friendly airports, and promoting the use of renewable energy sources.

**International Collaboration:** Foster international collaboration and cooperation to address aviation emissions on a global scale. This could involve joint research initiatives, technology sharing, and harmonizing emission standards across countries.

**Public Awareness and Education:** Raise public awareness about the environmental impact of aviation and encourage individuals to make environmentally conscious travel choices, such as opting for alternative modes of transportation or offsetting their carbon footprint.

## 7. FUTURE WORK

**Refinement of Machine Learning Models:** Continuously refine and improve machine learning models for predicting CO2 emissions from aviation activities. This involves incorporating additional features, refining algorithms, and validating model performance on real-world data.

**Exploration of Alternative Fuels:** Further research and development into alternative aviation fuels, such as sustainable biofuels or hydrogen, to reduce reliance on fossil fuels and mitigate CO2 emissions.

**Impact of Technological Advances:** Investigate the potential impact of emerging technologies, such as electric or hybrid-electric aircraft, on reducing aviation emissions and their feasibility for widespread adoption.

**Policy Evaluation:** Evaluate the effectiveness of existing environmental policies and regulations in mitigating aviation emissions and identify areas for improvement or expansion.

**Global Data Collaboration:** Foster collaboration among countries and international organizations to collect and share comprehensive data on aviation emissions, enabling more accurate modeling and analysis of environmental impacts on a global scale.

## 8. REFERENCES

1. *Zhang, Yongjun, Xu, Yifan, Sun, Huijun, Sun, Wei, & Zhang, Aiqin. (2021). Machine Learning Techniques for Aircraft Emission Prediction: A Review. Applied Sciences, 11(22), 10806.*
2. *International Air Transport Association (IATA) publications on environmental sustainability.*
3. *Zhang, Q., & Cao, J. (2015). Air pollution and control action in Beijing. Journal of Cleaner Production, 112, 1519-1527.*

## 9. ABBREVIATIONS

| | |
|---|---|
| CO2 | Carbon dioxide |
| RMSE | Root Mean Squared Error |
| MAE | Mean Absolute Error |
| R-Squared | Coefficient of Determination |
| SHAP | SHapley Additive exPlanations |

## 10. CHECKLIST

## Checklist of items for the Final Dissertation Report

**This checklist is to be duly completed, verified and signed by the student.**

| | | |
|---|---|---|
| 1. | **Is the final report neatly formatted with all the elements required for a technical Report?** | Yes |
| 2. | Is the Cover page in proper format as given in Annexure A? | Yes |
| 3. | Is the Title page (Inner cover page) in proper format? | Yes |
| 4. | (a) Is the Certificate from the Supervisor in proper format?<br>(b) Has it been signed by the Supervisor? | Yes<br>Yes |
| 5. | Is the Abstract included in the report properly written within one page? Have the technical keywords been specified properly? | Yes<br>Yes |
| 6. | Is the title of your report appropriate? **The title should be adequately descriptive, precise and must reflect scope of the actual work done.** Uncommon abbreviations / Acronyms should not be used in the title | Yes |
| 7. | Have you included the List of abbreviations / Acronyms? | Yes |
| 8. | Does the Report contain a summary of the literature survey? | Yes |
| 9. | Does the Table of Contents include page numbers?<br>(i). Are the Pages numbered properly? (Ch. 1 should start on Page # 1)<br>(ii). Are the Figures numbered properly? (Figure Numbers and Figure Titles should be at the bottom of the figures)<br>(iii). Are the Tables numbered properly? (Table Numbers and Table Titles should be at the top of the tables)<br>(iv). Are the Captions for the Figures and Tables proper?<br>(v). Are the Appendices numbered properly? Are their titles appropriate | Yes<br>Yes<br>Yes<br>Yes<br>Yes<br>Yes |
| 10. | Is the conclusion of the Report based on discussion of the work? | Yes |
| 11. | Are References or Bibliography given at the end of the Report?<br>Have the References been cited properly inside the text of the Report?<br>Are all the references cited in the body of the report | Yes<br>Yes<br>Yes |
| 12. | Is the report format and content according to the guidelines? The report should not be a mere printout of a Power Point Presentation, or a user manual. Source code of software need not be included in the report. | Yes |

**Declaration by Student:**

I certify that I have properly verified all the items in this checklist and ensure that the report is in the proper format as specified in the course handout.

B. Ankitha Yamini

| | |
|---|---|
| **Place:** Nalgonda | **Signature of the Student** |
| **Date:** 29.04.2024 | **Name:** B. Ankitha Yamini |
| | **ID No.:** 2022MT12088 |