Student Number: BP0288879

Is the UK on track to meet the fourth emissions budget of The Climate Change Act? Assessing the feasibility of meeting the 1,950 MtCO2e 2027 emissions cap.

Data Science Project Report

Contents

[List of Figures 2](#_Toc175637679)

[Section A – Data Science Project 3](#_Toc175637680)

[Research Question 3](#_Toc175637681)

[Executive Summary 3](#_Toc175637682)

[Project Background 3](#_Toc175637683)

[Data Description 3](#_Toc175637684)

[Data Source 3](#_Toc175637685)

[Data Characteristics 3](#_Toc175637686)

[Data Preparation 4](#_Toc175637687)

[Methodology 5](#_Toc175637688)

[Model Selection 5](#_Toc175637689)

[Implementation 5](#_Toc175637690)

[Evaluation Metrics 5](#_Toc175637691)

[Further Analysis 7](#_Toc175637692)

[Project Findings 9](#_Toc175637693)

[Conclusion 10](#_Toc175637694)

[Bibliography 12](#_Toc175637695)

[Appendix 13](#_Toc175637696)

[Section A – Data Science Project 13](#_Toc175637697)

## 

## List of Figures

[Figure 1 - Table showing the data types of each column 3](#_Toc175133249)

[Figure 2 - A screen capture showing the data source in its original format 4](#_Toc175133250)

[Figure 3 - A screen capture showing the data source following transformation 4](#_Toc175133251)

[Figure 4 - A Table showing the model options trialled for the forecast 4](#_Toc175133252)

[Figure 5 A table showing the ARIMA Model outputs for each Industry Group 5](#_Toc175133253)

[Figure 6 - Four plots highlighting trends and seasonality within the dataset 7](#_Toc175133254)

[Figure 7 - Two correlation plots 8](#_Toc175133255)

[Figure 8 - heatmap visualising the forecast errors across different industry groups 9](#_Toc175133256)

# Section A – Data Science Project

## Research Question

Is the UK on track to meet the fourth emissions budget of The Climate Change Act? Assessing the feasibility of meeting the 1,950 MtCO2e 2027 emissions cap.

## 

## Executive Summary

This project utilises data from the Office of National Statistics (ONS) on greenhouse gas emissions by industry, spanning 1990 to 2022. The analysis is designed for strategists from the Committee on Climate Change (CCC), who are responsible for ensuring emissions targets are evidence-based and independently assessed. This analysis is crucial for evaluating the UK's progress towards the 2027 emissions commitment, as mandated by the Climate Change Act. It will provide stakeholders with a time series forecast model to assess the likelihood of meeting the 2027 emissions targets.

## Project Background

This project employs Excel, Tableau, and Python for data analysis and visualisation, supported by PowerPoint for stakeholder presentations. It analyses UK CO2 emissions by industry from 1990 to 2022, aiding strategists in making evidence-based predictions essential for meeting the UK's legally binding carbon budgets under the Climate Change Act. Accurate predictions are crucial, as these budgets limit emissions over five-year periods and must be set at least 12 years in advance.

## Data Description

### Data Source

The project utilises the ONS dataset titled ‘A*tmospheric Emissions: Greenhouse Gases by Industry and Gas’,* providing an overview of UK territorial greenhouse gas emissions, published by the Department for Energy Security and Net Zero in June 2024. Given that “*CO2 has consistently been the dominant greenhouse gas in the UK*” (DESNZ, 2024a), the analysis will focus exclusively on CO2 emissions.

### Data Characteristics

The dataset provided by the ONS consists of multiple Excel file tabs, each containing varying levels of detail. The format of the data is consistent across all tabs, with each tab displaying emissions data for a distinct Green House Gas (GHG) type. Each tab contains two pivot tables:

* Industry Group Pivot Table: Summarises GHG emissions by broader industry group categories.
* Distinct Industry Pivot Table: Details GHG emissions by specific industries.

#### A table of data type Description automatically generatedData Types

Figure 1 - Table showing the data types of each column

#### Dataset Structure

The original dataset comprises of 21 rows and 33 columns. In the transformed dataset, 649 rows and 4 columns. The dataset contains no missing values.

#### Dataset Documentation

Collected and presented by the ONS, the dataset holds full documentation and collection methodology. See Appendix A for further detail.

### Data Preparation

A screenshot of a computer

Description automatically generatedThe dataset, initially in Excel from the ONS website, was transformed with Microsoft Power Query. Data was un-pivoted for Python compatibility, and cross-field validation ensured metric alignment. Uniqueness checks confirmed each industry matched a distinct year, detecting duplicates and missing values. Column names were updated for clarity.

Figure 2 - A screen capture showing the data source in its original format

A screenshot of a computer

Description automatically generated

Figure 3 - A screen capture showing the data source following transformation

## Methodology

### Model Selection

Time Series Analysis was selected for accurate predictions due to the dataset's time-dependent nature, despite some limitations like data variability. It effectively identifies patterns and trends, supported by confidence markers.

To remove user bias, a range of forecasting methods were considered to ensure the most accurate predictions for the dataset:

A white rectangular box with black text

Description automatically generated

Figure 4 - A Table showing the model options trialled for the forecast

After evaluating all industry groups, ARIMA was found to be the most suitable model and was used for forecasting. To enhance usability, the top three industry groups by volume were highlighted, and others were consolidated into an "OTHER" category.

### Implementation

#### Modelling

To begin this phase of the project, the data was imported from the transformed Excel file into a Google Collab notebook for analysis using Python. The full code can be found on GitHub [[Data-Science-Portfolio/C02EmissionsForecast - Python Code.ipynb at main · Amy-Vizard-Lovett/Data-Science-Portfolio (github.com)](https://github.com/Amy-Vizard-Lovett/Data-Science-Portfolio/blob/main/C02EmissionsForecast%20-%20Python%20Code.ipynb)]

### Evaluation Metrics

Here’s a detailed assessment of the ARIMA model results for each industry group:

A table with numbers and letters

Description automatically generated

Figure 5 A table showing the ARIMA Model outputs for each Industry Group

**OTHER:**

MAE and MSE: Both metrics indicate good accuracy in the forecast with relatively low error.

R²: High R² value suggests the model explains a large proportion of the variance in the data.

Forecast Trend: The forecast values show a stable trend, which is consistent with the historical data.

**Group C:**

MAE and MSE: Low MAE and MSE values show the model's predictions are quite accurate.

R²: Indicates a strong model fit, although slightly lower compared to “Other.”

Forecast Stability: The forecast values are very stable, which may reflect a consistent historical trend.

**Group D:**

MAE and MSE: The high MAE and MSE suggest the model is not as accurate for this industry group compared to others.

R²: While still good, the lower R² indicates that the model may not capture all the variance as effectively as for other groups.

Forecast Values: The forecast is stable but might be less reliable due to higher errors.

**Group H:**

MAE and MSE: The model has moderate accuracy, with lower error compared to "D" but higher than "Other" and "C."

R²: Shows a strong fit, though less impressive than "Other."

Forecast Values: The forecast is quite stable, aligning with historical trends.

**Summary:**

Best Model: The ARIMA model performs well for the industry group "Other," with the lowest MAE and highest R². It also does reasonably well for "C" and "H."

Model Challenges: The model struggles with industry group "D," as indicated by high error metrics and lower R².

**Model Recommendations:**

From the metrics, refining the ARIMA model or exploring other models (like SARIMA or GARCH) for the "D" group would be suggested. For industry groups with stable trends ("C" and "H"), ARIMA performs adequately, and forecasts are consistent.

### Further Analysis

#### Trend Analysis

A graph of different types of sales

Description automatically generated with medium confidence

Figure 6 - Four plots highlighting trends and seasonality within the dataset

**Original Series:** Displays raw data over time, serving as the baseline for analysis.

**Trend Component**: Shows a reasonable downward trend that aligns with the visual pattern in the original series.

**Seasonal Component:** Is flat and centred around zero, indicating no detected seasonal pattern. This could mean the data lacks seasonality, or the decomposition method used was not suitable.

**Residual Component:** Also, flat and centred around zero, suggesting the model might be good if the residuals are white noise. However, this flatness could also indicate overfitting or that the trend has captured all the information.

**Strengths:** The trend component appears appropriate, and the lack of strong residual patterns might indicate a good model fit.

**Potential Issues:** The flat seasonal and residual components could mean there's no seasonality or that the decomposition method didn't detect it effectively.

#### A comparison of a graph Description automatically generated with medium confidenceCorrelation Analysis

Figure 7 - Two correlation plots

**The Autocorrelation Function (ACF)** **plot**: Reveals strong influence of past values on future ones, with a gradual decline indicating potential non-stationarity and the need for differencing. **The Partial Autocorrelation Function (PACF) plot:** Shows significant spikes at the first three lags, suggesting an ARIMA model with three lags (p = 3) could be appropriate.

Overall, the analysis indicates possible non-stationarity, recommending differencing (d > 0) for ARIMA modelling. The plots guide the initial model choice of ARIMA (3, 1, 0), with further fitting and validation needed for accuracy.

#### Error Analysis

A screenshot of a graph

Description automatically generated

Figure 8 - heatmap visualising the forecast errors across different industry groups

In 2024, Industry group "D" shows a significant overestimation with a positive error of 1.4, while the "Other" category has a notable underestimation with a negative error of -1.7. Most other errors are relatively small, particularly as the model forecasts further into the future, where errors decrease.

**Overall Trend:** The forecast accuracy varies across different years and industry groups, with some years showing higher discrepancies. However, errors tend to decrease over time, suggesting either an improvement in the model's predictions or increased predictability in the dataset.

**Conclusion:** The significant errors, particularly in 2024 for "D" and "Other," highlight potential reliability issues in the model's forecasts for those points. Although the model's performance improves over time, further refinement is needed to enhance reliability, especially for critical decision-making scenarios.

## Project Findings

#### Conclusion on Model Reliability

The model demonstrates varying levels of reliability across different industry groups. For "Other" and "Group C," the model performs well, with low Mean Absolute Error (MAE) and Mean Squared Error (MSE), high R² values, and stable forecast trends. This indicates that the model is well-suited for these groups, effectively capturing the underlying data patterns and providing accurate predictions.

However, for "Group D," the model's reliability is notably lower, as evidenced by higher MAE and MSE values and a comparatively lower R². This suggests that the model struggles to fully capture the variability and trends within this industry group, leading to less accurate forecasts. "Group H" falls somewhere in between, with moderate accuracy and a strong fit, though not as robust as "Other" and "C."

#### Next Steps

**Model Refinement for Group D:** Investigate data characteristics and adjust the model for "Group D" to address high error rates. Experiment with alternative models or additional features to improve accuracy.

**Ensemble Methods:** Use ensemble forecasting techniques to combine predictions from multiple models, enhancing accuracy and reliability for underperforming groups like "D."

**Sensitivity Analysis**: Conduct sensitivity analysis to identify factors impacting forecast accuracy, enabling targeted improvements and better model tuning.

**Validation and Testing:** Perform out-of-sample testing and cross-validation to ensure model robustness and generalisability across different datasets, particularly for less reliable industry groups.

**Stakeholder Feedback:** Gather feedback from stakeholders to validate forecast usefulness and prioritise areas for refinement. As suggested by Walbaum (2024): *“Involve stakeholders in the model development from the get-go to get them comfortable and address any concerns early on.”*

Addressing these steps will improve model reliability and accuracy, leading to more confident decision-making.

## Conclusion

The research question regarding the UK's progress towards meeting the fourth emissions budget remains inconclusive due to limitations in the forecasting model used. The analysis demonstrated several challenges:

**Model Accuracy**: The ARIMA model used for forecasting showed high forecast errors, such as MAE and MSE, and inconsistent performance across industry groups, resulting in uncertainty in the forecasts. However, it was still considered the best available option, indicating that further refinement is needed to achieve more reliable forecast accuracy.

**Data Limitations**: The historical data available was limited in terms of granularity and coverage, impacting the model's ability to capture accurate trends.

#### Recommendations for Future Research

To address these limitations, the following improvements are recommended:

**Utilise Advanced Models:** Explore advanced forecasting techniques and machine learning to enhance accuracy.

**Enhance Data Collection:** Gather more recent and comprehensive data to improve the model.

**Refine Methodology:** Reassess model assumptions, add relevant variables, and consider individual models for each industry group. Suggested models include SARIMA, Random Forest, and XGBoost.

**Implications:** Current analysis indicates areas for further investigation and provides a foundation for more detailed analysis.

# Bibliography

**Azaria, N** (2023) *A Comprehensive Guide to Mean Absolute Percentage Error (MAPE)***.** Aporia. Available from:[A Practical Guide to Mean Absolute Percentage Error (MAPE) (aporia.com)](https://www.aporia.com/learn/a-comprehensive-guide-to-mean-absolute-percentage-error-mape/) [Accessed 15/08/2024]

**DSENZ**, Department for Energy Security & Net Zero (2024[a) *Background quality report.* Department for Energy Security and Net Zero. Available from: UK territorial greenhouse gas emissions background quality report (publishing.service.gov.uk)](https://assets.publishing.service.gov.uk/media/667bf005aec8650b10090057/ghg-emissions-background-quality-report.pdf) [Accessed 15/08/2024]

**DSENZ**, Department for Energy Security and Net Zero (2024b) *2022 UK Greenhouse Gas Emissions, Final Figures*. Department for Energy Security and Net Zero. Available from: [2020 UK Greenhouse Gas Emissions, Final Figures (publishing.service.gov.uk)](https://assets.publishing.service.gov.uk/media/65c0d15863a23d0013c821e9/2022-final-greenhouse-gas-emissions-statistical-release.pdf) [Accessed 15/08/2024]

**Foote, K** (2023)*The Impact of Poor Data Quality*. Dataversity.Available from:[The Impact of Poor Data Quality (and How to Fix It) - DATAVERSITY](https://www.dataversity.net/the-impact-of-poor-data-quality-and-how-to-fix-it/) [Accessed 16/08/2024]

**Tableau** (2024) *Time Series Forecasting: Definition, Applications, and Examples.* Salesforce Inc. Available from: [Time Series Forecasting: Definition & Examples | Tableau](https://www.tableau.com/learn/articles/time-series-forecasting#:~:text=Time%20series%20forecasting%20occurs%20when,drive%20future%20strategic%20decision%2Dmaking.) [Accessed 15/08/2024]

**Walbaum, T** (2024) *How to Maximize Your Impact as a Data Scientist.* Medium. Available from: [How to Maximize Your Impact as a Data Scientist | by Torsten Walbaum | Towards Data Science](https://towardsdatascience.com/how-to-maximize-your-impact-as-a-data-scientist-3881995a9cb1#:~:text=Impact%20doesn%27t%20materialize%20automatically,close%20the%20loop%20with%20stakeholders.). [Accessed 16/08/2024]

# Appendix

### Section A – Data Science Project

#### Data Documentation

The dataset used in this was collected and compiled by the Office of Nation Statistics (ONS) titled ‘Atmospheric emissions: greenhouse gases by industry and gas’. The version this project uses was released on 05/06/2024. The data was compiled by combining “data from the UK’s GHG Inventory with data from several other sources, including local energy consumption statistics, to produce a nationally consistent set of greenhouse gas emissions estimates at local authority level.” (DES&NZ, 2024a).

#### Data Quality

A Background Quality Report, publish by the Department for Energy Security & Net Zero, outlines that the data sourced for this project developed inline “with methods defined within international guidance provided to all countered via the IPCC [Intergovernmental Panel on Climate Change]” (DES&NZ, 2024b), a credible source used by the United Nations educate policy makers.

#### Methodology

The code used in Python to create the time series forecast can be found on GitHub, alongside an additional copy of Section A.

GitHub Code: [Data-Science-Portfolio/C02EmissionsForecast - Python Code.ipynb at main · Amy-Vizard-Lovett/Data-Science-Portfolio (github.com)](https://github.com/Amy-Vizard-Lovett/Data-Science-Portfolio/blob/main/C02EmissionsForecast%20-%20Python%20Code.ipynb)