Sahan Chowdhury data science project

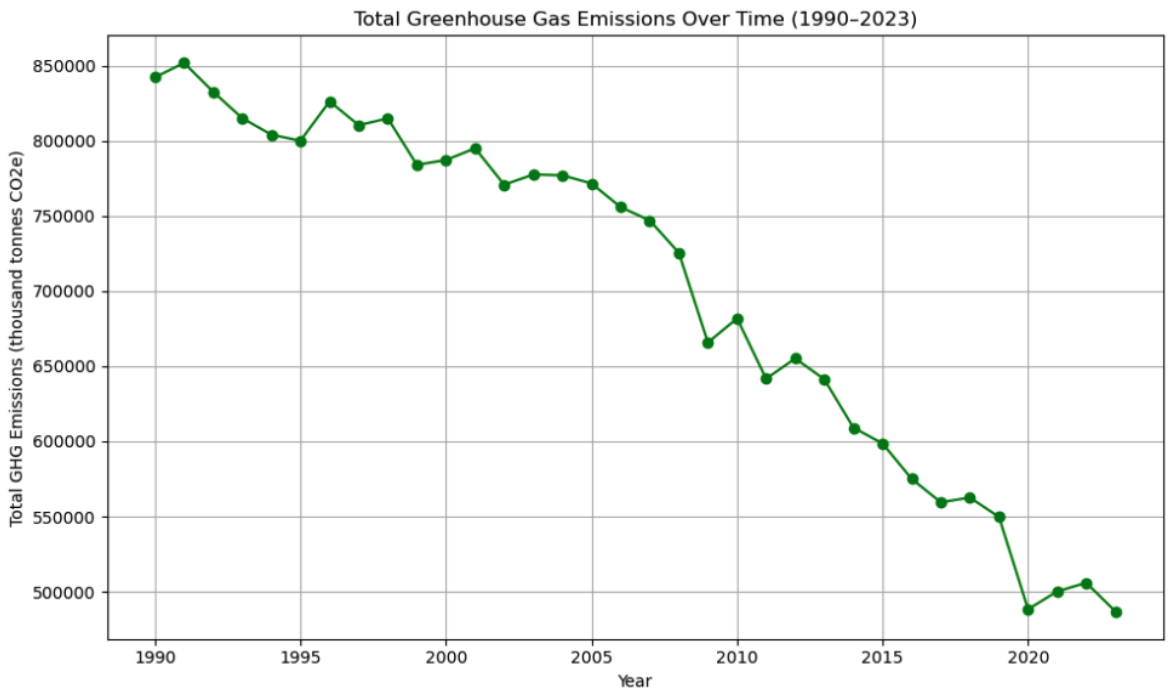
Investigating historical trends and future projections of UK Greenhouse Gas Emissions

1. The data

The dataset used in this project is the 'GHG total' sheet from the ONS atmospheric pressure dataset, covering emissions across 23 industries from 1990 to 2023. Descriptive statistics showed varying standard deviations between sectors, indicating that standardisation was needed for the machine learning models. Box plots and Z-scores were used to identify outliers, and although some extreme values were detected, they were retained due to their valid explanations. A threshold z score of +-2 was used.

2. Data Analysis

From figure(1) its evident emissions have decreased, as of 2023 total emissions are below 500,000 compared to 1990 when it was above 800,000



Figure(1): line graph of total emissions over time

A line plot was created figure(2) for each industry from table1 (see jupyter notebook) using a log scale to accommodate the wide range of emissions. Most industries showed a decline in emissions since 1990, with notable reductions in sectors like 'Electricity, gas, steam and air conditioning supply' (67.77%) and 'Water Supply' (66.98%). These declines were driven by regulatory changes, technological advancements, and shifts to renewable energy. For instance, as of 2024, 42% of UK energy comes from renewable energy (Cladco, 2024).

On the contrary, the 'construction' sector had an increase since 1990 of 37.10%. This can be explained due to the national growth in infrastructure development and urbanisation. As stated by the UN, materials used in construction, such as cement, steel, and aluminium, have a significant carbon footprint. Industries directly impacted by the effects of urbanisation, such as real estate and transport, also saw surges in emissions, particularly noteworthy was a single-year increase for 'transport and storage' of 39.56% alone in 2022.

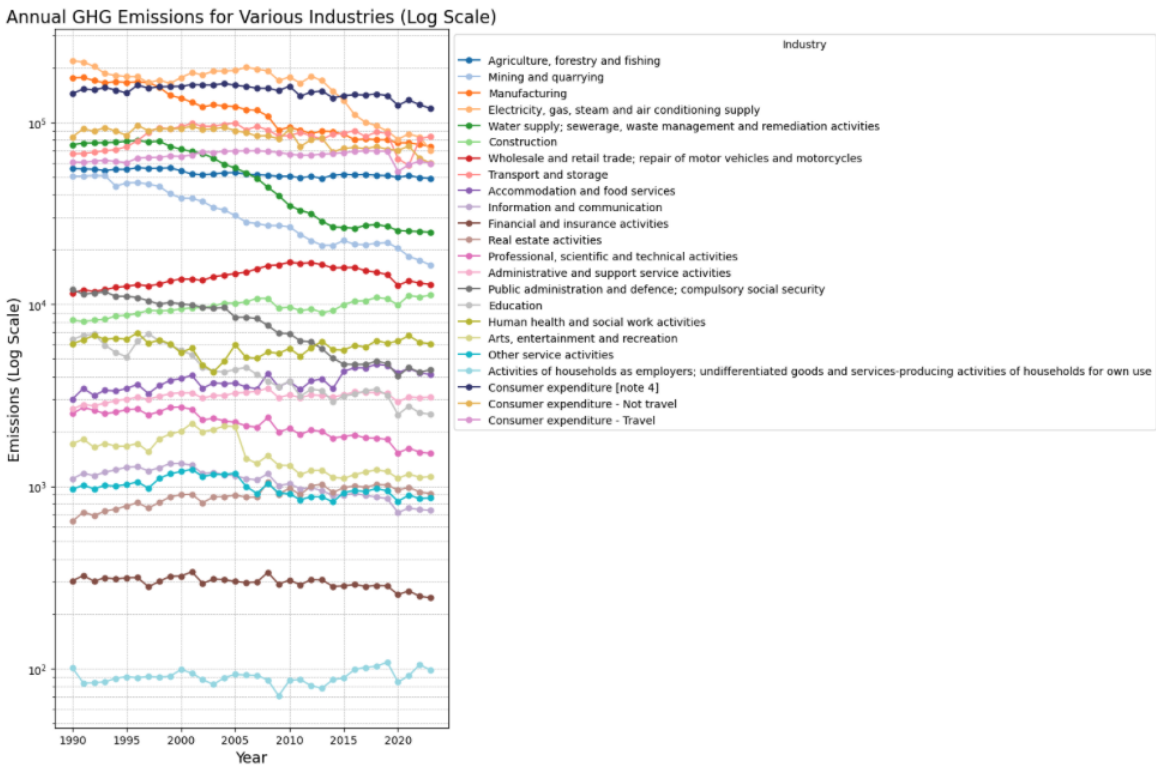


Figure (1): Line graph of emissions for different industries

Table 2 was used to create stacked bar charts (see Jupyter Notebook) to identify the main contributors to emissions for each sector. For instance, for the Transport and storage sector, air transport accounted for most of the emissions, with a notable increase over time. For electricity production, coal has significantly decreased over the years. However, the use of gas has increased, and for water supply, emissions were mainly from treatment and disposal. A correlation matrix was computed, and the strongest correlation was observed with the sectors mining and quarrying and manufacturing with a Pearson correlation coefficient of 0.98, implying that the models could treat these sectors together, while real estate showed the lowest of -0.88 with both 'manufacturing' and 'mining and quarrying'.

3. Model findings and visualisations

Initially, I explored several models based on similar studies to find the best approach for predicting greenhouse gas emissions. The first model I tested was linear regression, which is a method that models the relationship between variables by fitting a straight line to the data. It's simple, easy to understand, and works well when the relationship between variables is linear. The study by Ozdemir, Pehlivan, and Melikoglu (2024) showed that linear regression is a successful method for predicting emissions, achieving a low root mean squared error (RMSE) of less than 0.2036, indicating its effectiveness.

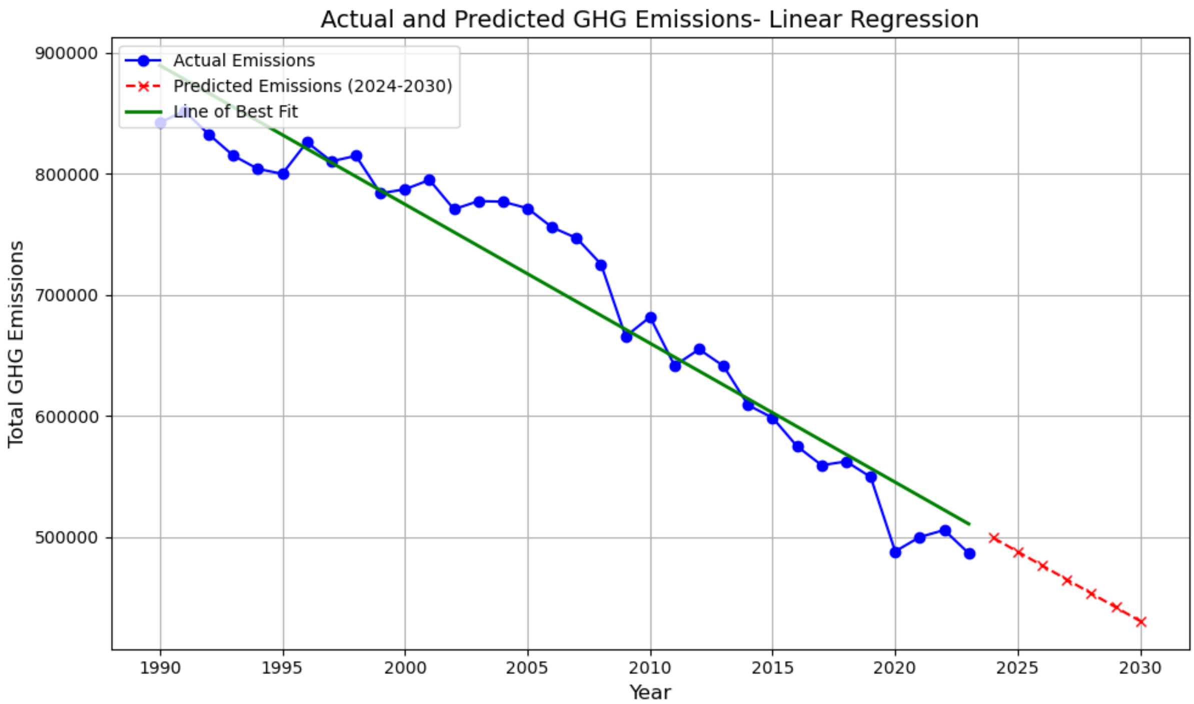
Next, I chose support vector regression (SVR). SVR is a machine learning method that works by finding the best hyperplane in a high-dimensional space to predict the target variable. It is particularly useful for small datasets and non-linear patterns. According to Qin et al. (2022), SVR works well when there are limited data points, making it a good choice for this project. In their study, SVR produced an RMSE of 37.43 mt, which showed that it could accurately predict emissions with small sample sizes.

After fitting the models, I evaluated their performance by calculating three key metrics: R-squared (R^2), mean absolute error (MAE), and mean squared error (MSE) as shown in the table below.

	Linear Regression	Support Vector Regression
R^2	0.9329	-0.1853
Mean Absolute Error	25031.9173	100198.8564
Mean Squared Error	911108081.9840	16095826593

Table(1)

The linear regression model achieved a high R^2 value, it also had a relatively low mean absolute error and squared error compared to Support Vector regression. As a result, the linear regression proved to be a suitable choice for this project.



Figure(2): predicted model

The predictions for total greenhouse gas (GHG) emissions in the UK from 2024 to 2030 suggest a continued decline in emissions over the coming years, by 2030, total emissions are expected to be 430,726.39.

These results indicate a steady reduction in emissions, which aligns with the UK government's commitment to achieving net zero emissions by 2050. This decrease could reflect the effectiveness of ongoing policies such as the shift towards renewable energy sources and stricter regulations on industries with high emissions, like electricity and manufacturing.

For the UK government, these predictions imply that their current strategies are on track to reduce emissions over the next decade, but further efforts may be needed to ensure more reductions in sectors like construction and transport, which have seen recent increases in emissions. The government could use these forecasts to adjust policy, especially in areas where emissions are not declining as quickly.

4. Conclusion

In conclusion, this project successfully analysed UK greenhouse gas emissions across various sectors using linear regression to predict future trends. While the model performed well, it assumes a linear relationship across all sectors, which may not fully capture the diverse emissions patterns in sectors like construction and transport. To explore non-linear patterns, I also tested Support Vector Regression (SVR), but it did not perform as well due to the small sample size. The project identified key sectors where emissions have decreased, such as electricity and water supply, while sectors like construction and transport have seen increases.

Future improvements to the model could include sector-specific predictions and the integration of additional variables, such as economic activity and population growth, to provide more accurate long-term forecasts. Furthermore comparisons of more models with tuned parameters could also enable to find the most optimal model, while looking into sectors could help find ways to reduce emissions.