Model Summary

Introduction

In recent years, the problem of climate change has been on the forefront of many governments of nations around the world. To tackle this problem, many countries including the United Kingdom have pledged to reduce their greenhouse gas (GHG) emissions to tackle climate change. Specifically, the UK government have committed in legislation to reducing their GHG emissions to net zero by 2050.

Brief Summary of Data

To assess whether the UK are taking the right action to reach this goal, I used the current edition of the “Atmospheric emissions: greenhouse gases by industry and gas” dataset provided by the Office for National Statistics to create a model that predicted the total GHGs emitted by the UK for the next five years.

This dataset contained eight spreadsheets. Seven of which each contained annual greenhouse gas emissions from 1990 to 2021 for seven different GHGs and the last spreadsheet contained the annual GHG emissions for all the GHGs. Each of the spreadsheets had two tables. Both tables contained data on the emissions of GHGs where one table separated GHGs emitted by industry section and the other table separated GHGs emitted by industry group within the industry sections.

As the goal was to predict total GHGs emitted, the spreadsheet containing total GHGs was the most beneficial. Also, the goal was to evaluate the UK, so I chose the table focusing on industry sections rather than industry group as it has less data unnecessary for this model.

Data Processing/Cleaning

For data processing, I used the Pandas library in Python to parse through the table and convert it into a Pandas DataFrame. In this table, there was a column containing an alphabetical index (A-T), an empty column, an empty row, and a missing column name for the column of industry sections. I removed the alphabetical index column and the empty column, then set the name for the column of industry sections to “Section” while setting it to be the index column and then removed the empty row.

A graph of gas emissions

Description automatically generated

Figure 1: Visualisation of annual greenhouse gas emissions 1990-2021.

Modelling Technique

To decide which modelling method I wanted to use, I created a visualisation of the annual GHGs emitted in Figure 1. This gives us a very vague idea on how to model GHGs, so I used the Seaborn package to make use of its residual graph function.

A graph with blue dots

Description automatically generated

Figure 2: Visualisation of the difference between GHG emissions and its best-fitting linear model.

In figure 2, I could see that a single linear model wasn’t satisfying the data. By the clear peak and linear gradients of the graph, the data prompted a hypothesis that the data could be modelled with two separate linear models. One linear model before 2008 and the other linear model for 2008 and after.

To test this hypothesis, I decided to use Sci-Kit Learn as it contains methods for creating linear models and model validation techniques. I used a train test split which is a method for randomly splitting the data into a dataset for training and a dataset for testing. By using a linear regression model created from the training data, I predicted the testing dataset and evaluated statistics between the predicted testing dataset and the real testing dataset. I did this for both the “before 2008” dataset and the “2008 and after” dataset. After evaluating the coefficient of determination and the root mean squared error for both, I concluded that my hypothesis was correct. This assumption would be the basis of my model.

For making a prediction on the GHGs emitted in the next five years, a secondary assumption had to be made. This assumption was that the model used for the “2008 and after” dataset could be used for the next five years. I could not use the same model that I used for validating the model because this model was created based on a random sample of the data instead of the population of the data. Therefore, I had to create a new model. Once again, I used Sci-Kit Learn’s linear modelling functions and used the resulting model to predict the next five years (2022-2026) of GHG emissions. Figure 3 shows the annual GHG emissions with these new predictions.

A graph with blue dots and white text

Description automatically generated

Figure 3: Visualisation of previous and provisional annual greenhouse gas emissions. 1990-2026

I also used the model to predict the GHG emissions during the year 2050 to make a conclusion on whether the UK government would reach their goal of net-zero emissions by 2050. The predicted GHG emissions in thousand tonnes in 2050 was 37911.5.

Analysis

The figures shown clearly indicate that there is a downwards trend of GHG emissions as we get closer to 2050. Therefore, there is evidence that the UK government are taking the right steps to achieve this. However, the model also predicts that the UK will not be net-zero by 2050 so the government needs to do more to reach this goal.

Limitations

My model is made up of two main assumptions. The first being that I could model the data with two linear models. The method I used to validate had extreme limitations as it involved randomly sampling a dataset with 18 points of data. Such a small number of data points make it difficult to validate this assumption.   
My second assumption was that I could use the second part of the model to predict the next five years, as well as the definitive answer to whether the UK government could reach the goal of net-zero emissions by 2050. Predicting the future is very difficult, but I could’ve reached these conclusions possibly a different method.

One possible method I could’ve used is by modelling each greenhouse gas separately then calculating the sum rather than modelling the sum of all.

Another similar method could be modelling each industry section separately and used the sum of these to reach my conclusions.

These methods would’ve likely reduced my chance of error and bias found in my model.