Problem Statement

As science has advanced, we have been able to measure Carbon Dioxide (CO_2) emissions for past years. Using global CO_2 emissions data, I developed an ARIMA model to predict CO_2 emissions for future years. This can allow corporations, countries and organizations to monitor their CO_2 emissions and potentially adjust their practices.

Data Wrangling

The raw dataset was obtained as a csv from the Global Carbon Project (Andrew, Robbie M.; Peters, Glen P.). The dataset had many empty entries and the types needed to be adjusted in order to be assessed properly. I also had to merge the dataset with a country list from Wikipedia in order to add the continents. Outlier data points were examined using a scatter plot. Besides missing values, no numerical data needed to be adjusted.

In order to reduce dimensionality, any features with 40% or more missing values were eliminated. I also dropped columns such as region and country ID codes. This made a cleaner data set that could be analyzed more efficiently.

Exploratory Data Analysis

With this cleaned dataset, I dove deeper into the relationships between the features. The feature I want to predict is the total CO_2 emissions. In figure 1, I graphed the total CO_2 emissions by year. The scatter plot shows that as time moves forward, CO_2 emissions increase, and start to increase at a higher rate around the 1920's

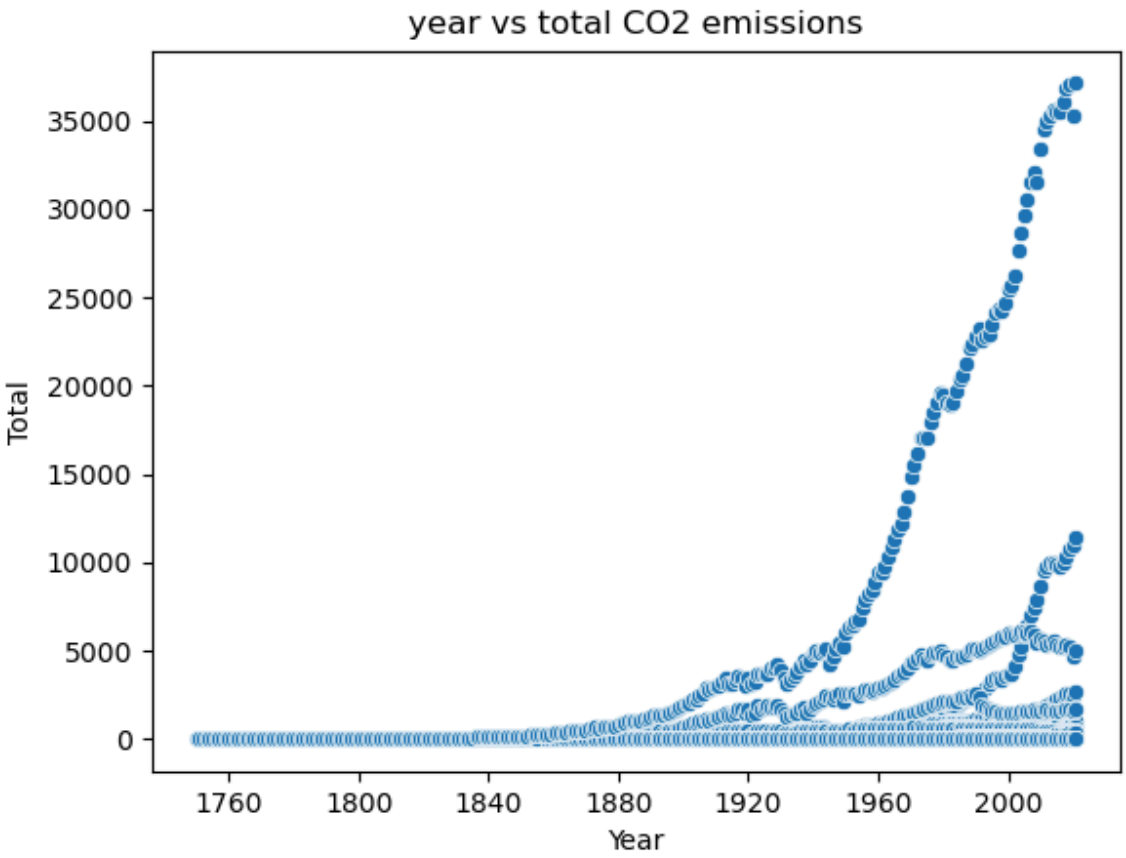

Figure 1

Next I made a similar scatter plot, and added hue to the continent. This is shown below in figure 2. This allowed me to differentiate between continents.

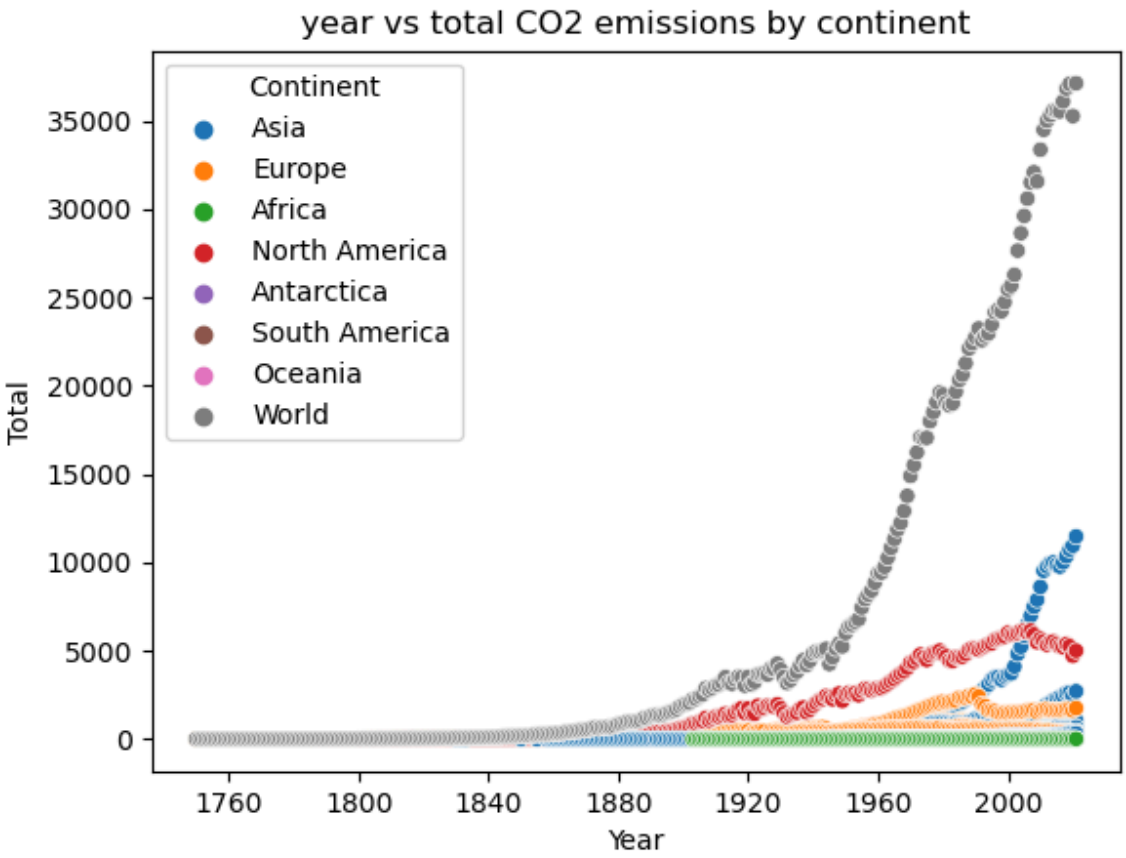

Figure 2

In figure 3, shown below, instead of total CO_2 emissions, I examined just coal CO_2 emissions by year. This scatterplot had a similar shape to the total CO_2 emissions graph. This finding makes me believe that total CO_2 emissions and the coal CO_2 emissions have a strong relationship.

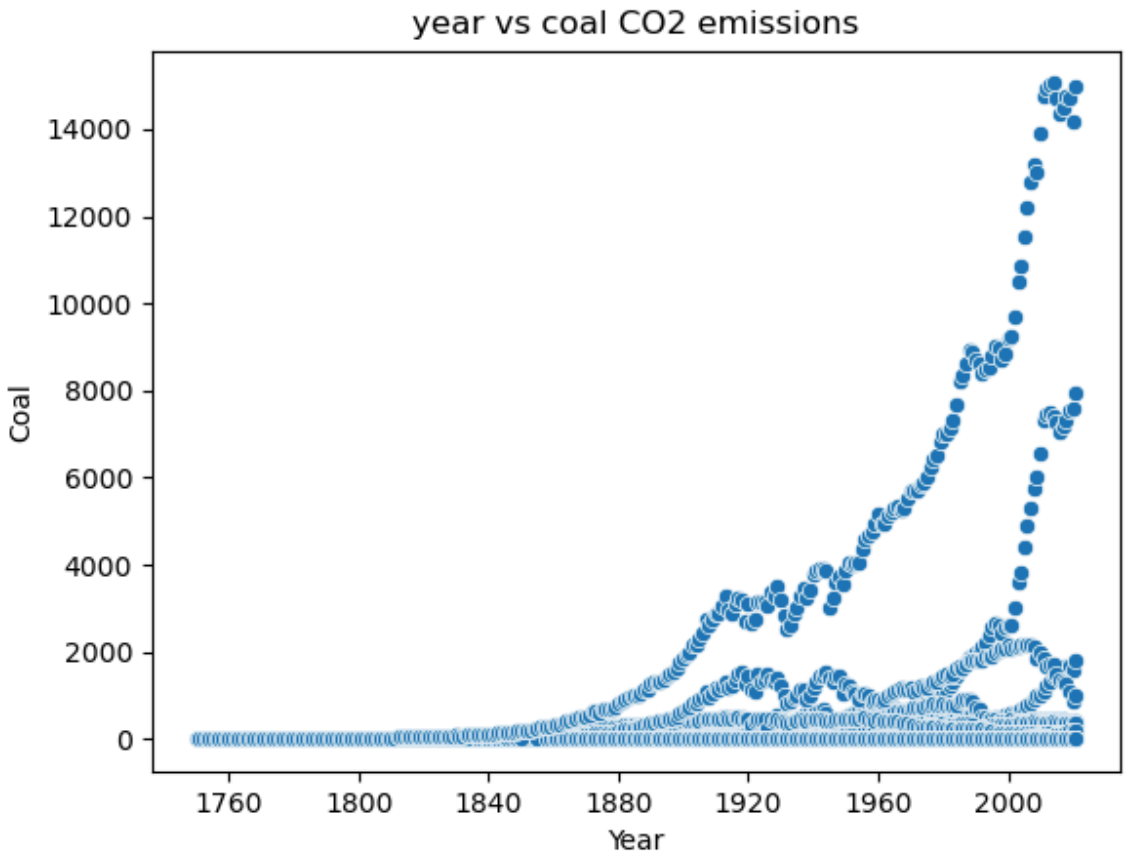

Figure 3

In figure 4, I wanted to see the relationship between the numerical features. So I created a heatmap of those columns. I can see that coal emissions and total emissions have a strong relationship, as do total emissions and oil emissions.

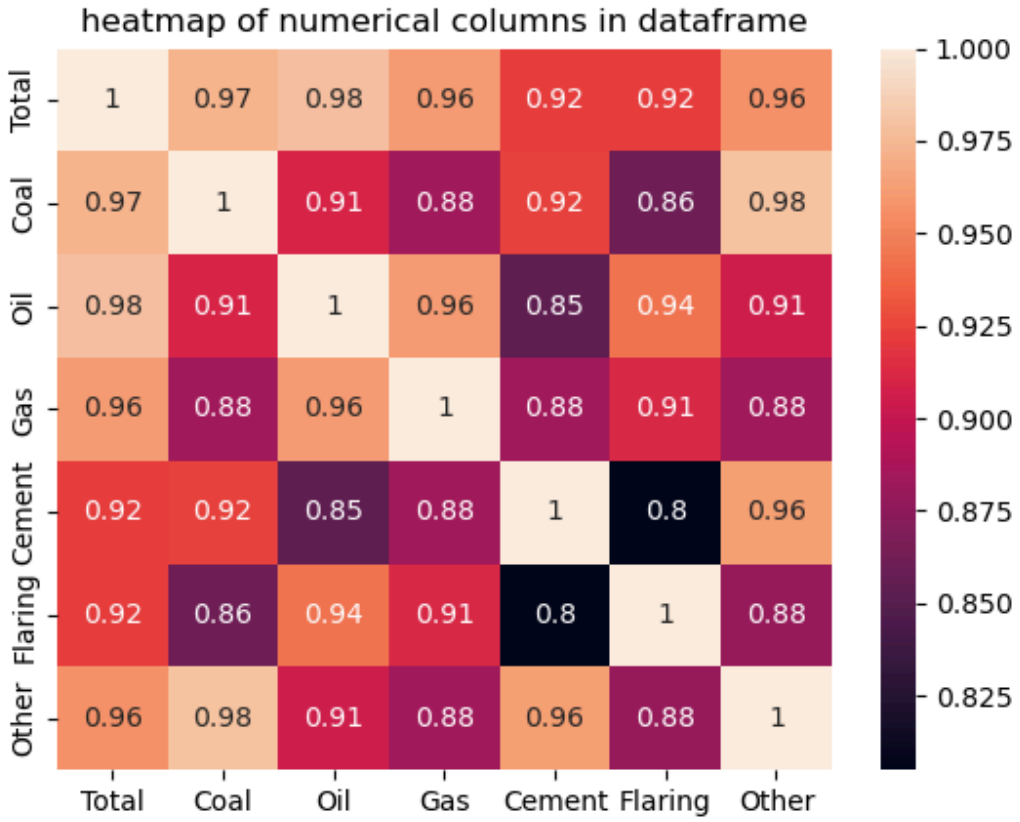

Figure 4

With these relatively strong relationships and the scatterplots showing the shape increasing as time goes by, I decided to look at this data as a time-series and analyze it as so.

In-Depth Analysis

As part of the time-series analysis, I decided to use the ARIMA model. Looking at the scatterplot of year vs total CO_2 emissions, the data is not seasonal. However, I did need to scale the data. So I decided to use a log scaling, and a standardscaler and see how each would affect the predictions. I also wanted to difference the data and check the autocorrelation and partial autocorrelation to find the order for the ARIMA model. As shown in figure 5, both the autocorrelation and partial autocorrelation drops off after 0.

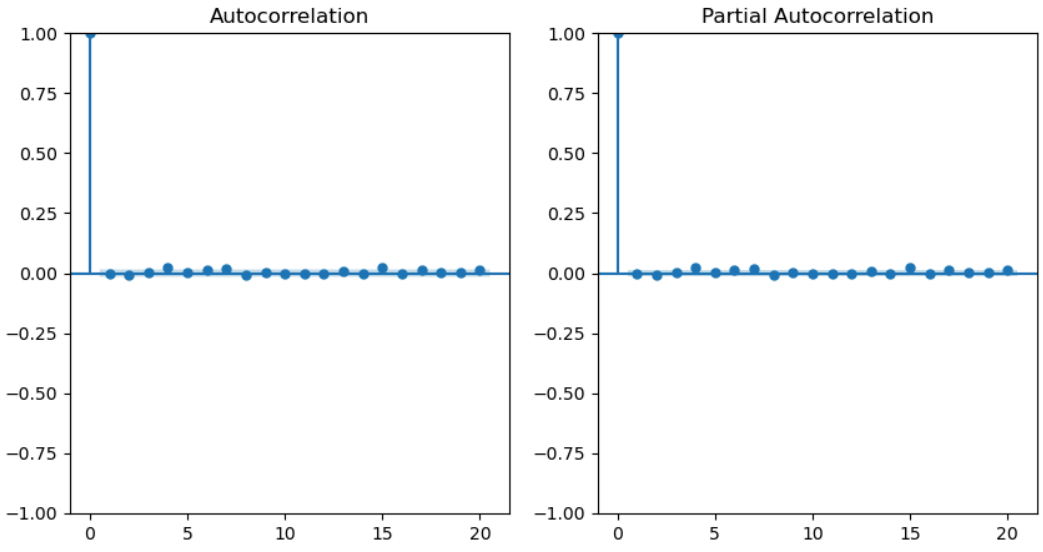

Figure 5

First, I tried the ARIMA model without scaling and used a time series split to test the model. Then I tried the scaled data, and then the log scaled data. For each of these, I found the Mean Squared Error, Mean Absolute Error and r^2. I then compared these to find the best model and then used AutoARIMA to find the best order for the ARIMA model. The standardscaler data had the best performance metrics. However, after looking at a plot for the model, as shown in figure 6, it is clear that the standardscaler model does not fit the data.

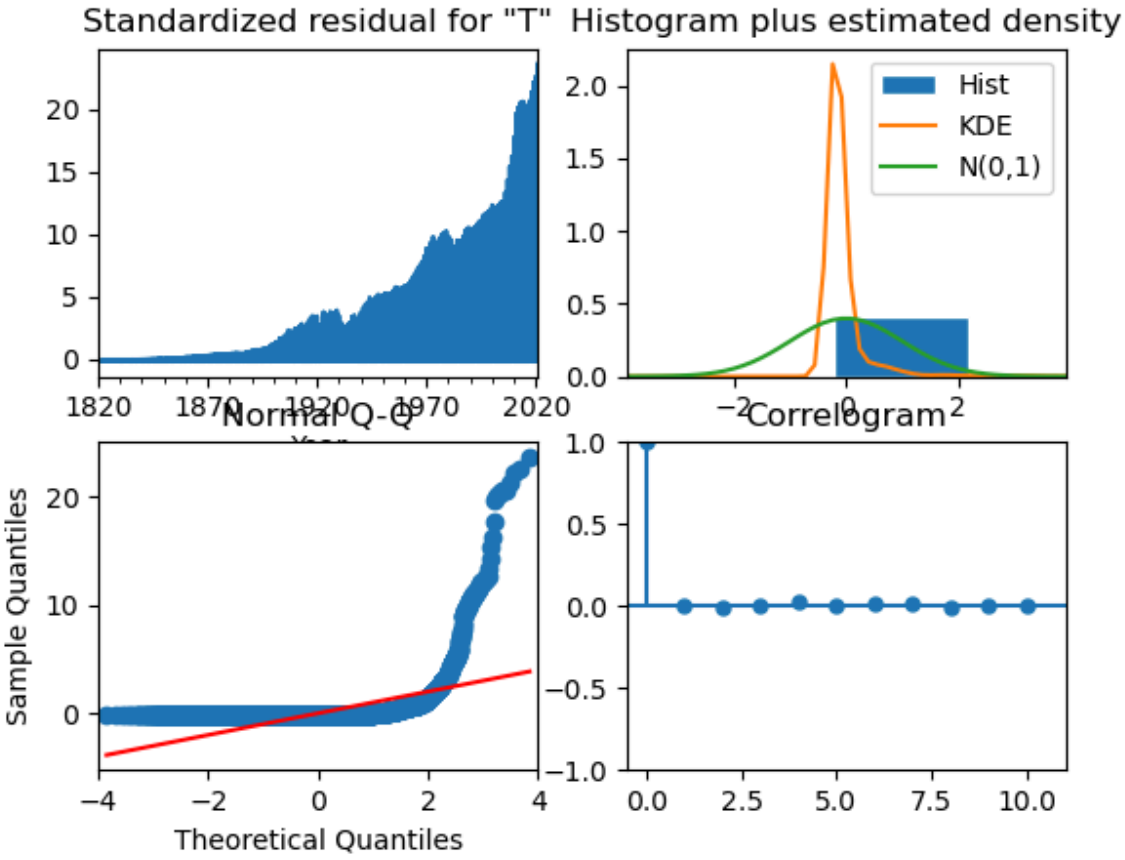

Figure 6

Figure 7, shown below is the plot diagnostics of the log scaled model, and it looks to fit the data much better than the standardscaler model. From here, I used the AutoARIMA model to see if any other model for this data would work better.

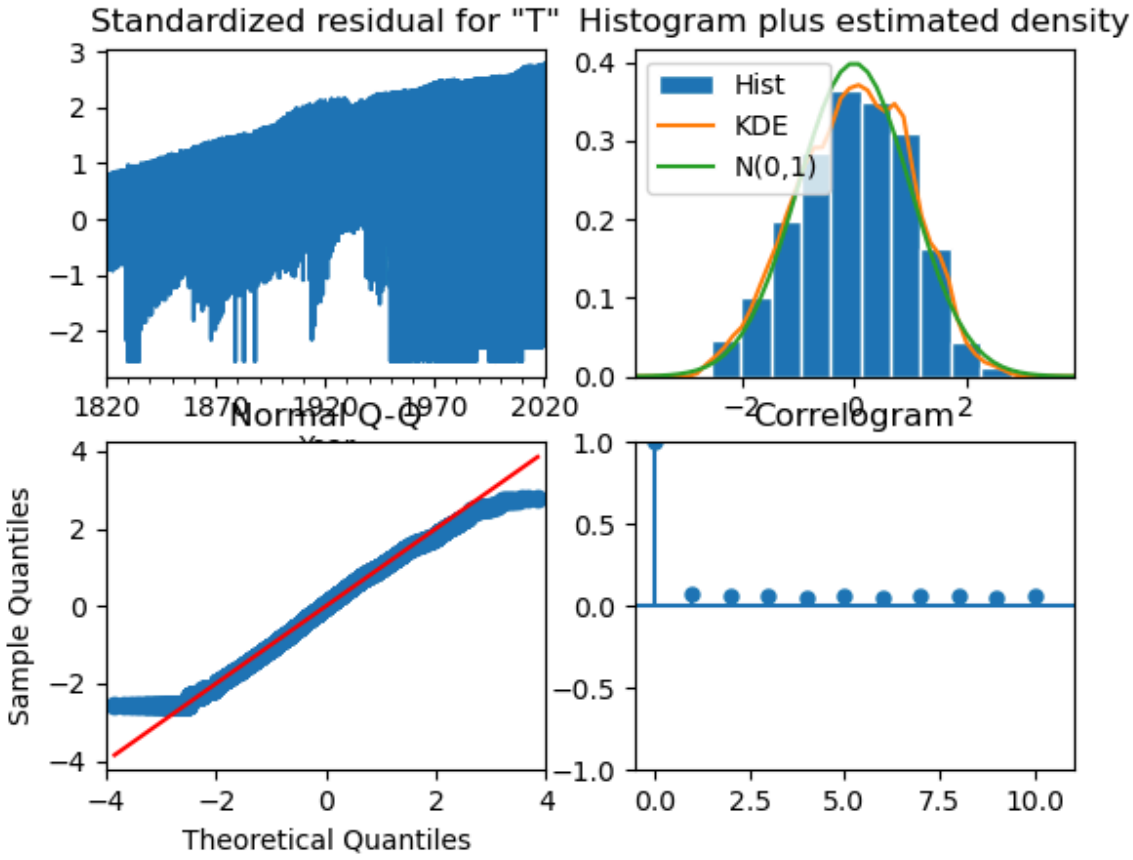

Figure 7

Model Selection

After using AutoARIMA's stepwise function to search for the best model, the results were the model ARIMA (2,0,2). So I used the time series split to check the performance metrics and compare to previous iterations, and sure enough the (2,0,2) order for the log scaled data had better mean squared error, mean absolute error, and r^2, than the (0,0,0) model I had tried earlier. Shown below in figure 8, is the plot diagnostics for the (2,0,2) model. The fit looks like there are no real outliers that would affect the metrics. Therefore the ARIMA (2,0,2) model with log transformation is the model that I have chosen for this data.

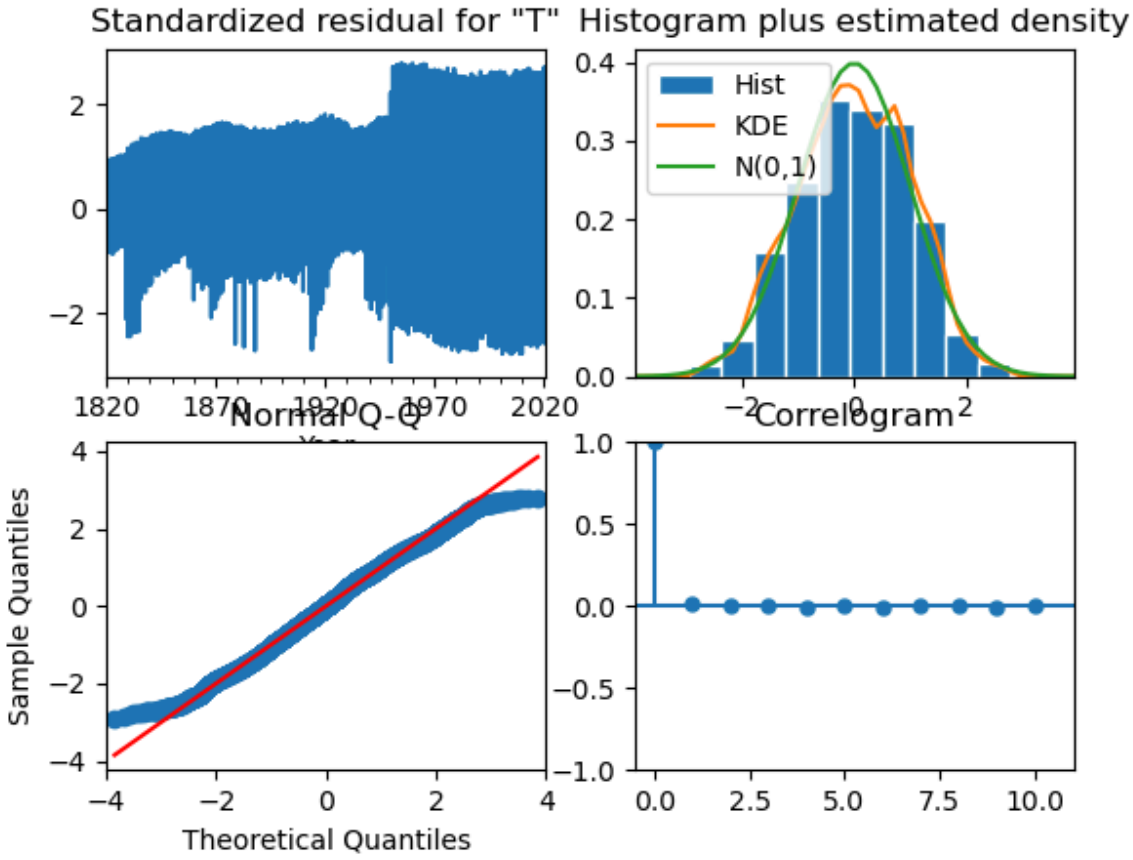

Figure 8

Recommendations on Further Use

Using this model, clients can compare the findings to compare to other greenhouse gas emission data in order to draw further conclusions on greenhouse gas data.

Clients can also use this data to model future CO_2 emission data to prepare for the road ahead. According to the Paris Climate Accords, many nations have agreed to try to keep the global surface temperature to below 2 degrees Celsius above pre-industrial levels. This model can help to show where the CO_2 emissions will be in the future and how the nations can adjust their policies to account for the emissions.

Clients could also use the findings to pair with other weather data to predict weather patterns in the future.

Future Research

Some ways this research can be continued is to model data for each continent to get a more representative model for each continent. Another way this can be continued is to account for the population of countries or continents to accurately predict emissions based on future population predictions.