Greenhouse Gas Contributors & Future Emission Predictions.

Quinn Maloney
Data Science Practicum 1
Regis University

Purpose of Analysis and Project Overview

Climate change is a very real crisis we are facing today and human emissions of co2 and other greenhouse gasses are a primary driver of climate change. I would like to dig deeper into this and determine what specifically is playing a role in these increased emissions and predict how these rates will change in years to come. Additionally, in order to effectively reduce emissions, it is important to figure out where these emissions are coming from. My goal in this practicum project is to determine the main contributors to greenhouse gas emissions and make future emission predictions.

Data Description and Preparation

The first dataset I used consisted of yearly emission values from the late 1700s to 2020, obtained from the *Our World in Data* website. This dataset consists of sixty column features for 248 countries/regions. Emission columns include 'co2', 'co2_per_capita', 'cement_co2', 'cement_co2_per_capita', 'coal_co2', 'coal_co2_per_capita', 'gas_co2', 'gas_co2_per_capita', 'oil_co2', 'oil_co2_per_capita', ‘methane’, ‘nitrous oxide’, etc. The large number of features allows us to predict future emission rates more accurately. The dataset consists of multiple data types.

The Quantitative Features consisted of: float and integer values for emission values and year of observation, respectively. Emission columns included 'co2', 'co2_per_capita', 'trade_co2', 'cement_co2', 'cement_co2_per_capita', 'coal_co2', 'coal_co2_per_capita', 'flaring_co2', 'flaring_co2_per_capita', 'gas_co2', 'gas_co2_per_capita', 'oil_co2', 'oil_co2_per_capita', 'other_industry_co2', 'other_co2_per_capita', 'co2_growth_prct', 'co2_growth_abs', 'co2_per_gdp', 'co2_per_unit_energy', 'consumption_co2', 'consumption_co2_per_capita', 'consumption_co2_per_gdp', 'cumulative_co2', 'cumulative_cement_co2', 'cumulative_coal_co2', 'cumulative_flaring_co2', 'cumulative_gas_co2', 'cumulative_oil_co2', 'cumulative_other_co2', 'trade_co2_share', 'share_global_co2', 'share_global_cement_co2', 'share_global_coal_co2', 'share_global_flaring_co2', 'share_global_gas_co2', 'share_global_oil_co2', 'share_global_other_co2', 'share_global_cumulative_co2', 'share_global_cumulative_cement_co2', 'share_global_cumulative_coal_co2', 'share_global_cumulative_flaring_co2', 'share_global_cumulative_gas_co2', 'share_global_cumulative_oil_co2', 'share_global_cumulative_other_co2', 'total_ghg', 'ghg_per_capita', 'total_ghg_excluding_lucf', 'ghg_excluding_lucf_per_capita', 'methane', 'methane_per_capita', 'nitrous_oxide', 'nitrous_oxide_per_capita', 'population', 'gdp', 'primary_energy_consumption', 'energy_per_capita', and 'energy_per_gdp'

The Qualitative features consisted of: categorical values for ‘country’ and ‘iso code’. When categorical values are needed for analysis, they will be transformed to dummy/indicator variables, but in most cases the only use for country is to divide the dataset into subsets based on country.

The second dataset I used consisted of two columns, ‘Sub-sector’ which were objects, and ‘Share of global greenhouse gas emissions (%)’. The data was obtained from the *World Resources Institute* and all data was from the year 2020. This dataset was useful in determining which production-based sectors were responsible for the most greenhouse gas emissions. It is important to know which productions are responsible for emitting the most greenhouse gasses so that we can begin thinking of new ways to reduce these emissions.

Exploratory Data Analysis (EDA) and Visualizations

Exploratory data analysis is the first step for any data analysis project. Since the target column for my predictions was the co2 column, I did most of my EDA on this column. Since I wanted to analyze co2 patterns and trends over the years for each country, I often divided my dataset, grouping rows together based on having the same country value.

Before doing any analysis on my data, I used EDA to see the distribution of co2 values for each country/region, what percentage of each column was missing data, and plotted to see if there were any visual trends in the data.

Figure 1 below represents the percentage of missing data in each column of the dataset. As we can see, there are quite a few columns missing over 50% of its data. Since the dataset has a large number of columns to begin with, I decided to drop columns that were missing over 80% of their data. These fifteen columns included 'trade_co2', 'flaring_co2', 'flaring_co2_per_capita', 'other_industry_co2', 'other_co2_per_capita', 'consumption_co2', 'consumption_co2_per_capita', 'consumption_co2_per_gdp', 'cumulative_flaring_co2', 'cumulative_other_co2', 'trade_co2_share', 'share_global_flaring_co2', 'share_global_other_co2', 'share_global_cumulative_flaring_co2', and 'share_global_cumulative_other_co2'.

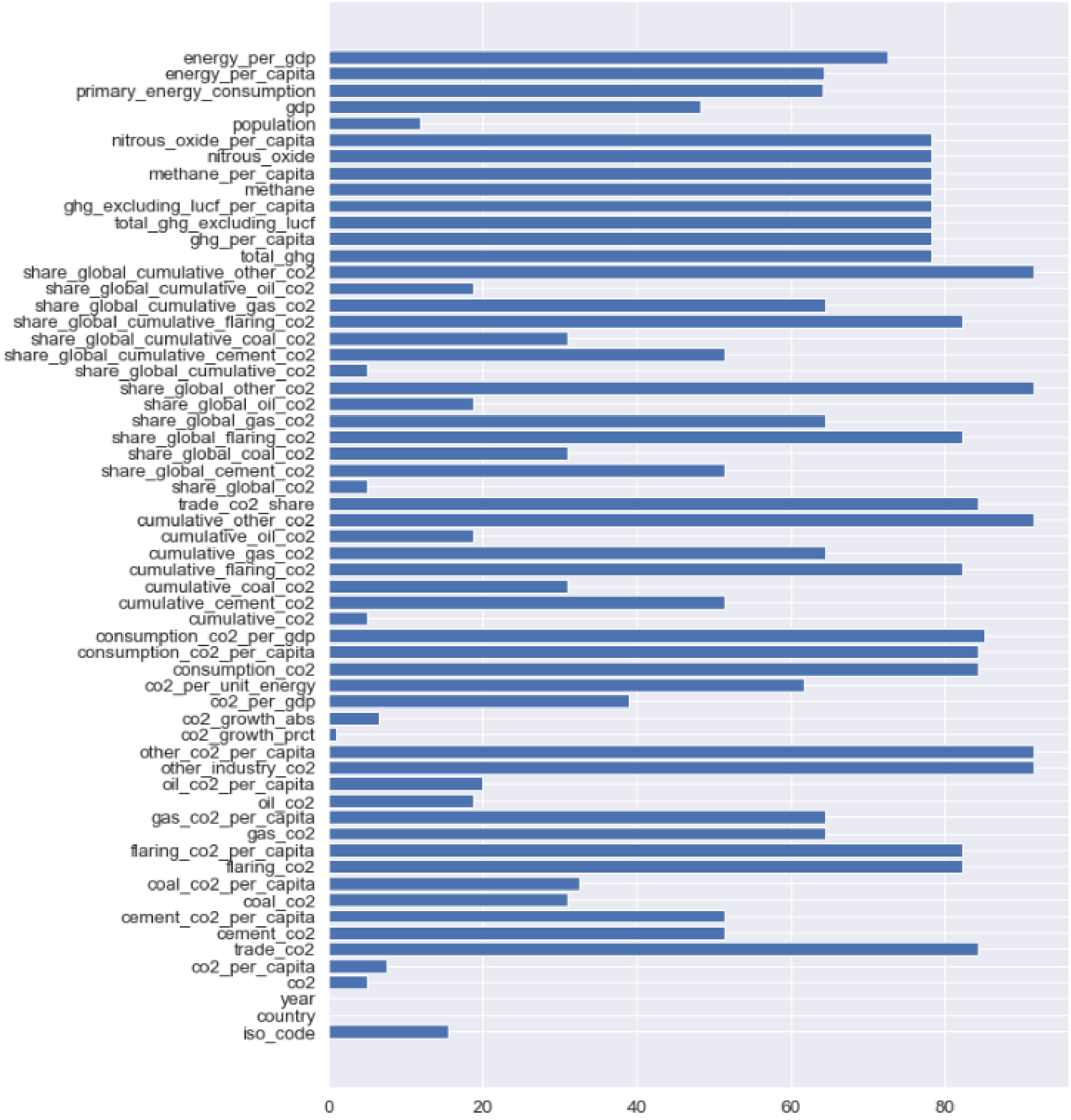


Fig. 1: Horizontal Bar Plot Showing the Percentage of missing values in each feature of the dataset.

Following this, I used Pearson's Correlation and Random Forest Regression to determine which features had the strongest correlation to co2. Figure 2 consists of two heatmaps. Figure 3 consists of two diagrams representing the outcome of the Random Forest Regression.

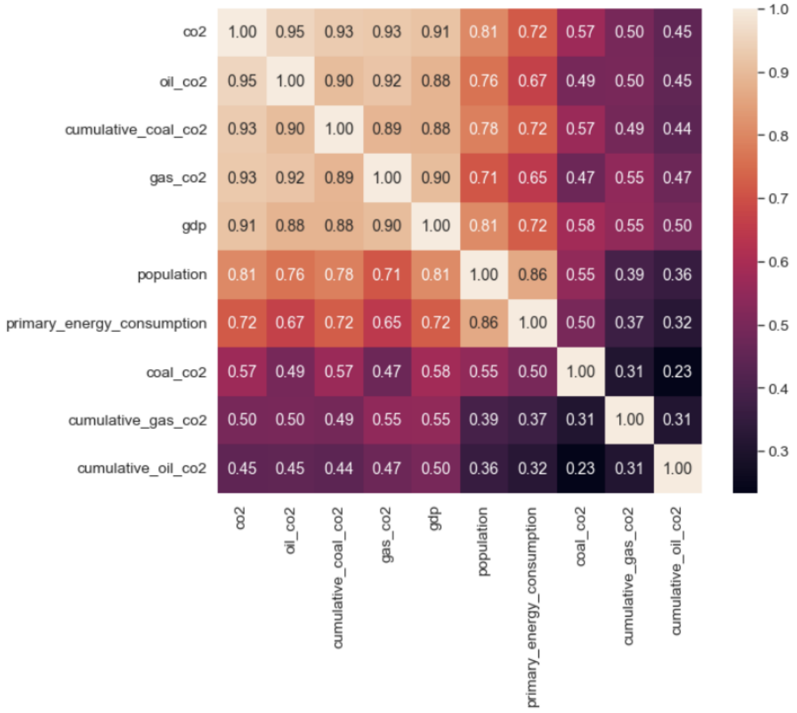


Fig. 2a: Correlation Heatmap Plot showing the ten most positively correlated features to our target feature(co2).

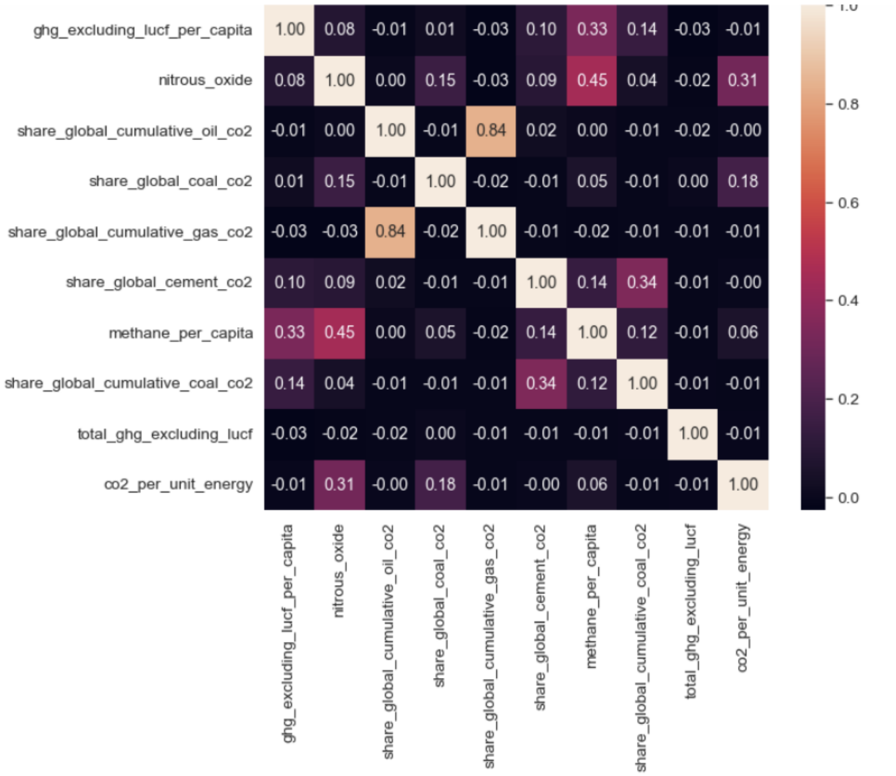
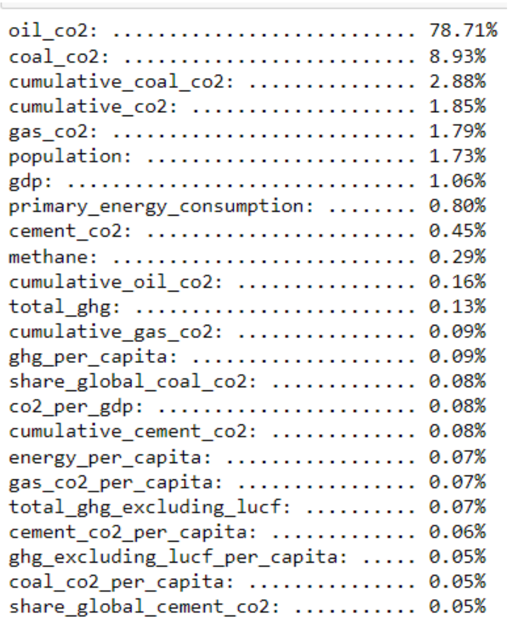


Fig. 2b: Correlation Heatmap Plot showing the ten most negatively correlated features to our target feature(co2).



```
oil_co2: .......................... 78.71%
coal_co2: ......................... 8.93%
cumulative_coal_co2: .............. 2.88%
cumulative_co2: ................... 1.85%
gas_co2: .......................... 1.79%
population: ....................... 1.73%
gdp: .............................. 1.06%
primary_energy_consumption: ....... 0.80%
cement_co2: ....................... 0.45%
methane: .......................... 0.29%
cumulative_oil_co2: ............... 0.16%
total_ghg: ........................ 0.13%
cumulative_gas_co2: ............... 0.09%
ghg_per_capita: ................... 0.09%
share_global_coal_co2: ............ 0.08%
co2_per_gdp: ...................... 0.08%
cumulative_cement_co2: ............ 0.08%
energy_per_capita: ................ 0.07%
gas_co2_per_capita: ............... 0.07%
total_ghg_excluding_lucf: ......... 0.07%
cement_co2_per_capita: ............ 0.06%
ghg_excluding_lucf_per_capita: .... 0.05%
coal_co2_per_capita: .............. 0.05%
share_global_cement_co2: .......... 0.05%
```

Fig. 3a: Random Forest Regression to determine feature importance for the co2 column

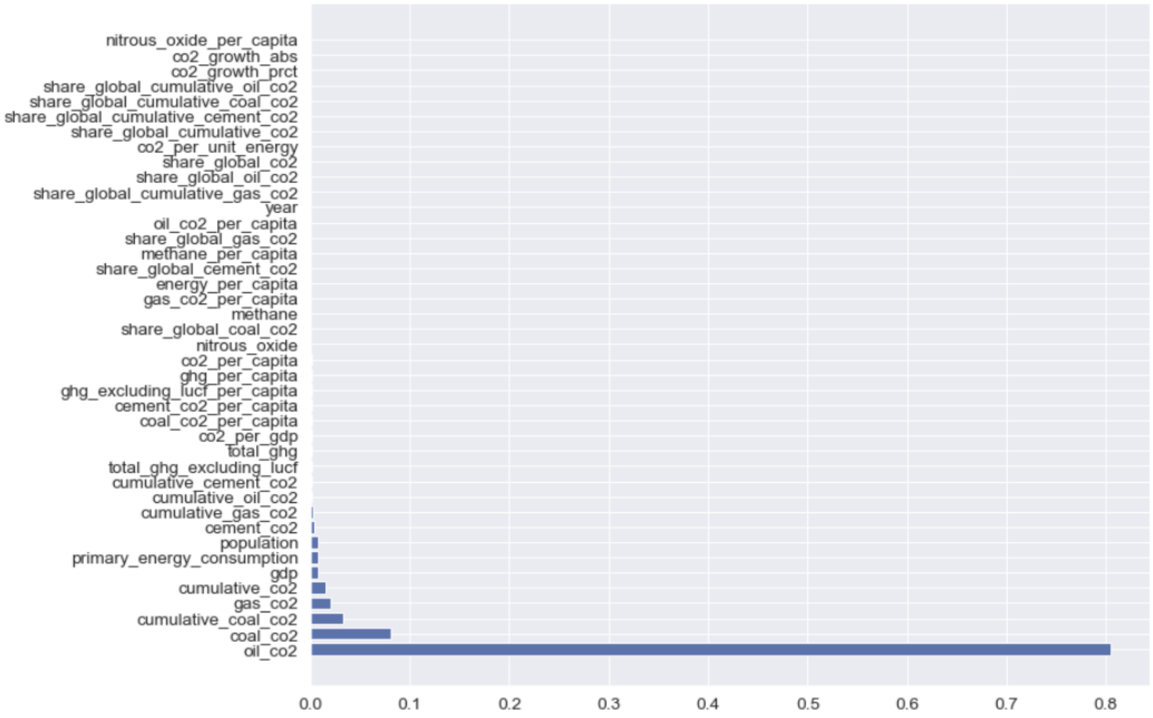


Fig. 3b: Bar Plot representation of Random Forest Regression to determine feature importance for the co2 column.

Some points to note from our Correlation Plots and Random Forest Regression results:

- 'oil_co2', ‘'cummulative_coal_co2', 'gas_co2', 'gdp', 'primary_energy_consumption', ‘coal_co2', 'cumulative_gas_co2', and 'cumulative_oil_co2' have a strong positive correlation with ‘co2’.

- 'share_global_cumulative_gas_co2', 'total_ghg_excluding_lucf', and 'co2_per_unit_energy' have a strong negative correlation with ‘co2’.

Tableau Visualizations

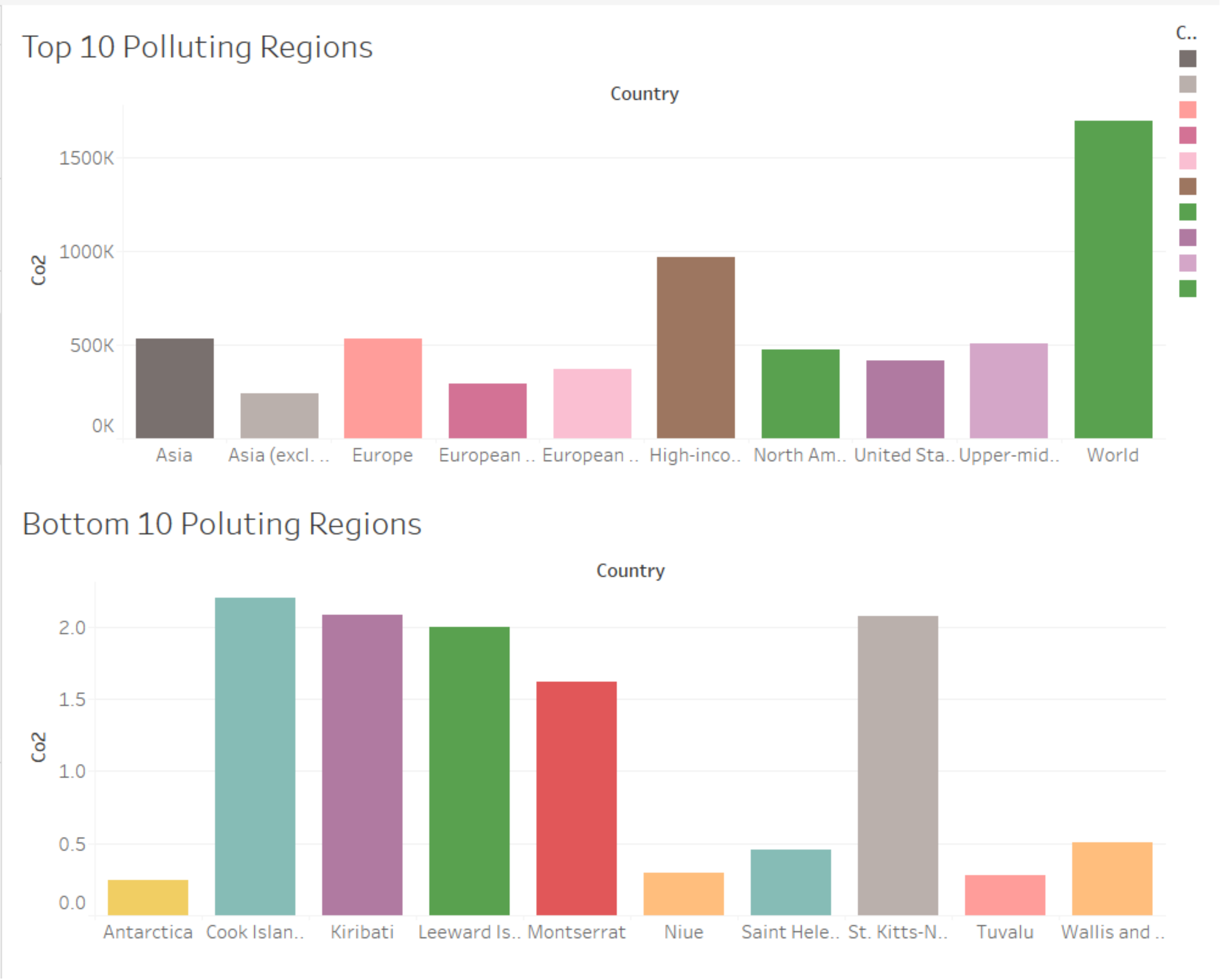
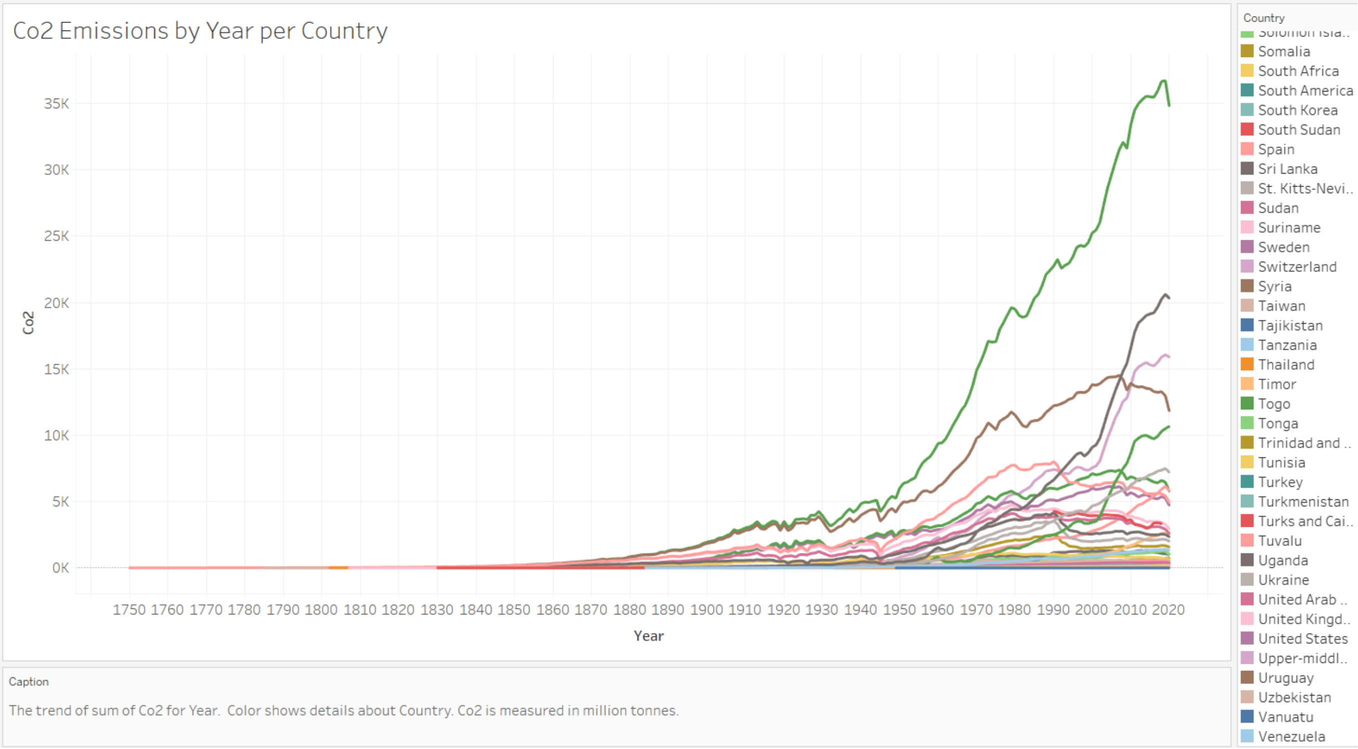


Fig. 4: Bar Chart representing the top and bottom 10 polluters.



The trend of sum of Co2 for Year. Color shows details about Country. Co2 is measured in million tonnes.

Fig. 5: Line Graph representing each country/region's co2 emissions per year.

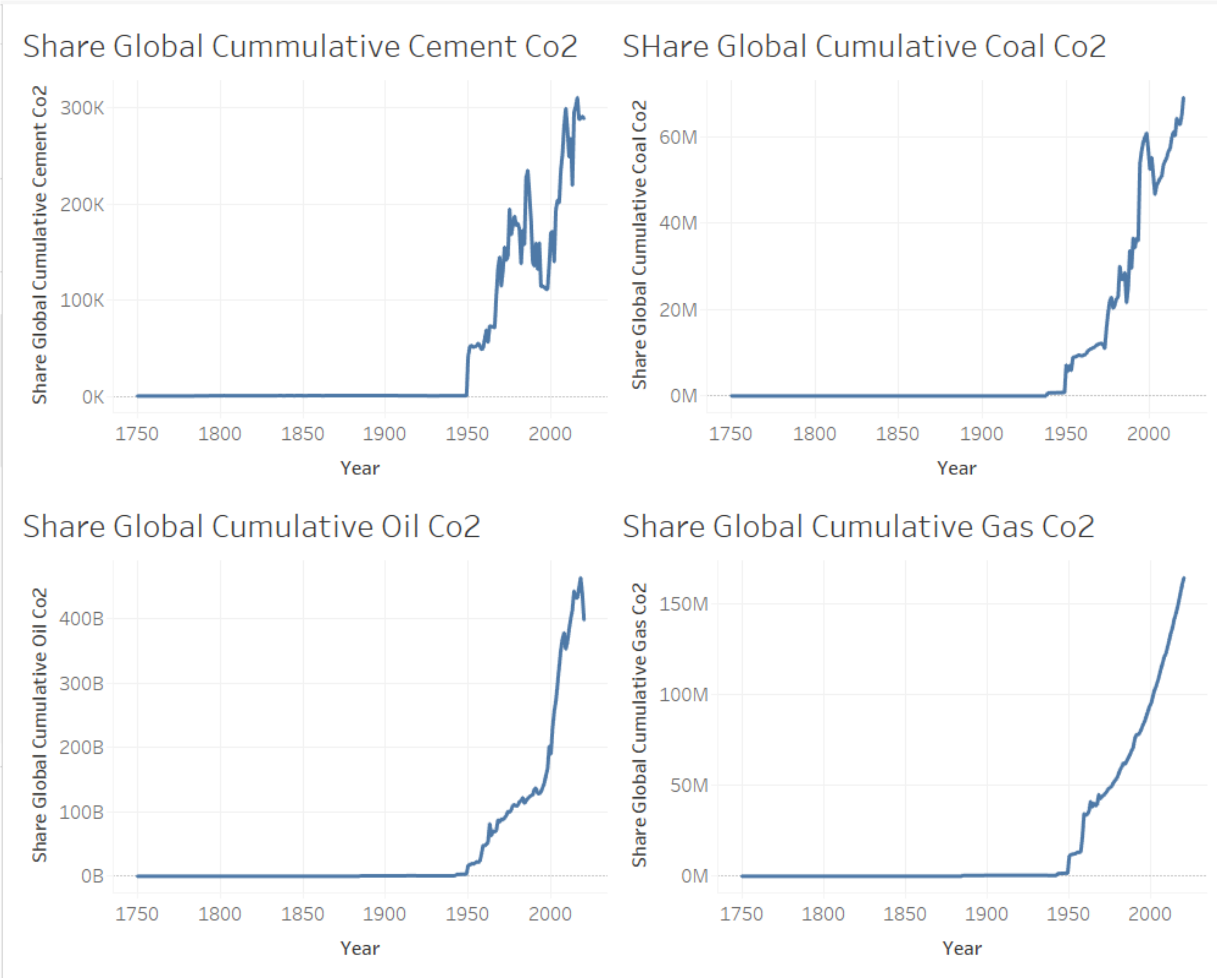


Fig. 6: Tableau Dashboard representing shares of cumulative cement, coal, oil, and gas co2.

Some important things to note from figure 6:

- Cumulative oil co2 is the largest co2 emission.
- Cumulative cement co2 accounts for the smallest portion of co2 emissions.

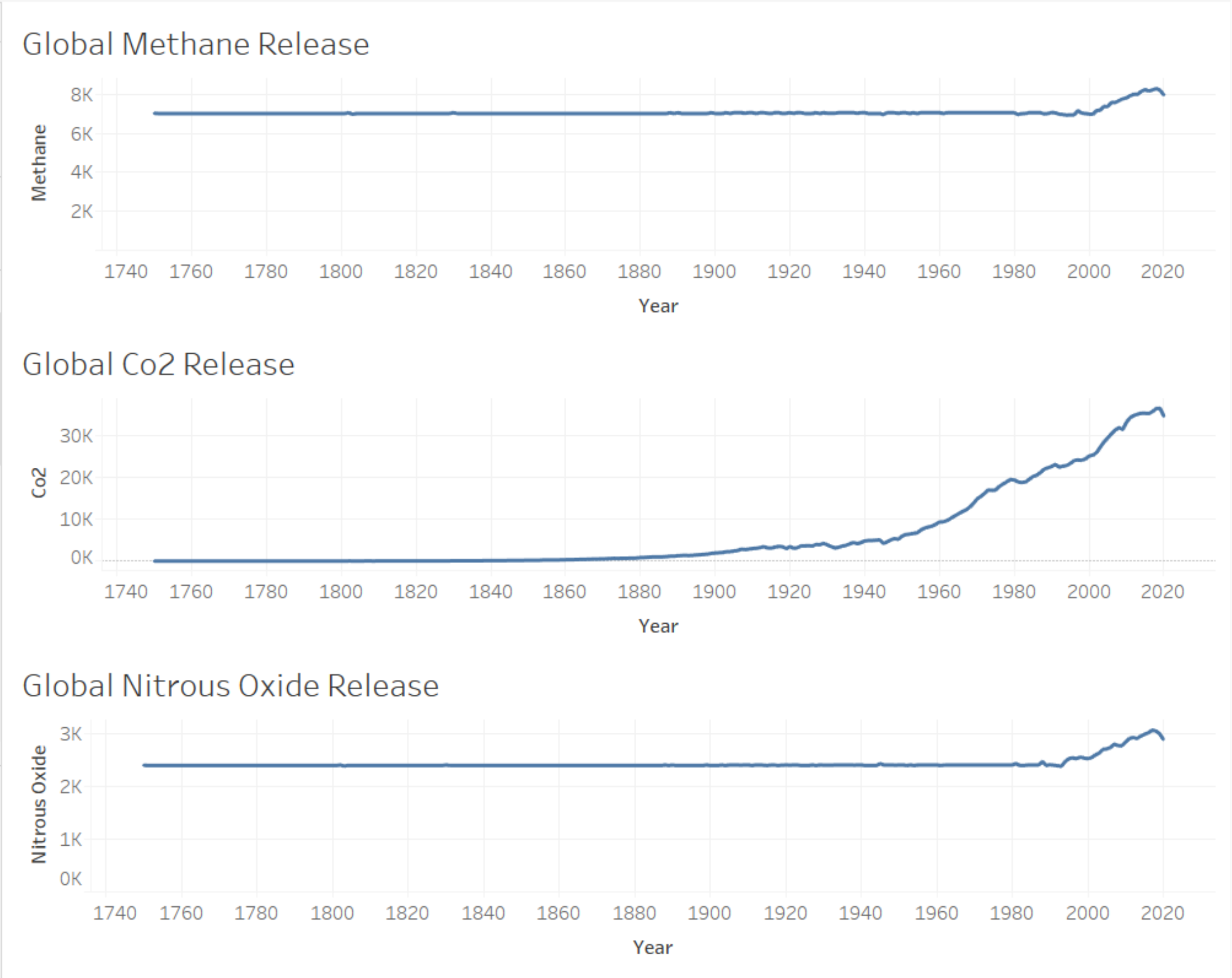


Fig.7 : Tableau Dashboard representing global methane, co2, and nitrous oxide release.

Handling Missing Data

After dropping columns with more than 80% of data missing, I then decided to drop the 'iso_code' column. Since this column did not add any useful information for the analysis and only included codes for the countries/regions, I decided it would be best to remove it. My next step to handle the remaining null values in my dataset was to scale my data and make a KNN Impute function. I separated my dataset into subsets by grouping rows which had the same country values. This ensured that the missing values for each country/region were filled in based on other values for that region. I then ran my subsets through my KNN function which filled in my remaining missing values.

Analysis

One of the main goals for my project was to understand where these greenhouse gas emissions were coming from. This is where my second dataset which consisted of percentages of emissions by sector for the year 2020 came in. Figures 8a and 8b below show what percentages of global greenhouse gas emissions are a result from which production sectors.

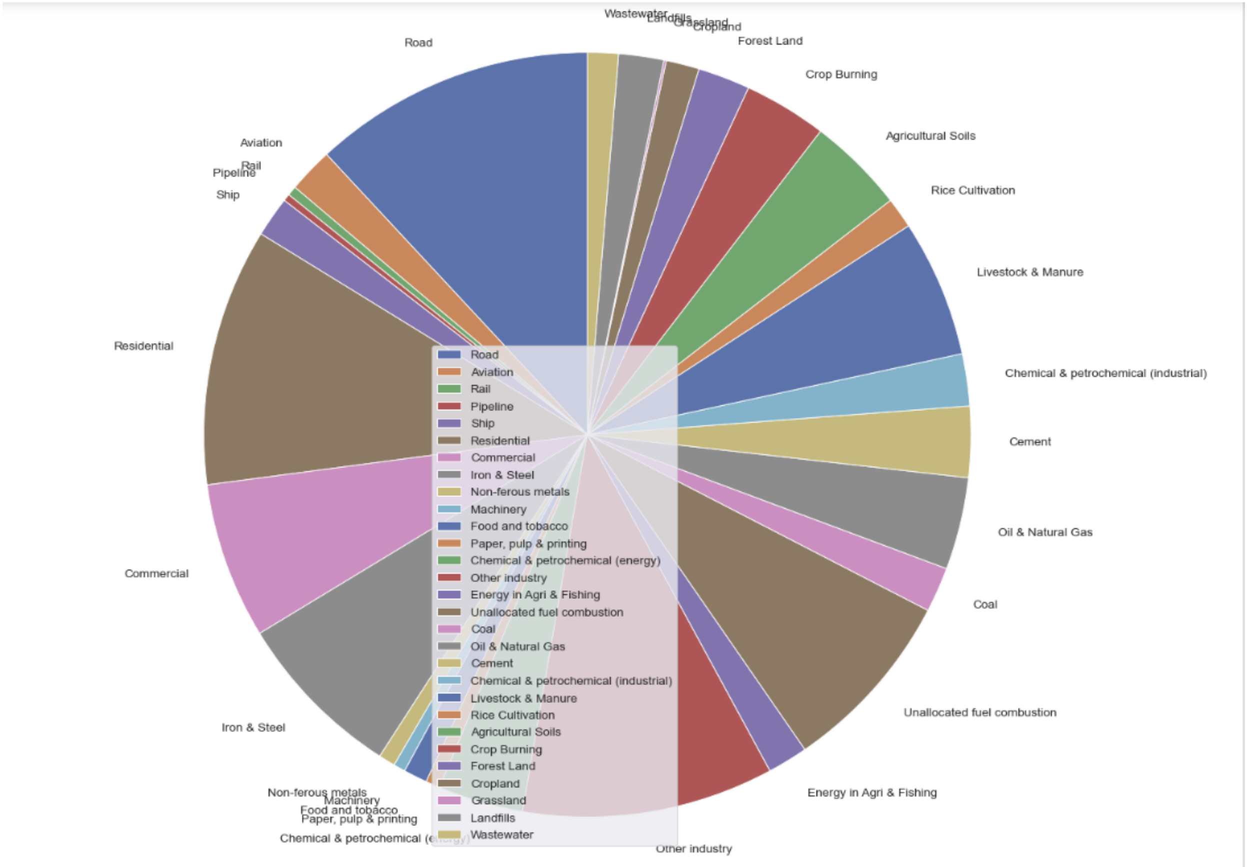


Fig. 8a: Pie Chart representing the percentage of greenhouse gas emissions responsible by each sector.

I then further categorized each data value into one of the following subsectors: ‘energy’, ‘industry’, ‘Agriculture, forest, and land use’, or ‘waste’. The figure below represents this.

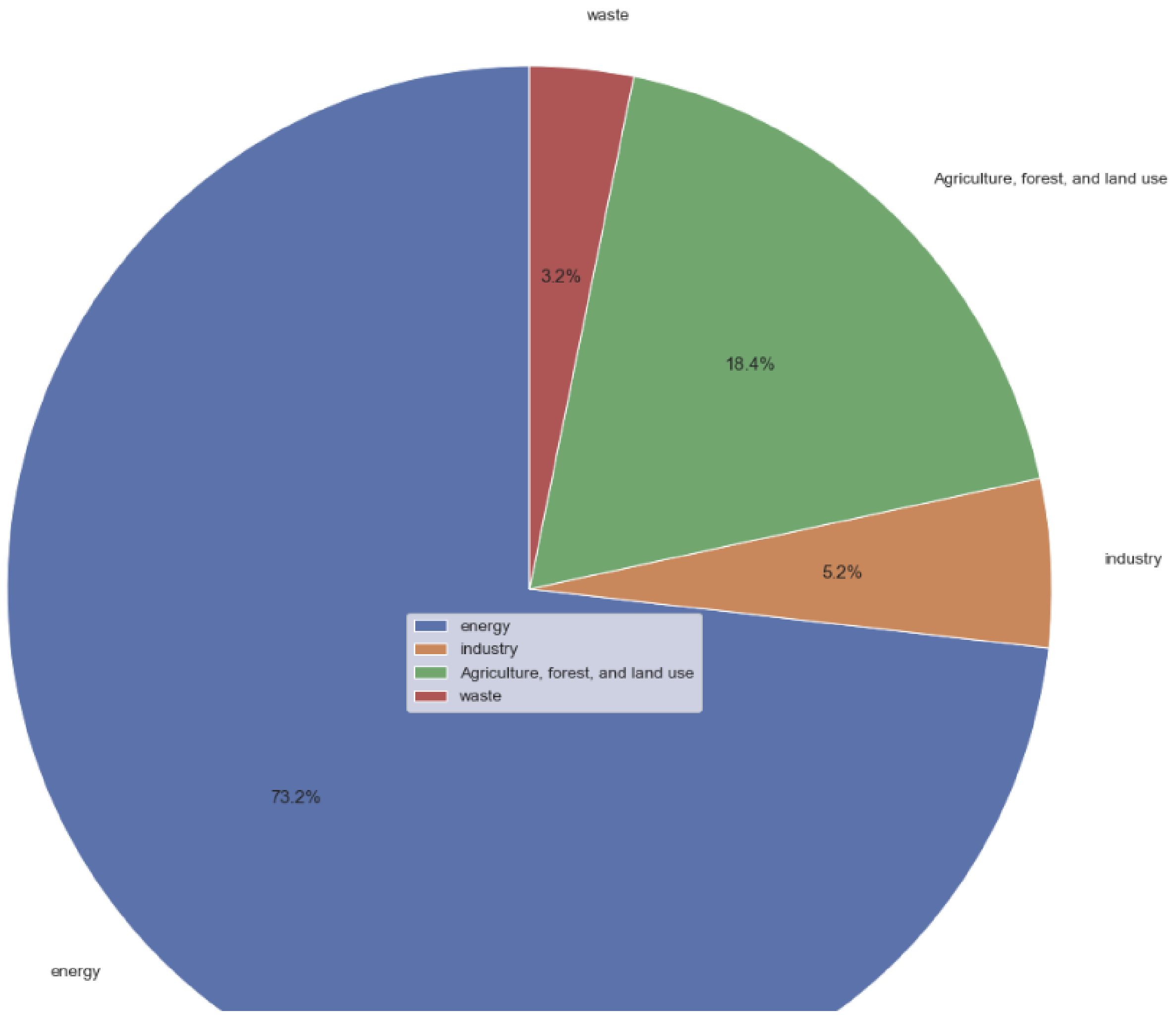
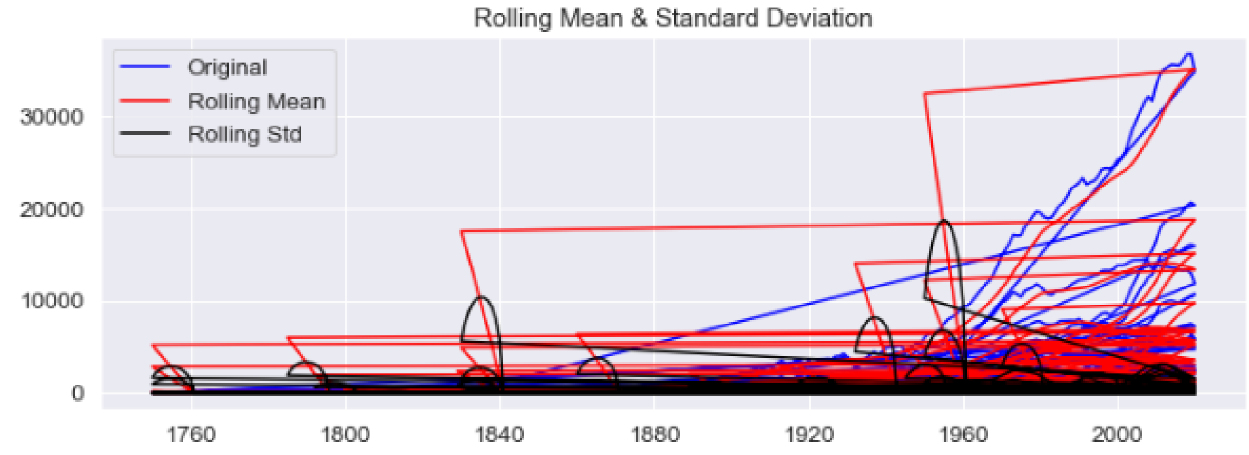


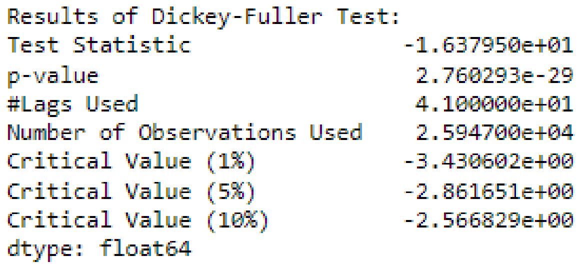
Fig. 8b: Pie Chart representing the percentage of greenhouse gas emissions responsible by each sub-sector.

Some important points to note from the two figures above: based on the year 2020,

- 73.2% of all greenhouse gas emissions are a result of energy productions.
- 18.4% of all greenhouse gas emissions are a result of agriculture, forest, and land use productions.
- 5.2% of all greenhouse gas emissions are a result of industry productions.
- 3.2% of all greenhouse gas emissions are a result of waste productions.

The main dataset I used only consisted of emission values up till the year 2020. Thus, the main analysis goal for my project consisted of predicting future co2 emission values for each country/region for the years 2021-2026. To do this, I used the ARIMA time-series prediction method. My first step in doing this was to convert my 'year' feature from integer values to datetime values and to then make the 'year' column the index of my dataset. Next, I needed to check to see if the co2 column was stationary or not. Figure 9 below shows the rolling mean and standard deviation graph which tested stationarity for the co2 feature.





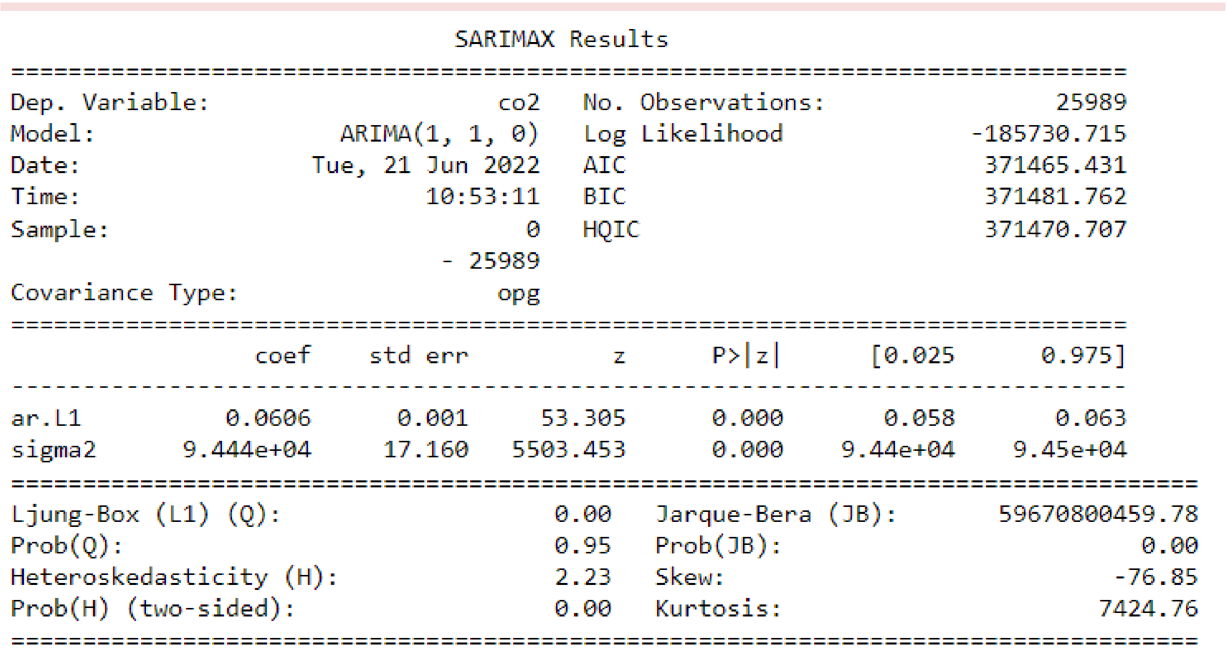
```
Results of Dickey-Fuller Test:
Test Statistic                -1.637950e+01
p-value                        2.760293e-29
#Lags Used                     4.100000e+01
Number of Observations Used    2.594700e+04
Critical Value (1%)           -3.430602e+00
Critical Value (5%)           -2.861651e+00
Critical Value (10%)          -2.566829e+00
dtype: float64
```

Fig. 9: Line Graph representing the Stationarity test for the co2 column.

The graph above shows that our data is not stationary because the mean is increasing even though the standard deviation is small, and the test statistic is greater than the critical value.

Since the data is not stationary, I used an ARIMA model with order = (1, 1, 0) for my predictions. ARIMA(1, 1, 0) is a first-order autoregressive model with one order of nonseasonal differencing and a constant term.

The four figures below, figures 10, 11, 12, and 13, are the results from fitting the ARIMA model.



```
                               SARIMAX Results
==============================================================================
Dep. Variable:                    co2   No. Observations:                25989
Model:                 ARIMA(1, 1, 0)   Log Likelihood             -185730.715
Date:                Tue, 21 Jun 2022   AIC                         371465.431
Time:                        10:53:11   BIC                         371481.762
Sample:                             0   HQIC                        371470.707
                              - 25989
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.0606      0.001     53.305      0.000       0.058       0.063
sigma2      9.444e+04     17.160   5503.453      0.000    9.44e+04    9.45e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):       59670800459.78
Prob(Q):                              0.95   Prob(JB):                         0.00
Heteroskedasticity (H):               2.23   Skew:                           -76.85
Prob(H) (two-sided):                  0.00   Kurtosis:                      7424.76
===================================================================================
```

Fig. 10: Display of ARIMA model results after fitting.

Figure 10 tells us that our model is significant in predicting co2 values because the P-Value in 'P>|z|' column should ideally be less than 0.05 for the model to be significant and our P-Value is 0.000.

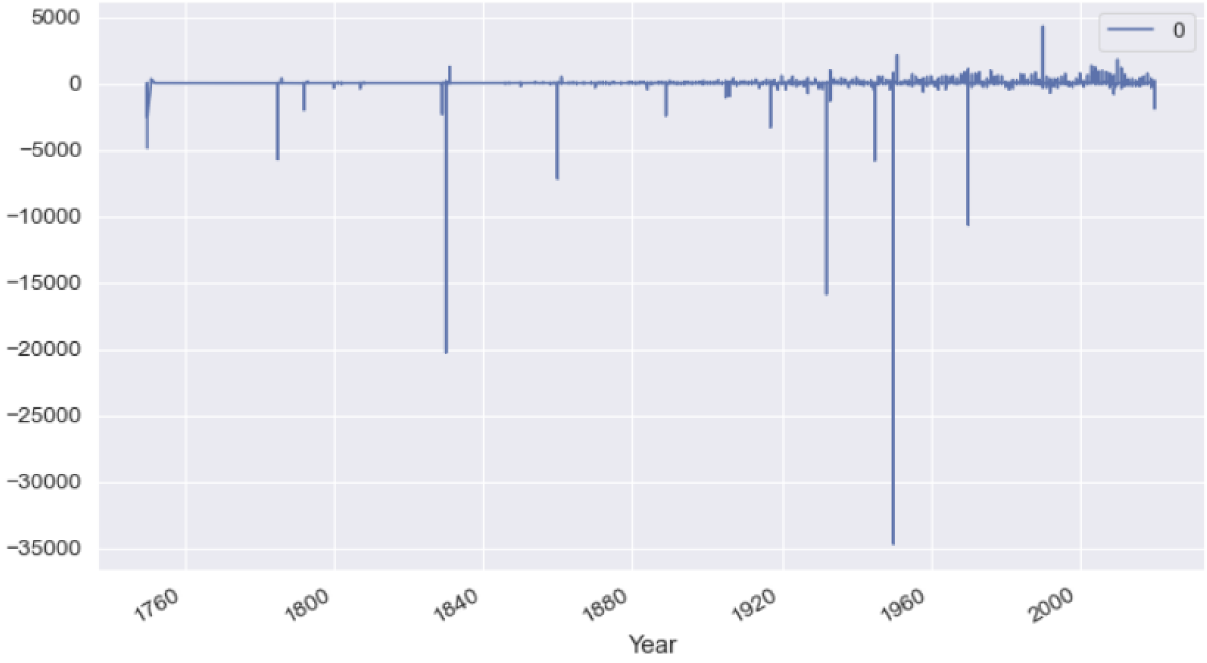


Fig. 11: Line Plot representing residual errors.

Figure 11 shows us that there may be a slight bias in the prediction since we have a non-zero mean in the residuals.

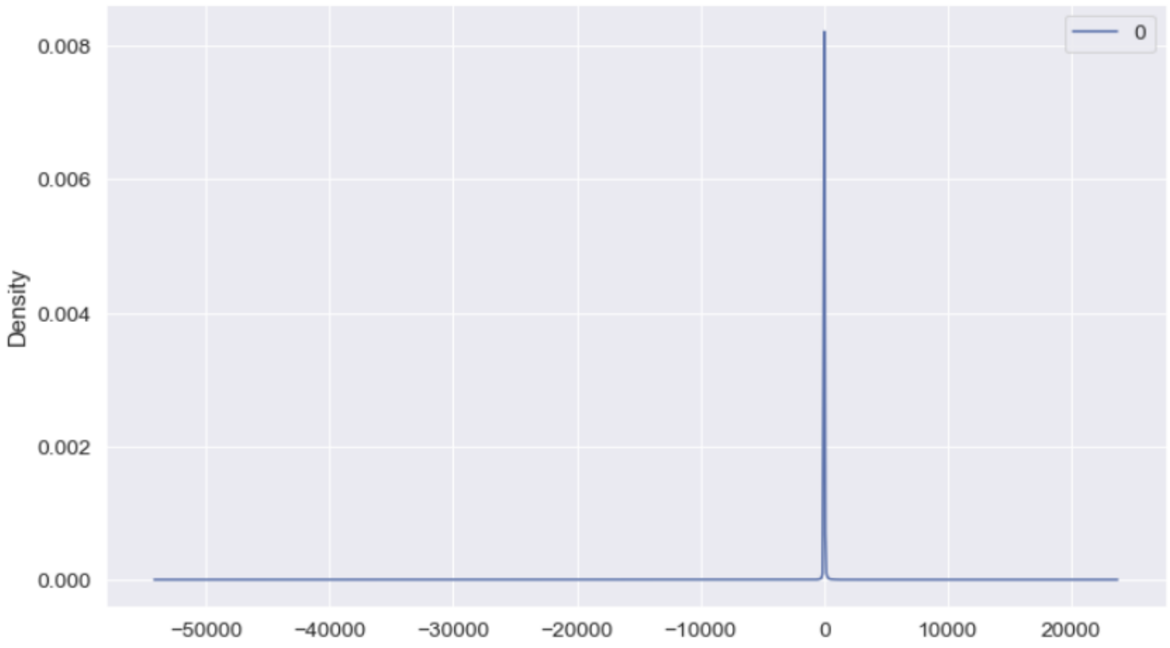
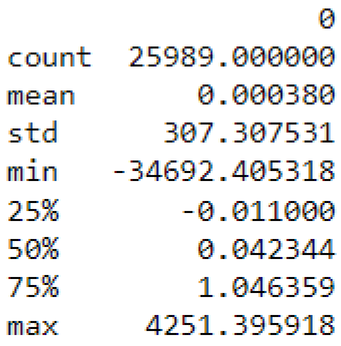


Fig. 12: Density Line Plot

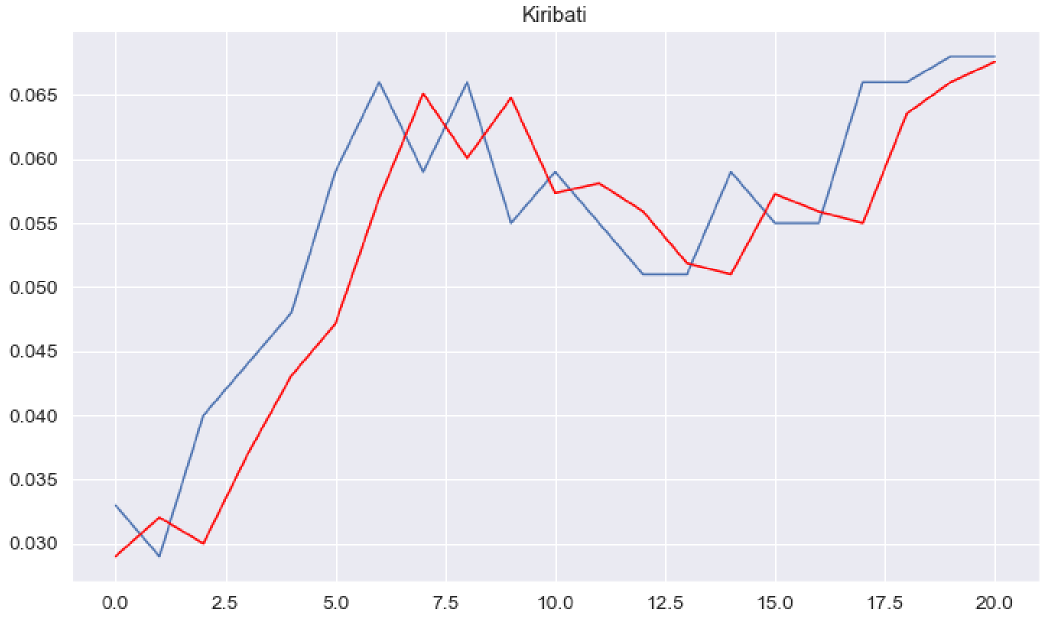
Figure 12 shows that our errors are Guassian and may be centered at 0.

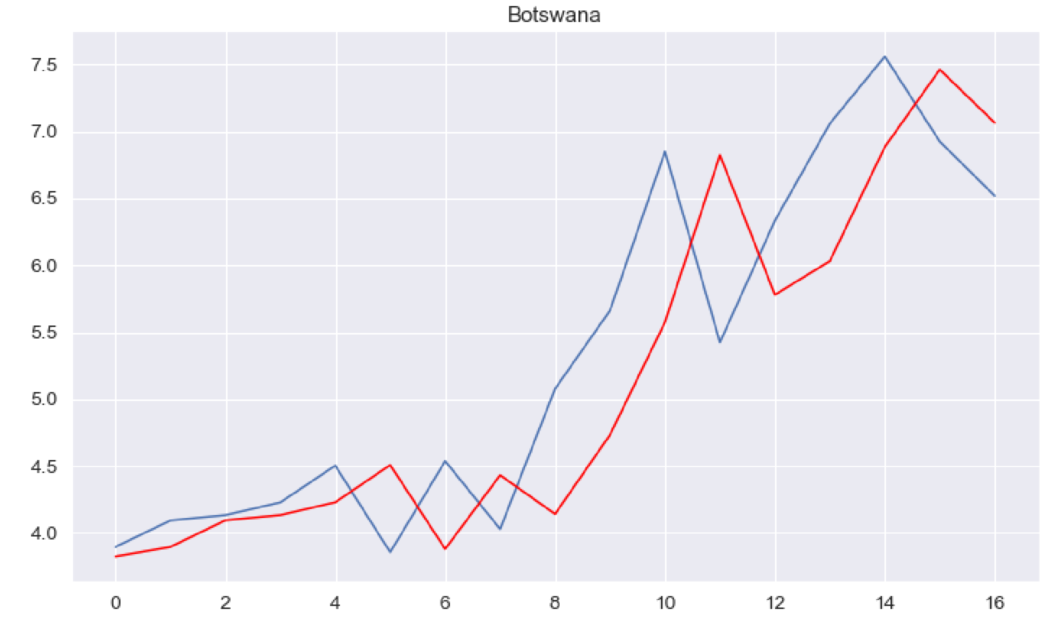


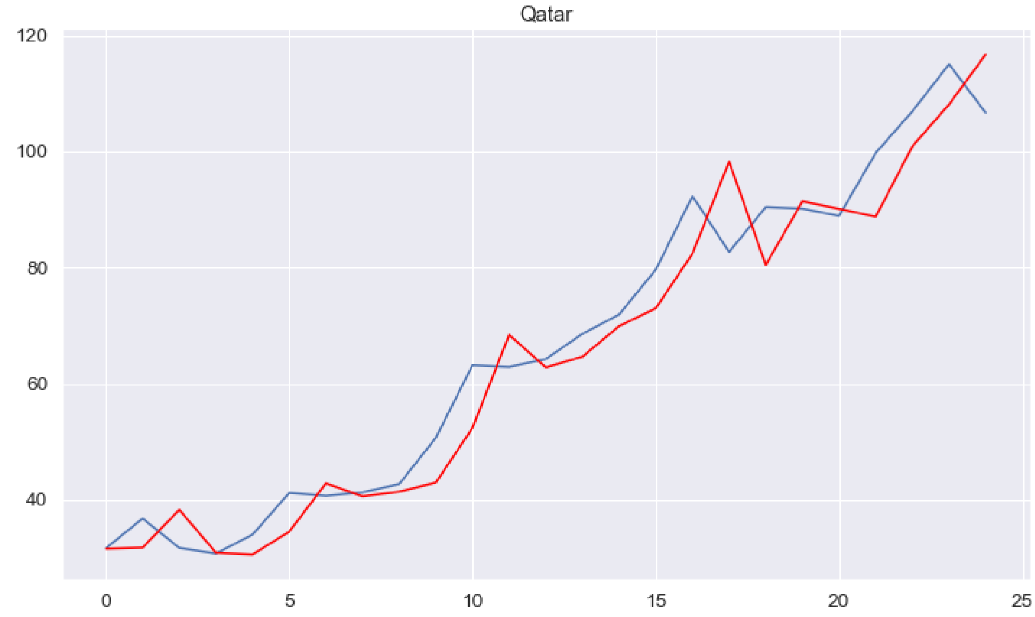
```
                  0
count  25989.000000
mean       0.000380
std      307.307531
min   -34692.405318
25%       -0.011000
50%        0.042344
75%        1.046359
max     4251.395918
```

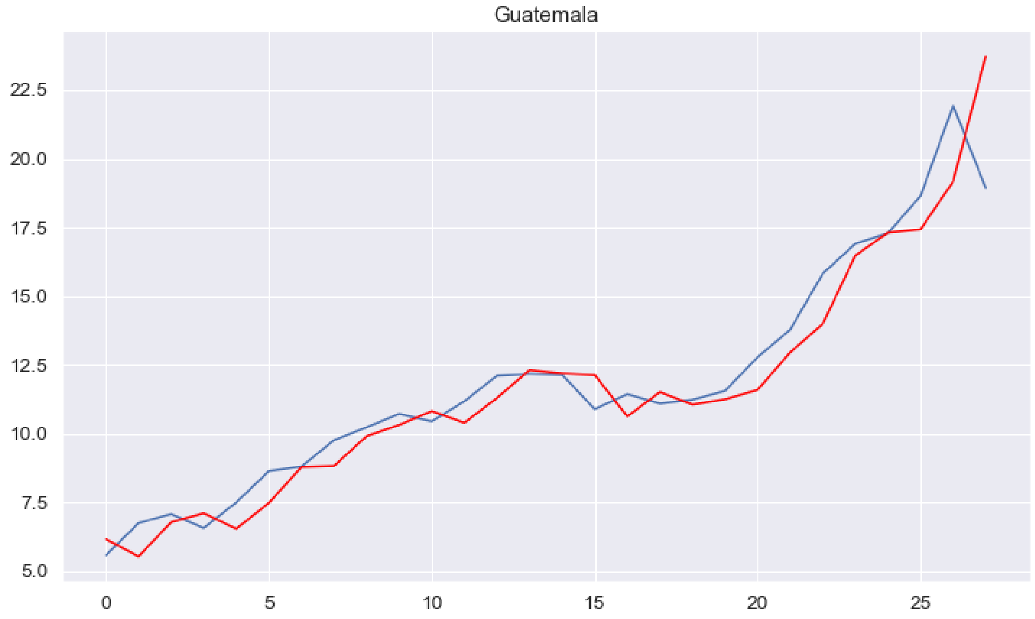
Fig. 13: Diagram representing different features of our ARIMA model.

After running subsets of the data (grouped by country) through the ARIMA model to fit it, I displayed the following graphs. For the graphs below in figure 14, the expected co2 values as provided in the dataset are represented in blue while the values predicted from the model are shown in red. Since I have 248 diagrams, one for each region, for fitting the model, I decided to only include a few of them in the report.


Kiribati
0.065
0.060
0.055
0.050
0.045
0.040
0.035
0.030
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0


Botswana
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
0
2
4
6
8
10
12
14
16


Qatar
120
100
80
60
40
0
5
10
15
20
25


Guatemala
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
0
5
10
15
20
25

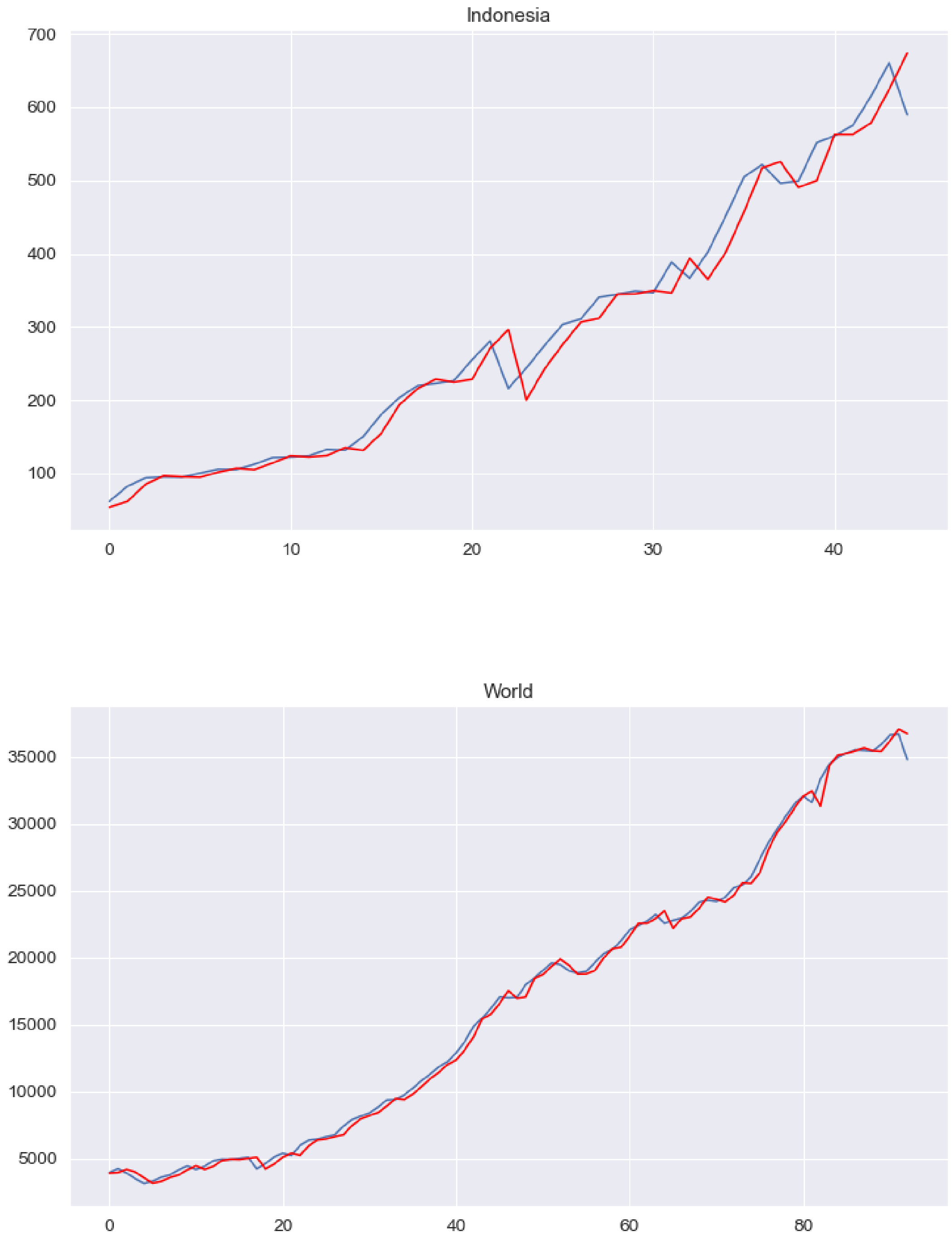


Fig. 14: Line Graph representing expected values in blue and ARIMA predictions in red.

Findings

Lastly, I used the ARIMA model to make co2 predictions for each country for the years 2021-2026. I then added the predictions to the original dataframe and displayed graphs of each country's original co2 values (represented in blue) and the future predictions values for years 2021-2026 (represented in red). The figure below shows ‘World’’s original and predicted values.

```
--------------------------------------------------------------------------------
World
```

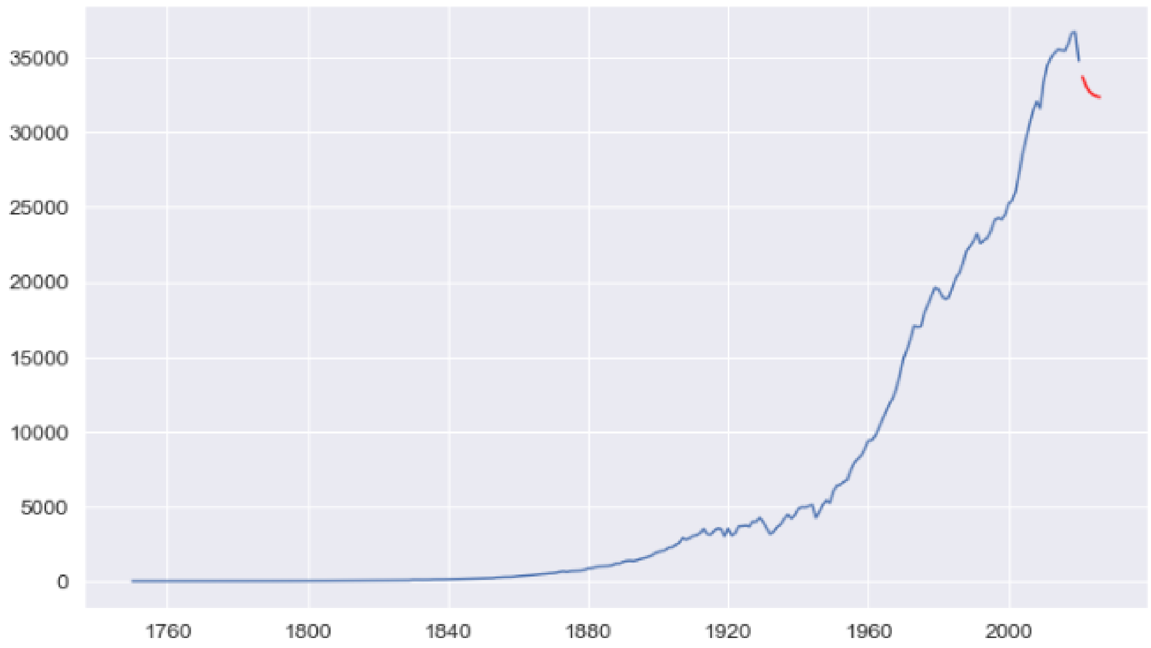


Fig. 15: Line Graph representing the original co2 values for years 1760-2020 (shown in blue) for World and the predicted ARIMA values for years 2021-2026 (shown in red).

Conclusions

We can see that there is indeed a correlation between human production and emission rates due to the increase in co2 emissions starting with the Industrial Revolution. We have found that as of 2020, energy production accounts for 73.2% of all global greenhouse gas emissions causing it to be the largest sector for emissions. Lastly, we were able to make predictions for future emission rates based on previous trends for different regions.

Future Steps

To continue this project, my next steps would be to:

- Complete ARIMA time series predictions for years 2021-2026 on ‘cement_co2', 'coal_co2', 'gas_co2', 'oil_co2', 'nitrous_oxide', and 'methane' columns.
- Compare predicted results with actual recorded values as the years go on.
- Discover solutions to reduce emission rates.
- Discover more factors which lead to the increase or decrease in emission rates.

References

ARIMA Model In Python| Time Series Forecasting #6|. (2020, September 5). [Video]. YouTube. https://www.youtube.com/watch?v=8FCDpFhd1zk

Brownlee, J. (2020, December 10). *How to Create an ARIMA Model for Time Series Forecasting in Python*. Machine Learning Mastery. https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

Introduction to ARIMA models. (n.d.). Duke.Edu. https://people.duke.edu/%7Ernau/411arim.htm

J. (2018, June 21). *Time Series Forecast : A basic introduction using Python.* Medium. https://medium.com/@stallonejacob/time-series-forecast-a-basic-introduction-using-python-414fcb963000

Data

O. (2021–2022). *GitHub - owid/co2-data: Data on CO2 and greenhouse gas emissions by Our World in Data* [Dataset]. Our World in Data. https://github.com/owid/co2-data

Ritchie, H. (2020, May 11). *Emissions by sector* [Dataset]. The World Resources Institute. https://ourworldindata.org/emissions-by-sector