Applied Data Analysis & Research Methodology

MOD007893

SID: *2167364*

a.r.u. | Anglia Ruskin University

Impact analysis of renewable energy adoption in the UK to achieve net-zero goals by 2050

Abstract

The UK has a lofty target of having Net Zero carbon emissions by 2050. Carbon emissions into the environment are getting worse every day. Coal and fossil fuels account for most of the UK's fuel use. These are the two main causes of carbon emissions. The United Kingdom also uses coal and other fossil fuels to produce its electricity. The need to use renewable fuels is crucial for the future. It will take a significant shift in the way energy is produced and used to meet this goal. The UK must also increase both its production and consumption of renewable energy. We will do data analysis for this project on numerous datasets pertaining to energy usage and production in the UK. These data sets were obtained from the official website of the UK government. We can discover trends, patterns, and insights that can inform policy decisions and support the transition to a net zero world by evaluating these datasets. To assess the data, we will specifically use innovative machine learning algorithms. We will be utilizing a variety of visualization tools, such as Power BI, to make the predictions easier for everyone to understand. We will use time series analysis in several ways to analyze the data. To anticipate the use and production of fuels accurately and successfully, time series analysis is a useful technique. Data from earlier years to the present are analyzed when performing a time series analysis. By evaluating this, it is possible to project future consumption, production, and emission levels. Python, a strong computer language and tool for data analysis, is the programming language utilized in this project. I hope this effort is successful in assisting with the Net Zero emission goal.

Introduction

As per the UN, After the Paris Agreement 2015, countries around the world have set net-zero targets, which aim to balance the amount of greenhouse gas produced with the amount removed from the atmosphere. Transitioning to renewable energy sources is seen as a key solution for reducing greenhouse gas, as they have a low carbon footprint and are sustainable over the long term.

The report that follows provides a thorough overview of a project that uses structured data to build a prognostication model. The project analyses a dataset made available by the UK's National Statistics department using Python code in the Google Colab environment. The study lists the characteristics of the data pipeline, such as acquiring data requirements, gathering, and organizing data, cleaning, and preprocessing data, analyzing, and interpreting data, and developing design and development methodologies. The goal of the project is to develop a machine learning algorithm that uses the Autoregressive Integrated Moving Average (ARIMA) approach for time-series forecasting to analyze trends in fuel consumption and carbon emissions from 1995 to 2022. The report describes the many steps in the data pipeline and displays the flowchart for the ARIMA model. To minimize carbon emissions, the report's conclusions, and suggestions, which are based on the data analysis and interpretation, emphasize the need to concentrate on achieving net zero in the household and transportation sectors.

Data pipeline attributes:

Data Requirement Gathering:

We use structured data in our project, which is made up of numerous CSV and Excel data

files, and we use Python code in the Google Colab environment to analyze the data. Over the past few years, we have developed a tendency to do statistical analysis, and using the results, we have created a model for prognostication.

Data Collection and Organizing:

The dataset we use is published by UK's National statistics department. These datasets consist of data from 1995 to 2022. In this project, we have chosen the monthly and annual reports from 1995 to 2022.

Parameters of dataset include,

- Year
- Total fuel used.
- Renewables
- Coal
- Oil
- Gas
- Nuclear
- Low carbon
- Fossil fuels
- Carbon Absorption
- Carbon Emission
- Transport

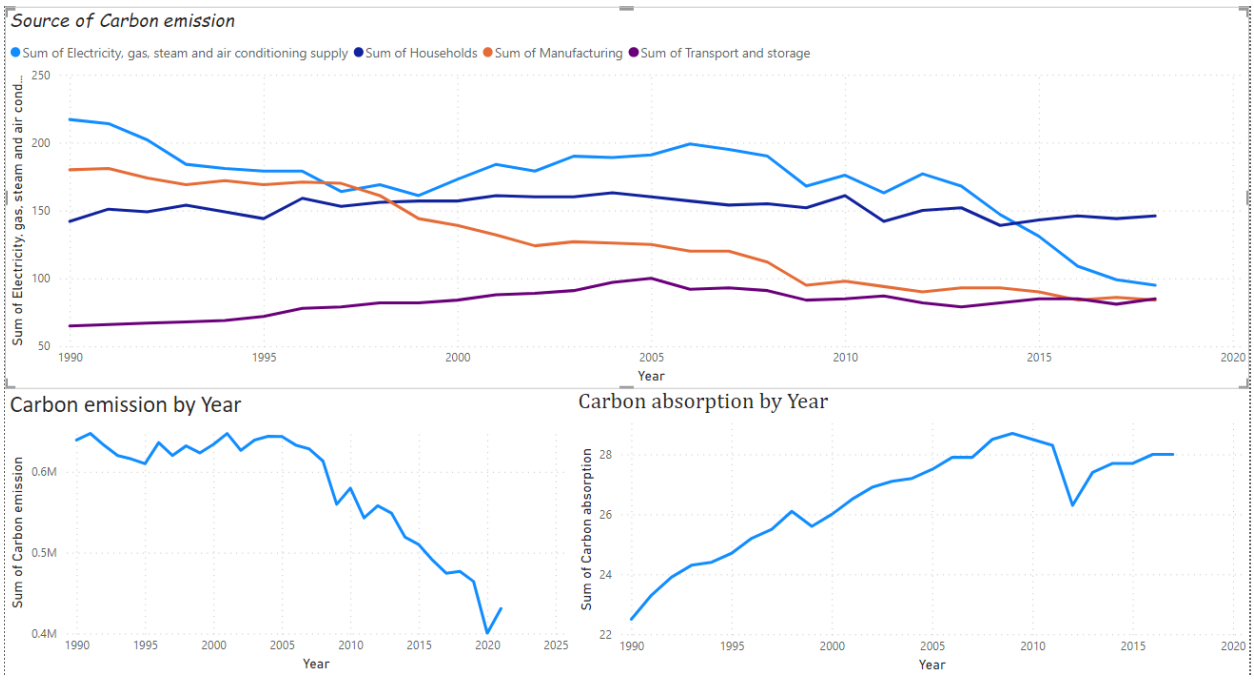
Data Cleaning and Pre-processing:

Once we receive the data, cleaning the data before processing it and applying a model is mandatory. Cleaning data includes looking for null values, removing outliers, checking whether all data has a similar data type, and removing duplicate data. Cleaning data will improve the efficiency of the data.

There were no null values in the dataset we used.

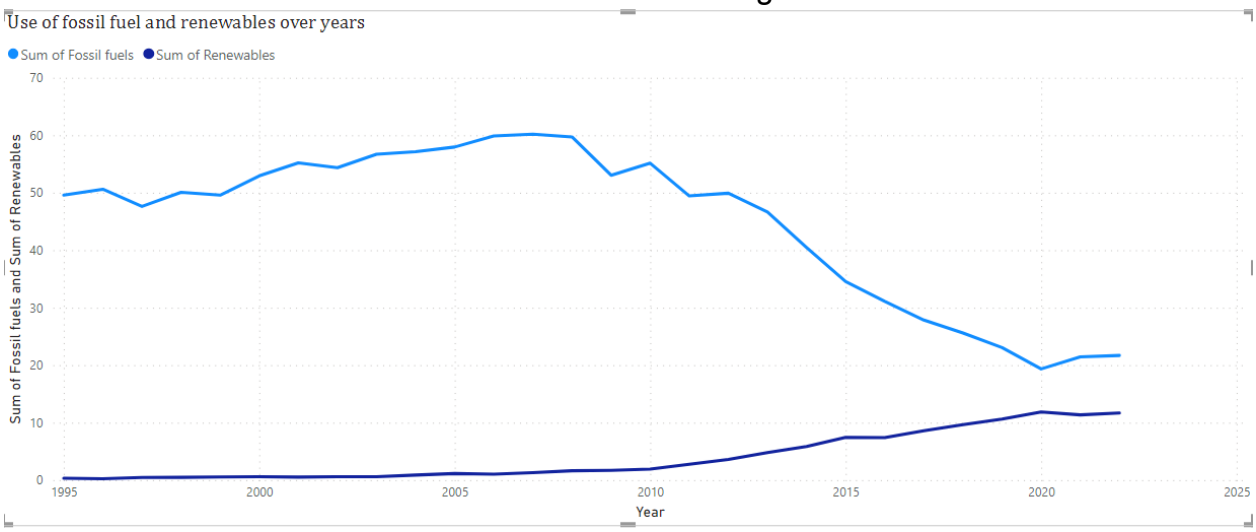
Data Analysis and Interpretation:

The next stage after data cleansing is data analysis and interpretation. For analysis, we use various libraries like pandas, NumPy, matplotlib, seaborn, etc. The result of the analysis is given below.



From the image, we can understand that the emission rate is decreasing, and the absorption rate is increasing. But the rate of absorption is much less compared to the rate of emission. So, we conclude that we need more efficient absorption measures to balance the absorption-emission rate. We can also use renewable energy, which is said to be clean energy to reduce carbon emissions, which can help in neutralizing the emission-absorption rate gradually.

We can also understand how each sector contributes to carbon emissions. Electricity, which was the major contributor to carbon emissions, is seen to be drastically decreasing over the years. A similar decrease can be seen in the manufacturing sector also, whereas both the household and transport sector do not seem to follow the trend. These two sectors are the areas where we need to concentrate more on attaining net zero.



From the above image, we can see that the use of fossil fuels has been decreasing over the years and the use of renewable energy has been increasing over the years. If this trend

follows there is a high chance that fossil fuel usage will be reduced completely and will be completely replaced by renewables.

Design and development (Methodology)

After data analysis, we get a detailed idea about our dataset. From the original data set, we form two separate datasets using only the columns required for doing the machine learning algorithm.

Year	Renewables
1995-01-01	0.34
1996-01-01	0.25
1997-01-01	0.47
1998-01-01	0.49
1999-01-01	0.56
2000-01-01	0.61

	Year	Total fuel used
0	1995-01-01	71.31
1	1996-01-01	73.17
2	1997-01-01	70.07
3	1998-01-01	73.67
4	1999-01-01	72.35
5	2000-01-01	73.18

The other two datasets have a smaller number of columns, and all are significant columns, so it is not necessary to do feature selection for other datasets. After feature selection, we need to deploy the model in an algorithm.

Time series analysis and regression methods are being used in this case. A time series is a collection of data points where time is an integral part of the event from beginning to end. Time series analysis is the process of using certain tools and procedures to analyze time series data. Making predictions based on historical data allows one to forecast the future using both previous and current data

We are going to apply one of the most used methods for time-series forecasting, known as ARIMA, which stands for Autoregressive Integrated Moving Average. The ARIMA model combines moving average (MA), autoregressive (AR), and differencing (I) models. Utilizing a series of historical data to forecast future values is known as an autoregressive (AR) component. The number of historical values utilized for prediction, or p, indicates the order of the AR component.

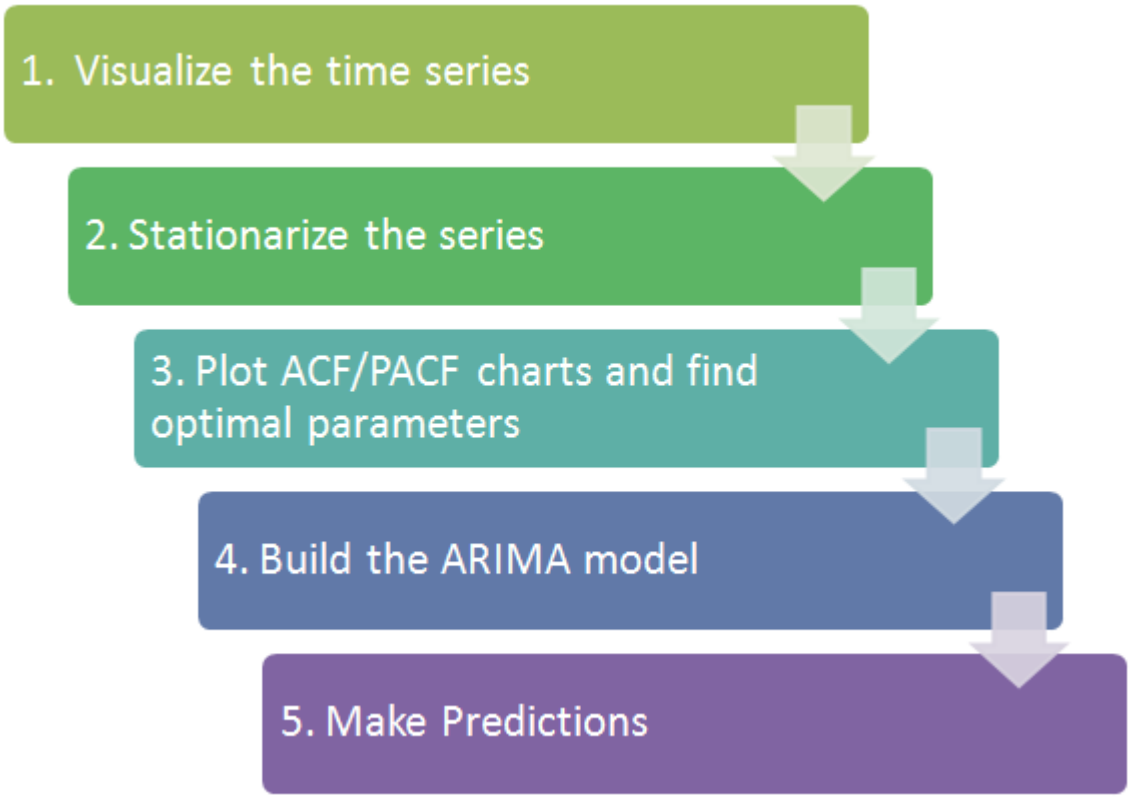
The process of using differences between subsequent observations to make the series stationary is referred to as differencing (I) in this sentence. D, which stands for the number of times the series is differed, represents the order of differencing. The term Moving Average (MA) component describes the process of forecasting future values using historical errors. The number of prior errors used for prediction, or q, indicates the order of the MA component.

The following phases are commonly included in a data pipeline for a project on time series analysis using an ARIMA model: data collecting, data cleaning and preprocessing, exploratory data analysis, model training, model evaluation, and visualizing.

The first stage is to gather pertinent data for time series analysis on fuel usage. This can entail gathering information from several databases or open APIs. The information is gathered from the website of the UK's national statistics office. Following that, the data is cleaned and pre-processed to weed out any unnecessary or missing information and to get it ready for the ARIMA model. The data has undergone the necessary modifications. The next step after

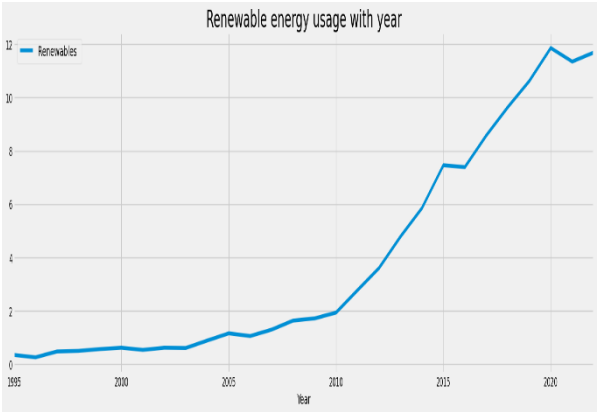
completing this is to analyze exploratory data to put the model into practice. In this step, some data visualizations are carried out to determine the data's trend.

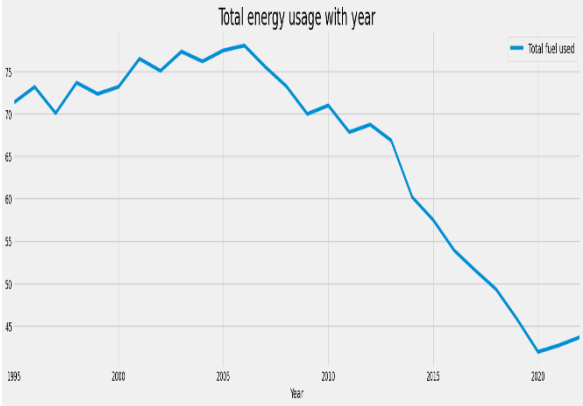
The model is trained on both fossil fuel and renewable fuel data and prediction is done.
I have given below the flowchart for the ARIMA model



FLOWCHART FOR ARIMA MODEL

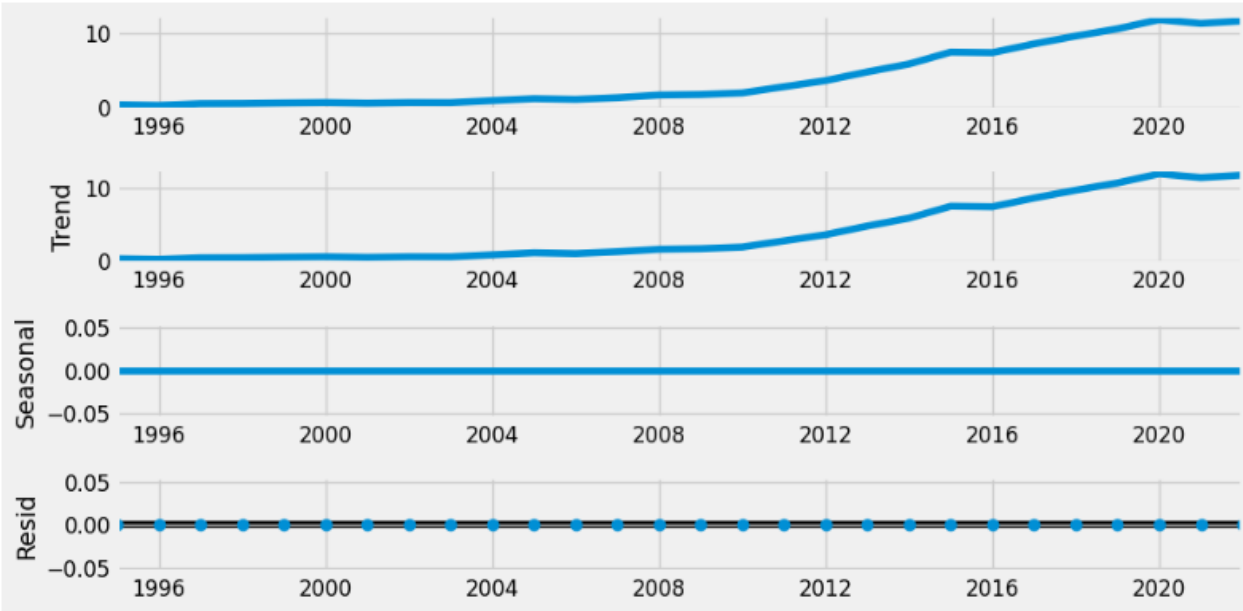
Step 1: Visualizing the series.
We are using matplotlib to visualize the series.





We can see that renewable fuel use is increasing whereas fossil fuel use is decreasing.
We are also doing seasonal decomposition in this step. We can decompose a time series into trend, seasonal, and remainder components. ("A Multi-Indexed Logistic Model for Time Series") The series can be decomposed as an additive or multiplicative combination of the base level, trend,

seasonal index, and residual.



Step 2: Stationarize the series

In its most basic form, stationarity refers to the absence of change in the statistical characteristics of a process producing a time series. It simply means that the series' evolution over time does not impact the evolution of the evolution itself. As a result, the algebraic equivalent is a linear function, not a constant one. A linear function's value changes as x increases, but its rate of change stays the same since it has a constant slope.

Step 3: Plotting ACF and PACF

Time series analysis and forecasting frequently make use of autocorrelation and partial autocorrelation plots. These plots show how much observation in a time series is correlated with observations made at earlier steps. The strength of the connection between two variables is summed up statistically. We can determine the correlation between time series observations and lag observations, or measurements from earlier time steps. A serial correlation, also known as an autocorrelation, is created when time series observations are correlated with values from earlier iterations of the same series.

The Autocorrelation Function, abbreviated as ACF, is a plot of a time series' autocorrelation by lag. This diagram is also known as an autocorrelation plot or a correlogram. An observation in a time series and observations at earlier time steps are related in a partial autocorrelation, which summarizes this relationship without the effects of intervening data. Both the direct correlation and the indirect correlations make up the autocorrelation between an observation and an observation from a previous time step. These indirect correlations follow a linear relationship between the observation and observations made at subsequent time steps.

The partial autocorrelation function aims to eliminate these indirect correlations. This is the partial autocorrelation's intuition, without getting into the math. An observation in a time series and observations at earlier time steps are related in a partial autocorrelation, which summarizes this relationship without the effects of intervening data. Both the direct correlation and the indirect correlations make up the autocorrelation between an observation and an observation from a previous time step. These indirect correlations follow a linear relationship

between the observation and observations made at subsequent time steps.
The partial autocorrelation function aims to eliminate these indirect correlations. This is the partial autocorrelation's intuition, without getting into the math.

Step 4: Build the ARIMA model.

ARIMA (p, d, q) is the notation used to represent ARIMA models. Seasonality, trend, and noise are taken into consideration by these three parameters in data:

```
p = d = q = range(0, 2)
pdq = list(itertools.product(p, d, q))
seasonal_pdq = [(x[0], x[1], x[2], 12) for x in list(itertools.product(p, d, q))]

print('Examples of parameter combinations for Seasonal ARIMA...')
print('SARIMAX: {} x {}'.format(pdq[1], seasonal_pdq[1]))
print('SARIMAX: {} x {}'.format(pdq[1], seasonal_pdq[2]))
print('SARIMAX: {} x {}'.format(pdq[2], seasonal_pdq[3]))
print('SARIMAX: {} x {}'.format(pdq[2], seasonal_pdq[4]))
```

```
Examples of parameter combinations for Seasonal ARIMA...
SARIMAX: (0, 0, 1) x (0, 0, 1, 12)
SARIMAX: (0, 0, 1) x (0, 1, 0, 12)
SARIMAX: (0, 1, 0) x (0, 1, 1, 12)
SARIMAX: (0, 1, 0) x (1, 0, 0, 12)
```

This is the parameter selection process for our ARIMA Time Series Model for fuel use. Our objective is to use "grid search" to identify the ideal collection of parameters that gives our model the highest performance.

```python
for param in pdq:
    for param_seasonal in seasonal_pdq:
        try:
            mod = sm.tsa.statespace.SARIMAX(df_select,
                                            order=param,
                                            seasonal_order=param_seasonal,
                                            enforce_stationarity=False,
                                            enforce_invertibility=False)

            results1 = mod.fit()

            print('ARIMA{}x{}12 - AIC:{}'.format(param, param_seasonal, results.aic))
        except:
            continue
```

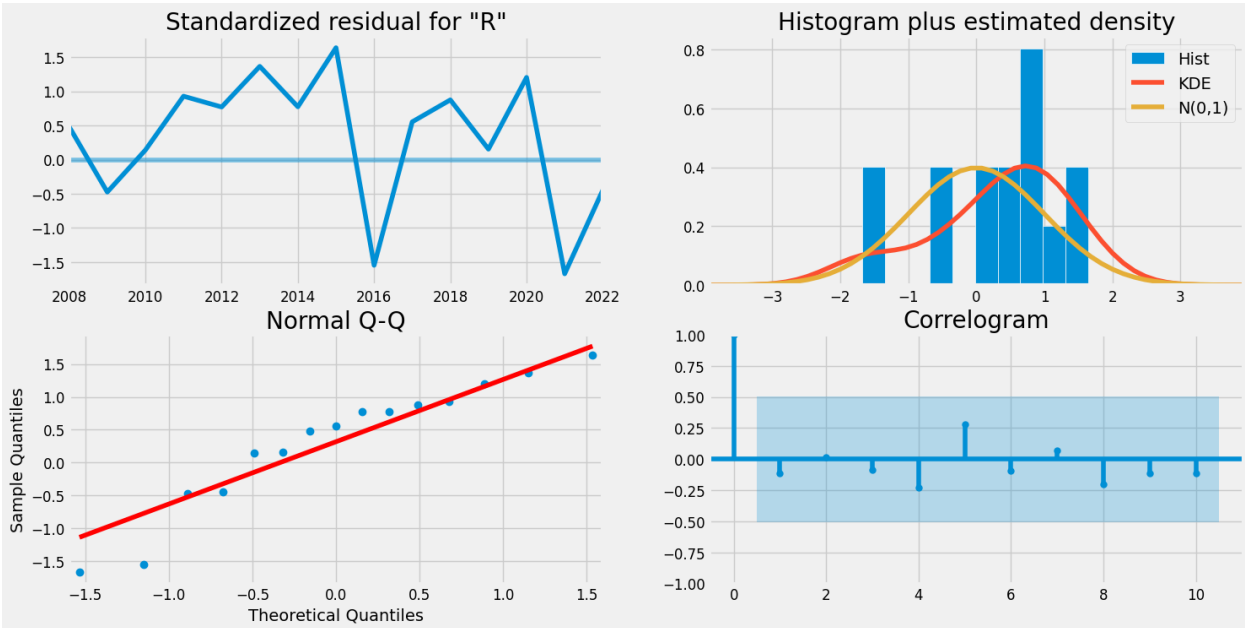
```
ARIMA(0, 0, 0)x(0, 0, 0, 12)12 - AIC:35.5804700694779
ARIMA(0, 0, 0)x(0, 0, 1, 12)12 - AIC:35.5804700694779
ARIMA(0, 0, 0)x(0, 1, 0, 12)12 - AIC:35.5804700694779
ARIMA(0, 0, 0)x(0, 1, 1, 12)12 - AIC:35.5804700694779
ARIMA(0, 0, 0)x(1, 0, 0, 12)12 - AIC:35.5804700694779
ARIMA(0, 0, 0)x(1, 0, 1, 12)12 - AIC:35.5804700694779
ARIMA(0, 0, 0)x(1, 1, 0, 12)12 - AIC:35.5804700694779
ARIMA(0, 0, 0)x(1, 1, 1, 12)12 - AIC:35.5804700694779
ARIMA(0, 0, 1)x(0, 0, 0, 12)12 - AIC:35.5804700694779
ARIMA(0, 0, 1)x(0, 0, 1, 12)12 - AIC:35.5804700694779
ARIMA(0, 0, 1)x(0, 1, 0, 12)12 - AIC:35.5804700694779
ARIMA(0, 0, 1)x(0, 1, 1, 12)12 - AIC:35.5804700694779
ARIMA(0, 0, 1)x(1, 0, 0, 12)12 - AIC:35.5804700694779
ARIMA(0, 0, 1)x(1, 0, 1, 12)12 - AIC:35.5804700694779
ARIMA(0, 0, 1)x(1, 1, 0, 12)12 - AIC:35.5804700694779
```

The next process in this step is fitting the model.

```python
mod = sm.tsa.statespace.SARIMAX(df_select,
                                order=(1, 1, 1),
                                seasonal_order=(1, 1, 0, 12),

                                enforce_invertibility=False)

results1 = mod.fit()

print(results1.summary().tables[1])
```

```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.8658      0.259      3.347      0.001       0.359       1.373
ma.L1         -0.3043      0.568     -0.536      0.592      -1.417       0.808
ar.S.L12       0.7529      0.629      1.196      0.232      -0.481       1.987
sigma2         2.3221      4.591      0.506      0.613      -6.675      11.320
==============================================================================
```

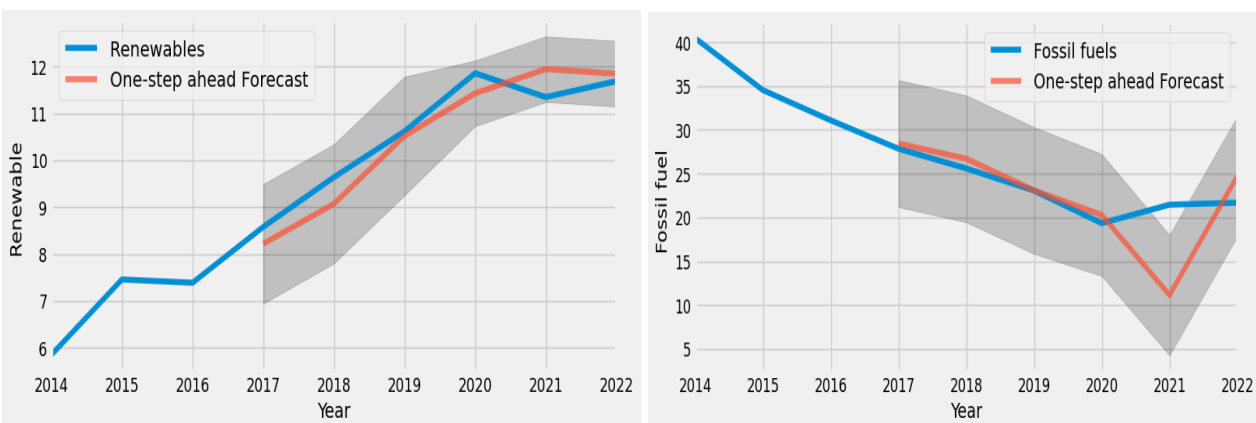
Then we do the model diagnostics to find any unusual behavior.



Even though our model is not perfect, we can see that the residuals are normally distributed.

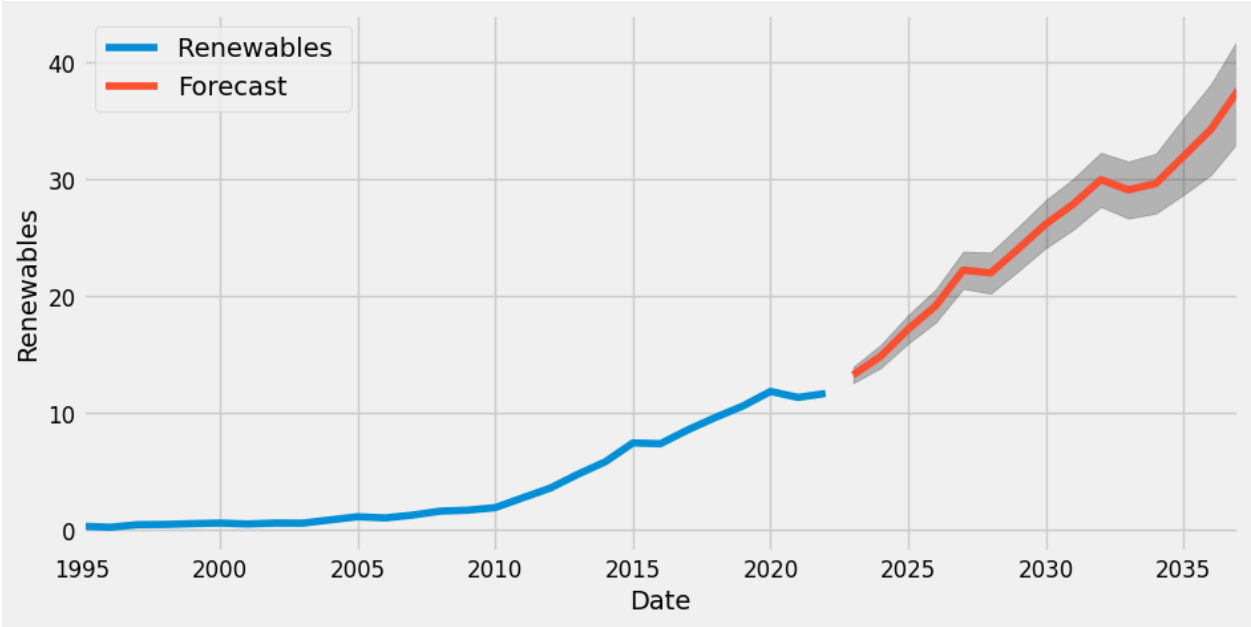
STEP 5: Making Predictions

To help us understand the accuracy of the forecast, we compare predicted energy use to real values in the time series. For this, we set the forecast so that it starts from 2017 to the end of the data.



The line plot displays the observed values in relation to the predictions from the rolling forecast. Overall, our predictions closely match the actual numbers.

Producing and visualizing forecasts



Our model accurately represented gasoline use. It is only normal for us to lose faith in our principles as we project deeper into the future. The confidence intervals produced by our model, which expand as we go further into the future, reflect this.

Apart from the ARIMA model, we are also using the random forest regressor and power bin for predicting future data.
Now we must check the error in this prediction. For that, we are using the Root Mean Squared Error, rf_score, etc.

Data Visualization

Visualization is the most effective method of displaying the expected data. Data visualization is the process of presenting data and all the information we have acquired graphically. A graph can make the data simple to grasp. To visualize the data that we forecasted, we utilized a variety of charts.
To visualize the data, I utilized the Power BI desktop application along with matplotlib and seaborn. Microsoft offers a software service called Power BI that enables customers to see and analyze data from numerous sources. For data modeling, data transformation, data visualization, and report preparation, it offers a range of tools and functionalities.

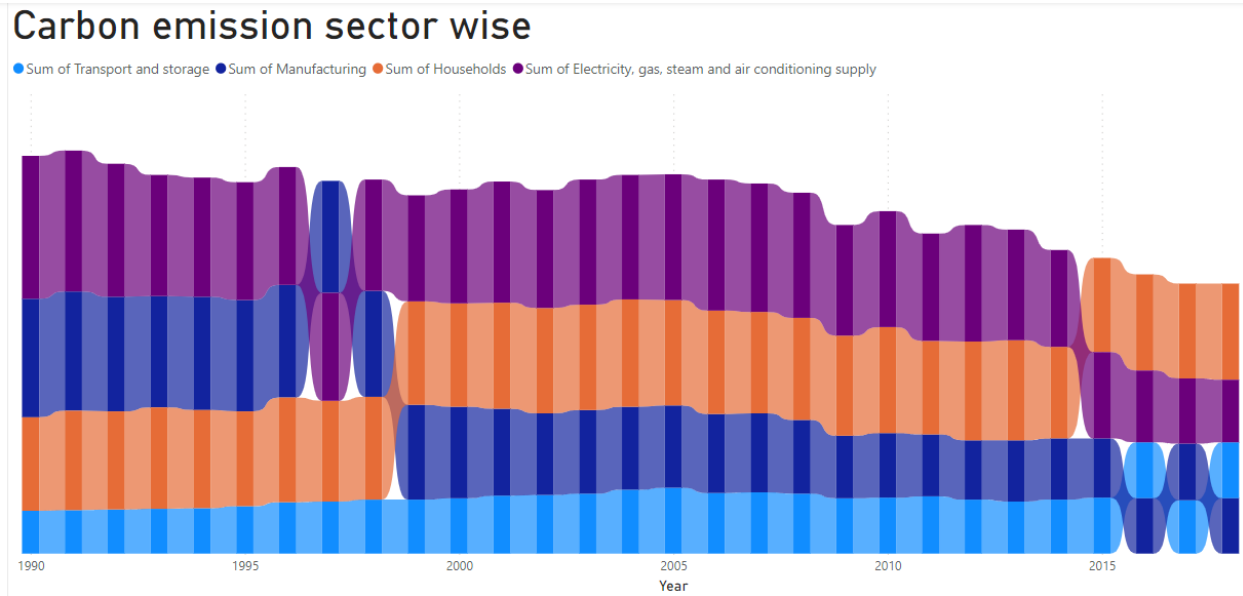


Fig 1. Shows how each sector contributes to carbon emissions. Electricity, which was the major contributor to carbon emissions, is seen to be drastically decreasing over the years. A similar decrease can be seen in the manufacturing sector also, whereas both the household and transport sector do not seem to follow the trend.

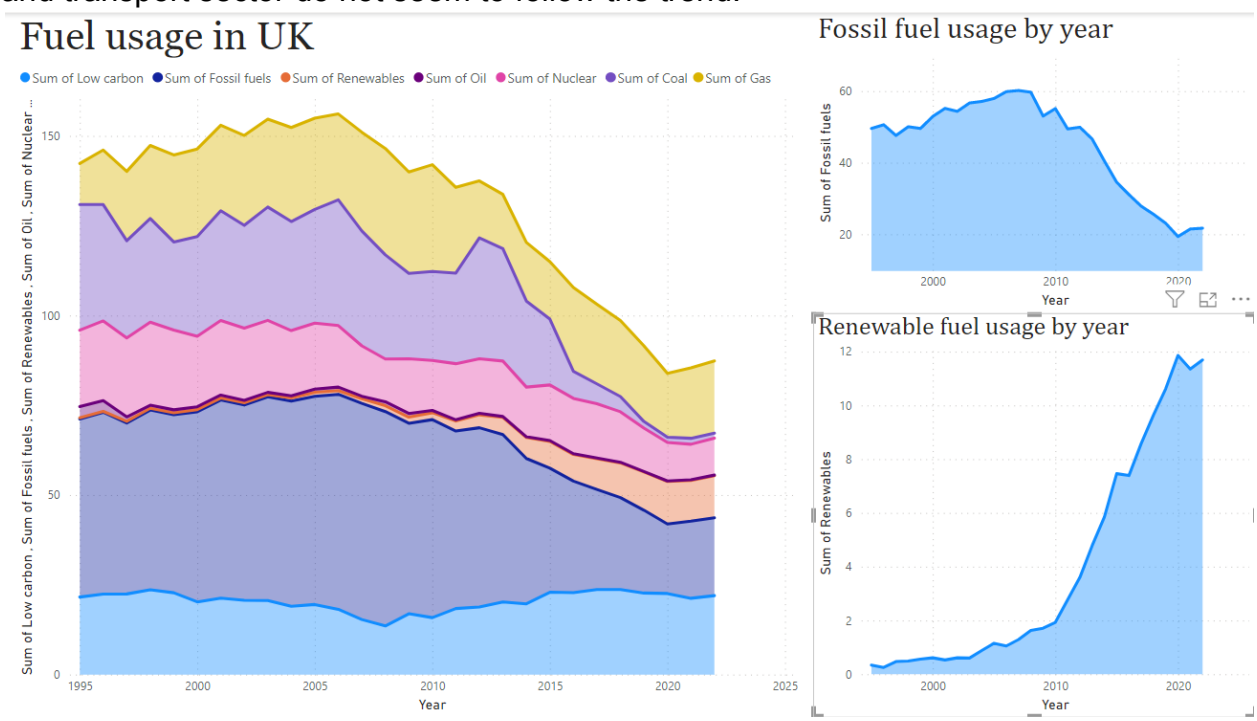


Fig 1. Shows the fuel used over the years. It also shows the trend of how the use of all fuels except renewables and fossil is being decreased.

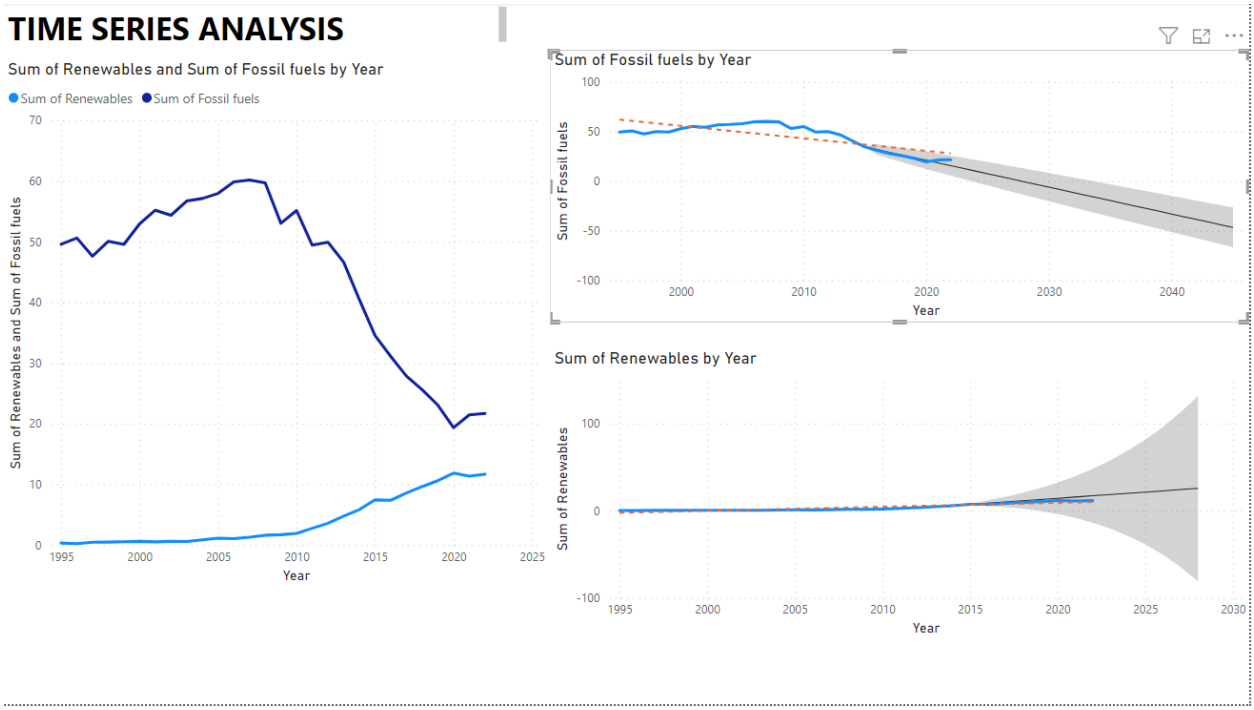


Fig 1. Shows the comparison of fossil fuel and renewable energy usage. It also shows the forecast for fossil fuel usage and renewable fuel usage. We can see the trend that fossil fuel usage is decreasing, and renewable fuel usage is increasing.

From the fig, we can reach a clear assumption that at a particular point before 2050, fossil fuel usage will be neutralized by renewable fuel usage. Renewable energy sources will be capable enough to replace fossil fuels.

Data Confidentiality, Privacy, and Any GDPR (General Data Protection Regulation) Issue

GDPR, data privacy, and confidentiality are significant challenges in the UK. People have a right to know what personally identifiable information is being collected about them, why it is being collected, and how it will be used under the Data Protection Act of 2018. Additionally, they have the right to request access to their personal data and, in some cases, the correction, erasure, or restriction of that data. Before collecting and processing a person's personal information, a company in the UK is required to have that person's consent. Additionally, the company is required to store the information securely and guard it against unauthorized access and disclosure.

The National Statistics Department of the UK Government's website, which is freely accessible online for research purposes, provided the data that we used for this project. The nation's data protection laws are not broken by it. The additional datasets are freely accessible in Kaggle and adhere to the guidelines regarding data security. The data collection we used does not contain any personal information that might be used to identify a specific person.

Evaluation

The purpose of this research was to use Time Series Analysis and the ARIMA model to anticipate how much fossil fuel and renewable fuel will be used by the UK by the year 2050 to meet the country's Net Zero goal. The utilization has been forecasted and the model has been successfully trained. We learned from this that the UK can reach its net zero goal by the year 2050. According to the forecasted data, the UK will use no fossil fuels by the year 2050, while

the use of renewable fuels would rise dramatically.
The usage statistics from 1995 to 2022 and the anticipated data from 2023 to 2050 are shown in Fig. 1.3. We can infer from this graph that the use of renewable fuels is increasing rapidly and has surpassed the use of fossil fuels in earlier years.
When analyzing the situation of fossil fuels, utilization was declining, and according to the projected figures, usage will cease entirely in 2050. And by the year immediately after 2035, less fossil fuel will be used than renewable fuel. This is depicted in Fig. 1.6. All these points to the likelihood that the UK will achieve its goal.
The forecast we received from the model we utilized was excellent. However, there is a probability that the prediction will be incorrect. The Mean Squared Error and Mean Absolute Error approaches are used to determine the prediction's error. Our forecasts have a mean squared error of 0.0023 and a mean absolute error of 0.0304. Our prediction's error is therefore quite small.

Conclusion

By identifying trends and patterns in the data that could guide policy decisions and tactics, this data analysis project on the use of fossil fuels and renewable fuels to reach the net zero target has provided insights into the current condition of net-zero targets.
We were able to predict the consumption of fossil fuels and renewable fuels using the ARIMA model, which revealed that the UK is making great progress towards meeting net-zero targets and is anticipated to do so by the middle of the twenty-first century.
The project was focused only on fossil fuel usage and renewable energy usage. As a next step, I hope to do the time series forecast for carbon emission and absorption and find whether they will neutralize each other before 2050. And how much renewable energy should be produced for this purpose? The results of which can be used for practical application. As of supplying the predicted amount of energy into the microgrid for energy production