



*Division of Computing Science and Mathematics
Faculty of Natural Sciences
University of Stirling*

***Visualization and Predictive Modelling of
Energy Consumption & Carbon Costs at
the University of Stirling***

Rashmi Raveendran

**Dissertation submitted in partial fulfilment for the degree of
Master of Science in Big Data**

September 2023

Abstract

Problem: Addressing carbon emissions is a significant concern in today's world, and as an educational institution, the University of Stirling plays a crucial role in researching and mitigating these emissions to minimize environmental risks. Given the substantial volume of energy data available from various sources, deriving meaningful insights can be an immensely challenging task. Therefore, the implementation of a tool for visualizing energy consumption patterns becomes indispensable for the university. Such a tool enables a thorough analysis of the data, identification of inefficiencies, and the ability to make well-informed decisions based on it. Additionally, it has the potential to enhance energy management practices and facilitate further research in this field. Ultimately, the adoption of such a tool contributes to the university's commitment to promoting sustainable practices, which not only aids in reducing environmental impact but also leads to cost savings.

Objectives: The objective of the paper is to develop a visualization tool for the University while also forecasting energy consumption for the next 5 to 10 years. This tool will provide the University with insights into current consumption patterns, enabling a critical assessment of pertinent issues and a commitment to a more sustainable future. Additionally, the aim is to design the tool to be adaptable for future updates, ensuring its utility for long-term use with forthcoming data.

Methodology: A comprehensive analysis of the data was conducted to carefully clean and extract the necessary information from the data. This included a review of visualization techniques and forecasting methods, contributing to the creation of the dataset. Furthermore, a detailed examination of the approach for effectively utilizing the data in meaningful visualizations was undertaken. The choice of PowerBI for visualization was based on its accessibility and the flexibility it offers for tailoring to the project requirements. Additionally, the selection of the SARIMAX model was based on its capability for fine-tuning, accommodating various features such as time series order, seasonal order, and the inclusion of feature variables, as well as its capability to achieve stationarity. Overall, careful consideration of the data's nature and project requirements guided the decision-making process regarding the platform to be employed.

Achievements: The primary objective of the paper was to create a visual dashboard and generate future energy usage predictions. To achieve this, a PowerBI dashboard has been developed for visualization and employed a forecast model for predictive analysis. The dashboard was thoughtfully customized to enable in-depth analysis of energy consumption, initially categorized by energy sources and further refined by building and consumption year filters. Furthermore, we have also generated 5-year future predictions using advanced forecasting techniques. The utilization of the provided 10-year historical data has been maximized, extracting comprehensive insights. In response to end-user requirements, the dashboard's scope has been expanded to incorporate Waste and Travel data while also establishing a Standard Operating Procedure for keeping PowerBI tables up to date with future data. Overall, the project successfully met its objectives, effectively leveraging data to provide valuable insights into the energy usage of the University of Stirling.

Attestation

I understand the nature of plagiarism, and I am aware of the University's policy on this.

I certify that this dissertation reports original work by me during my university project except for the following:

The energy consumption data has been provided by the University of Stirling

The other Scottish universities CO2 emission data has been collected from the open source repository of Sustainable Scotland Network.

Signature

A handwritten signature in black ink, appearing to be 'Zach' followed by a stylized flourish.

Date: 7th September 2023

Acknowledgements

I would like to express my immense gratitude to Professor Rachel Norman, my dissertation supervisor, for her invaluable support, encouragement, and motivation that played a pivotal role in bringing my project to its successful conclusion. Additionally, I extend my gratitude to the entire Computing Science Department at the University of Stirling for imparting essential knowledge during my course. Without the support of the professors, the learning process would have undoubtedly been quite challenging. Finally, as a technology enthusiast, I also wish to acknowledge the entire online community for providing me with up-to-date knowledge articles, research papers, technology blogs, and forums. And, on a personal level, I am immensely thankful to my family for their unwavering motivation and encouragement, both in my decision to pursue a master's degree and throughout my journey in academia.

Table of Contents

ABSTRACT	II
ATTESTATION	III
ACKNOWLEDGEMENTS	IV
TABLE OF CONTENTS.....	V
LIST OF FIGURES	VII
LIST OF TABLES	IX
ABBREVIATIONS	X
1 INTRODUCTION	1
1.1 BACKGROUND AND CONTEXT	1
1.1.1 PHASE I: Data Analysis and Visualization	2
1.1.2 PHASE II: Forecast Modelling	2
1.2 SCOPE AND OBJECTIVES	3
1.3 ACHIEVEMENTS.....	4
1.4 OVERVIEW OF DISSERTATION	5
2 STATE-OF-THE-ART	6
2.1 CARBON EMISSION ANALYSIS.....	6
2.1.1 Carbon Emission in Educational Sector	7
2.2 IMPORTANCE OF DATA SCIENCE IN CARBON EMISSION ANALYSIS.....	7
3 METHODOLOGY.....	9
3.1 INTERACTIVE DASHBOARD.....	9
3.1.1 Intra-Campus Variability.....	9
3.1.2 CO2 Emission Benchmarking among Scottish Universities	10
3.2 ENERGY CONSUMPTION FORECAST	10
3.3 PROCESS WORKFLOW	11
4 DATA PREPARATION & INTEGRATION.....	12
4.1 SCOPE AND CONTEXT OF DATA.....	12
4.2 DATA COLLECTION & INTEGRATION	13
4.3 DATA PREPARATION & PREPROCESSING	14
5 DASHBOARD WITH POWERBI FOR VISUALIZATION	17
5.1 DATA TRANSFORMATION & AGGREGATIONS	17
5.2 DESIGN PRINCIPLE	19
5.3 CASE STUDY: DASHBOARD ANALYSIS	21
5.3.1 Analysing Monthly Trends	21
5.3.2 An Example Based on Building	22
5.3.3 Analysing Yearly Trend	23
5.3.4 Conclusion from the Case Study	24
5.4 CHALLENGES WITH POWERBI DASHBOARD.....	25
6 TIMESERIES FORECAST MODEL USING MACHINE LEARNING	26

6.1	FEATURE ENGINEERING & SELECTION	26
6.2	INTERPRETATION OF TIMESERIES FEATURES	29
6.2.1	<i>Moving Average</i>	29
6.2.2	<i>Seasonal Decomposition</i>	30
6.2.3	<i>Stationarity of the Data</i>	30
6.2.4	<i>Auto Correlation</i>	31
6.3	TIMESERIES MODEL TRAINING	32
6.3.1	<i>Model Assumptions & Hyperparameters</i>	32
6.3.1.1	SARIMAX	33
6.3.1.2	Auto-ARIMA	34
6.3.1.3	Exponential Smoothing	35
6.3.1.4	Prophet	36
6.3.2	<i>Comparison of Predictions & Evaluation Metrics Across All Models</i>	38
6.4	FINAL TIMESERIES MODEL	39
6.4.1	<i>Final Model Interpretation</i>	39
6.4.2	<i>Final Model Results</i>	40
6.5	CHALLENGES IN MODEL TRAINING	41
7	CONCLUSION	42
7.1	SUMMARY	42
7.2	CRITICAL EVALUATION	43
7.2.1	<i>Data Integration</i>	43
7.2.2	<i>PowerBI Dashboard</i>	43
7.2.3	<i>Timeseries Forecast Model</i>	44
7.3	FUTURE WORK	44
7.3.1	<i>Efficient Data Storage & Management</i>	44
7.3.2	<i>Live Data and Dashboard</i>	45
7.3.3	<i>Integration of Timeseries Model with Dashboard</i>	45
	REFERENCES	46
	APPENDIX 1 - PYTHON LIBRARIES USED	49
	APPENDIX 2 – USER GUIDE TO APPEND DATA IN POWERBI TABLES	50
	APPENDIX 3 – INSTALLATION GUIDE FOR POWERBI DESKTOP	52

List of Figures

FIGURE 1.1 THE ENERGY DASHBOARD FOR THE UNIVERSITY OF STIRLING	4
FIGURE 3.1 VISUALIZATION SAMPLE ON THE ENERGY DASHBOARD	9
FIGURE 3.2 DASHBOARD SHOWING TRENDS ACROSS SCOTTISH UNIVERSITIES	10
FIGURE 3.3 THE PROCESS WORKFLOW OF THE PROJECT.....	11
FIGURE 4.1: FIRST SET OF RAW DATA FROM THE UNIVERSITY'S UTILITY RECORDS.....	12
FIGURE 4.2 SECOND SET OF RAW DATA COLLECTED FROM PUBLIC SOURCES	12
FIGURE 4.3: A SAMPLE OF GHG CONVERSION FACTOR FROM THE YEAR 2022	13
FIGURE 4.4: SAMPLE RAW DATASET WITH ELECTRICITY CONSUMPTION DATA (2021-2022)	13
FIGURE 4.5: RAW DATA TO PANDAS DATA FRAME - BEFORE CLEANING	14
FIGURE 4.6: SAMPLE ANNUAL DATA (2021-2022) FOR ELECTRICITY AFTER CLEANING - STAGE 1.....	15
FIGURE 4.7: SAMPLE COMBINED ELECTRICITY DATA (2005-2022) AFTER CLEANING - STAGE 2.....	15
FIGURE 4.8 UNIVERSITY TREND DATASET.....	16
FIGURE 5.1 DATE-DIMENSION TABLE	17
FIGURE 5.2: TABLES & RELATIONSHIPS IN POWERBI.....	17
FIGURE 5.3 SAMPLE OF THE DATA ADDED TO POWERBI	18
FIGURE 5.4 A SAMPLE AFTER TRANSFORMING USING 'UNPIVOT COLUMNS'	18
FIGURE 5.5: AFTER DATA FINAL TRANSFORMATION IN POWERBI	18
FIGURE 5.6 LAYOUT OF THE DASHBOARD.....	19
FIGURE 5.7 SAMPLE OF NAVIGATION PANE.....	19
FIGURE 5.8 SAMPLE OF SLICERS ON DASHBOARD	20
FIGURE 5.9 SAMPLE OF TOGGLE OF ENERGY AND CO2 PAGE.....	20
FIGURE 5.10 SAMPLE OF DISPLAY CARDS ON THE DASHBOARD	20
FIGURE 5.11 CHP MONTHLY TREND.....	21
FIGURE 5.12 ELECTRICITY MONTHLY TREND	21
FIGURE 5.13 GAS MONTHLY TREND.....	22
FIGURE 5.14 OIL MONTHLY TREND.....	22
FIGURE 5.15 ELECTRICITY CONSUMPTION IN ALAN GRANGE	23
FIGURE 5.16 ELECTRICITY CONSUMPTION IN CAMPUS MAINBOARD	23
FIGURE 5.17 YEARLY CHP CONSUMPTION TREND	24
FIGURE 5.18 YEARLY CO2 EMISSIONS FROM CHP	24
FIGURE 6.1 CHP CONSUMPTION DATA.....	26
FIGURE 6.2 DATE INDEXING SAMPLE	26
FIGURE 6.3 TARGET & FEATURE COLUMNS IN THE DATA	27
FIGURE 6.4 CORRELATION WITH TARGET VARIABLE	27
FIGURE 6.5 TRAIN TEST SPLIT	28
FIGURE 6.6 ANOMALIES (RED LABELS) FOUND IN THE TRAINING DATA	28
FIGURE 6.7 FINAL PRE-PROCESSED TRAINING DATA.....	29
FIGURE 6.8 MOVING AVERAGE AND STANDARD DEVIATION OF TRAIN DATA.....	29
FIGURE 6.9 SEASONAL DECOMPOSITION OF THE TRAIN DATA	30
FIGURE 6.10 RESULTS OF THE ADF STATIONARITY TEST.....	31
FIGURE 6.11 AUTO-CORRELATION PLOT	31
FIGURE 6.12 PARTIAL AUTO-CORRELATION PLOT.....	32
FIGURE 6.13 SARIMAX RESIDUAL PLOT	34
FIGURE 6.14 SARIMAX SUMMARY STATISTICS.....	34
FIGURE 6.15 AUTO-ARIMA RESIDUAL PLOT	35

FIGURE 6.16 AUTO-ARIMA SUMMARY STATISTICS..... 35

FIGURE 6.17 EXPONENTIAL SMOOTHING SUMMARY STATISTICS..... 36

FIGURE 6.18 CROSS-VALIDATION RESULTS FROM PROPHET MODEL 37

FIGURE 6.19 COMPONENT PLOT OF PROPHET MODEL..... 38

FIGURE 6.20 PREDICTIONS COMPARISON ON TEST DATA 38

FIGURE 6.21 PREDICTION COMPARISON OF PROPHET ON TEST DATA 39

FIGURE 6.22 COMPARISON OF EVALUATION METRICS RESULTS..... 39

FIGURE 6.23 COMPARISON OF FORECASTED RESULTS WITH TEST DATA 39

FIGURE 6.24 FORECAST RESULT ON FINAL MODEL 40

FIGURE 6.25 TREND OF FORECASTED RESULT 40

List of Tables

TABLE 4.1: BREAKDOWN OF VARIABLES IN THE DATASET 13

TABLE 4.2: INITIAL DATA CLEANING STAGES TO CREATE A DATASET FOR EACH ENERGY SOURCE 15

TABLE 6.1 HYPERPARAMETERS USED IN SARIMAX 33

TABLE 6.2 HYPERPARAMETERS USED IN PROPHET..... 36

TABLE 7.1 USED PYTHON LIBRARY VERSIONS 49

Abbreviations

Abbreviation	Definition
CO2	Carbon dioxide
GHG	Green House Gas
CHP	Combined Heat and Power
ARIMA	Autoregressive Integrated Moving Average
SARIMAX	Seasonal Auto-regressive Integrated Moving Average with eXogenous factors
LSTM	Long Short-Term Memory
SMAPE	Symmetric Mean Absolute Percentage Error
MAPE	Mean Absolute Percentage Error
SSN	Sustainable Scotland Network
ADF Test	Augmented Dickey-Fuller Test
ACF	Autocorrelation Function
PACF	Partial Autocorrelation Function
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
ETL	Extract, Transform, Load
IPCC	Intergovernmental Panel on Climate Change
UK	United Kingdom
RFID	Radio Frequency Identification
IoT	Internet of Things
AR	Augmented Reality
ANN	Artificial Neural Networks
SVM	Support Vector Machine
ETS	Exponential Smoothing

1 Introduction

Campus Energy Consumption & Carbon Footprint Analysis: Driving Sustainability, Reducing Environmental Impact & Fostering a Greener Future

The analysis of energy usage and carbon cost is critical in today's world for a multitude of reasons, including its impact on quality of life, economic development, and productivity. The study of carbon footprint is a broad subject, influenced by numerous factors. These factors include natural events like wildfires and volcanic eruptions, agricultural practices, and even deforestation; however, a significant portion of carbon emissions stems from the economic sector, comprising electricity generation, industries, residentials and transportation, as noted in [34]. Nonetheless, since this project is primarily focused on energy consumption within the education institution sector, this research study and report are confined to that specific scope. Hence assessing the energy patterns with the University will enable us to spot inefficiencies, improving sustainability and resulting in cost savings. Additionally, the study of carbon footprint also aids in reducing greenhouse emissions, which have a substantial impact on climatic changes and global warming. Moreover, it supports businesses and organizations in developing energy policies and utilizing renewable energy sources, all of which will ultimately result in a zero-carbon future that will result in an ecologically conscious future. The UK government has established goals to reduce carbon emissions to net zero by 2050, with a minimum reduction of 68% by 2030 (compared to the baseline year 1990) [1]. Furthermore, given that it plays a significant role in an array of environmental concerns, investigating energy consumption is essential in the context of contemporary climate change mitigation and sustainability [2]. Apart from that, the CO₂ emissions provided throughout the research are calculated by using the GHG conversion factors, which are the standardized values “derived from assessing main GHG contributors - energy-related emissions and process-related emissions and calculated separately based on the fundamental research by the IPCC (Intergovernmental Panel on Climate Change)” [27]. Each year, a set of conversion factors for different energy sources is generated based on scientific research to serve as a benchmark value to assess CO₂ emissions.

In conclusion, this project aims to develop the tools to do an in-depth analysis of energy usage within the University of Stirling, while also highlighting the significance of data science and analytics within the realm of environmental sustainability demonstrating the application of several techniques in carbon footprint analysis.

1.1 Background and Context

The primary driver for studying energy use and carbon costs as a university student and as an individual is a sense of environmental responsibility and a desire to put that awareness into action by supporting my immediate surroundings and their sustainability initiatives. An initial check indicated that the University of Stirling, which is situated on a 360-acre campus [3], utilizes a significant amount of energy because of its various buildings, lecture halls, libraries, offices, and residence halls, as well as it has approximately 17000 students and 1500 staff members [4]. Even though the University has certain sustainable practices in place that contribute to its sustainability goals, to adopt more effective energy-saving measures and reduce carbon emissions, an in-depth investigation of its patterns and trends is required. The dataset from the University of Stirling's original submission to SSN (Sustainable Scotland Network), which included data on electricity, water, gas, oil, waste, fleet, waste, and travel together with their carbon emissions from the years 2007 to 2005, has been used for the research. This analysis can offer insights into instances where the consumption of energy is at its peak, the causes of the increase in

emissions, and potential areas for optimization. In addition, we attempt to identify patterns, trends, and correlations in the data using data analytics techniques and prediction algorithms.

1.1.1 PHASE I: Data Analysis and Visualization

During this stage, our objective is to delve into historical energy usage and associated carbon emissions data of the University of Stirling ranging from the year 2005 to 2022. By adopting this approach, we will have the opportunity to reveal patterns, trends, and connections that can guide us to valuable insights for making informed decisions regarding the optimization of resources and the advancement of sustainability initiatives. Using aggregation methods and algorithms, the complex dataset has been cleaned, reshaped, and filtered to extract insightful information from it. Interactive visualizations have been developed to examine how much energy is consumed throughout the day or in certain regions of a building, as well as to compare and benchmark our performance with that of other Scottish universities. Employing forecasting and predictive analysis tools, it was possible to produce forecasts for the next five years, which may lead to decision-making on resource allocation and planning.

1.1.2 PHASE II: Forecast Modelling

Forecast models used in the analysis's second phase, focusing on spotting anomalies in consumption patterns and properly forecasting future energy demands, have greatly aided research on energy use and the carbon footprint of the University of Stirling. The models were developed using historical consumption data and carbon pricing, resulting in data-driven insights that help with sustainability initiatives, informed decision-making, and the reduction of carbon emissions. Additionally, by forecasting the confidence intervals of the data points, we gain insights into the extreme higher and lower limits that these values may attain. This information allows for prudent decision-making while considering potential risks. Furthermore, it facilitates the implementation of effective energy conservation strategies and cost-control measures, all guided by predictive insights. In essence, forecasting serves as an effective approach to the management of sustainable energy resources, ultimately reducing the environmental impacts.

1.2 Scope and Objectives

The previous section has explained the significance of carbon analysis and its contributions to environmental sustainability, along with outlining the planned methodologies for conducting this analysis. Consequently, the project's scope revolves around the creation of an interactive visualization tool tailored for the University's management of energy consumption data.

The project's defined scope and objectives are listed as follows,

- Gathering data on energy consumption and carbon emissions from the University of Stirling.
- Gathering end-user business requirements to comprehend specific preferences and ensure the inclusion of essential features and functionalities in the project.
- Develop a user-friendly dashboard tailored for the University, offering data analytics and insights.
- Developing a predictive model for projecting energy consumption over the next five years.
- Soliciting feedback from end-users to address any unmet requirements.
- Documenting instructions, if any, for future dashboard usage, such as the addition of future data into the dashboard.

This emphasises how crucial it is to accomplish the above-mentioned objectives, as given the dispersed energy data, conducting in-depth data analysis necessitates the creation of a dedicated dataset and a dashboard that will consolidate all data into one central location and enable effective utilisation and management of energy-related objectives. This effort also provides a thorough analysis of the University's energy usage status. Leveraging data analytics tools and programming languages, particularly PowerBI and Python, will enable the process of data collection and aggregation as well as the provision of analytical insights. The objective also includes forecasting energy consumption, which will aid in predicting energy usage for the next five years, thereby reinforcing the analysis and enhancing the efficiency of energy management.

1.3 Achievements

The project's goal was to centralize energy data and develop an interactive visualization and analytical tool, aligning with the objectives outlined in Section 1.2. This objective has been successfully realized, beginning with the data collection process at the university and extending to meeting end-user needs through the dashboard. By soliciting and incorporating end-user feedback, the alignment of the dashboard and analytics with the university's sustainability objectives and usability has been ensured. By obtaining the data from reliable sources and resolving any discrepancies during the preprocessing stage, the data's accuracy and dependability have been ensured, which guarantees the accuracy of the data when it is presented on the dashboard. Additionally, the use of standard GHG conversion factors outlined in Section 1 helped maintain the accuracy of CO₂ measurements. By defining procedures for integrating new, unforeseen data as well as developing a prediction model capable of predicting the next five years' data, it has been ensured that the dashboard's long-term usability has been accomplished. Furthermore, the dashboard is easily operable by the end user due to the easy availability of development tools or licences for all university staff. The accuracy of the data and the dashboard's flexibility for easily incorporating new data for future years ensure compliance. Despite challenges throughout the project journey, these hurdles encouraged the adoption of creative strategies, leading to a user-friendly solution usable by non-technical individuals as well. Additionally, it has been designed to smoothly incorporate upcoming studies or potential enhancements.

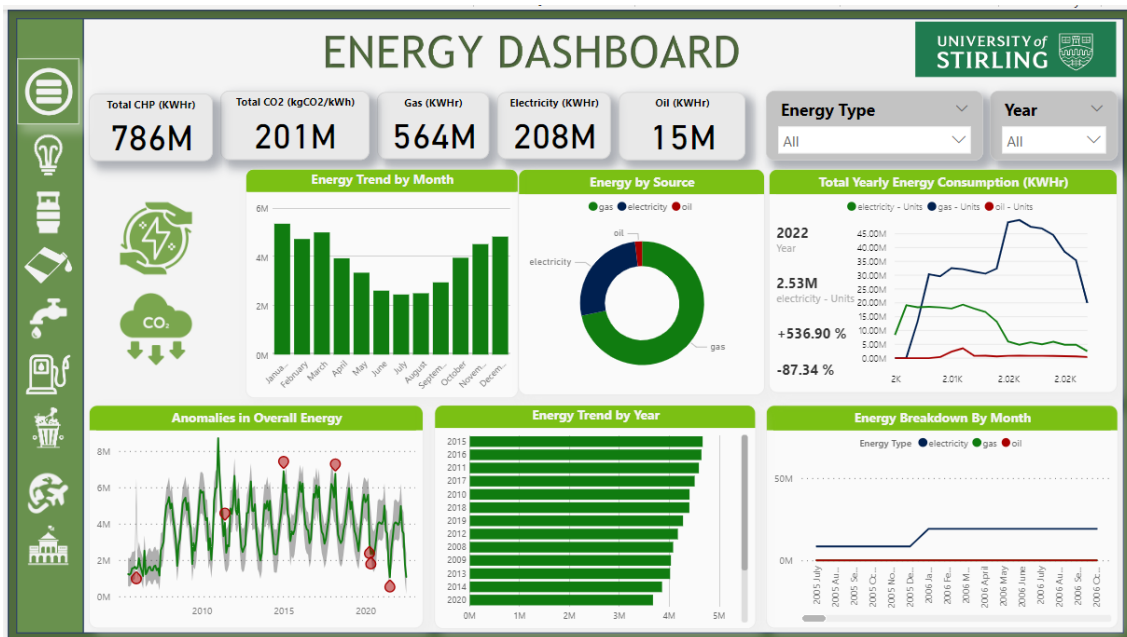


Figure 1.1 The Energy Dashboard for the University of Stirling

1.4 Overview of Dissertation

This section summarises the research journey that follows the project Introduction,

Section 2 – State of the Art: This section delves into the existing research and analyses within the specific field. It also highlights how prior research has contributed to the project's solution and discusses ideas for improvement drawn from these works.

Section 3 – Methodology: This section details the phases of the project, along with the methods and tools employed in designing and constructing the solution.

Section 4 – Data Requirements: This section provides detailed information regarding the scope and context of the data necessary for building the solution. It also outlines the data collection process and the initial preprocessing steps undertaken to create a dataset suitable for the project.

Section 5 – Integration with PowerBI for Visualization: Here, a detailed explanation of the creation of the PowerBI dashboard from scratch is provided, using the data generated in Section 4. This section covers dashboard design specifications and presents a case study showcasing real-life applications of the dashboard and the insights that end-users may derive during their analysis.

Section 6 – Time Series Forecasting Using Machine Learning: This section presents an in-depth exploration of the development of the time series model. It discusses the various models tested, the selection of the final model, and specifications, along with the forecast results.

Section 7 – Conclusion: This final section, provides an overarching summary of the project. Critical evaluation of each phase of the solution, and recommendations for improving existing solution as well as potential future developments for the designed solution.

2 State-of-The-Art

This section is broken down into two sections, the first of which provides a review of research studies on energy consumption and related advancements, and the second of which comprises an analysis of state-of-the-art research that suggests data analytics methods that have already been used in the relevant energy consumption field.

2.1 Carbon Emission Analysis

According to the study carried out in [33],

“Energy-related carbon emissions from UK manufacturing have fallen by approximately 2% per annum over the period 1990–2007.”

While this finding pertains specifically to the manufacturing sector in the UK, this long-term trend suggests a potential gradual decrease in emissions. It serves as a favourable outcome, illustrating how proactive measures and shifts in practices can contribute to enhanced sustainability. Moreover, in the pursuit of addressing global challenges such as climate change and global warming, the study of energy utilization and its optimization represent initial steps towards mitigating excessive energy consumption and its side effects. Accurate energy forecasting should rely on historical data [40] recognizing that each sector exhibits distinct factors impacting carbon emissions.

For instance, [28] provides a recent study of the energy data in the year 2022 on carbon emissions within different service industries in China. The findings indicate that a significant portion of these emissions stems from the combustion of coal and diesel fuel. Notably, this proportion has risen from 27.78% in 2007 to 55.21% in 2017. It's noteworthy that the emission coefficients for coal and diesel are greater than those for natural gas. Natural gas, in contrast, only contributes to 1% of the total emissions in the comprehensive study, with a specific connection to the service industry sector. In addition to that, according to a study cited in [28], key factors of carbon emissions encompass technology, population, economy, and urbanization, underscoring the necessity of developing strategies to cut down carbon emissions, including the implementation of carbon trading mechanisms, a particularly suitable approach for the industrial sector. Likewise, numerous research efforts have delved into comprehending the factors responsible for carbon emissions and developing strategies for mitigation. Similarly, the residential sector assumes a distinct yet significant role in contributing to carbon emissions.

A comparative analysis of residential energy consumption in China, India, and the United States in 2020, as described in [30], highlights the impact of urbanization on increased residential energy usage. This urbanization indirectly led to higher consumption of gas and electricity, while there was a shift away from coal. Specifically, urbanization contributed to a 28% overall increase in energy consumption, with a 32% rise in electricity usage and a 35% increase in gas consumption, coupled with a significant 98% decrease in coal consumption. This reinforces the study conducted in 2006 in [31] which identified key factors driving the rise in carbon emissions, including “technology, affluence, energy and economic structure, and population composition” [31]. It underscores economic growth as the primary driver of carbon emissions.

It is evident that each sector plays a substantial role in carbon emissions, each having its unique contributions. Despite the fact that a study in [28] suggests that coal has a higher conversion coefficient than natural gas, the research presented in [30] argues that the reduction in coal usage and the increased reliance on gas have contributed to higher emissions. This remains a topic of debate, but it is clear that each energy source has its environmental drawbacks that must be addressed appropriately. Moreover, it has motivated our research to establish a path to

identify the primary sources contributing to emissions, compare them with historical data and traditional methods, and seek ways to reduce emissions based on factors such as the type of energy source, emission coefficients, source significance, and other influencing factors.

2.1.1 Carbon Emission in Educational Sector

The study of the educational sector's contribution to carbon emissions presents an intriguing avenue of research. Numerous studies have delved into identifying the primary sources of carbon emissions and have sought to develop strategies for their reduction within educational institutions. In one such study referred from [32], an analysis of carbon emissions within higher education institutions in the UK was conducted. The analysis compared Russell Group universities across three key factors: “institutional targets, the use of normalized or absolute data, the presence of interim targets and monitoring, alignment with sector targets, and commitment from high-level management” [32] and the study has concluded that electricity consumption was a significant contributor to carbon emissions, followed by transportation. It was also noted that reducing overnight electricity usage had a positive impact on emission reduction. Therefore, investigating the primary drivers of emissions at the University of Stirling, as detailed in Section 7.1, could be a fascinating area of research with potential for corrective actions.

Furthermore, a case study examining carbon management in various universities, as discussed in [35], revealed that only a limited number of institutions possess an in-depth understanding of carbon emissions and related activities. The study also identified challenges related to time constraints, study costs, and data reliability during the research process. These challenges warrant attention in our analysis. Additionally, the research outlined in [36] provides extensive research on existing smart energy systems within university campuses. These systems are similar to smart buildings. The research discusses efforts to modernize the smart grid system, known as the smart energy system, which aims to deliver clean and low-carbon electricity. Moreover, the concept of the smart campus, which incorporates technologies such as RFID, IoT, cloud computing, augmented reality, and sensor technologies, is explored. These technologies collectively work towards monitoring environmental conditions and promoting sustainability through energy conservation and optimization.

In a broader context, all these research endeavours converge toward a common objective: enhancing sustainability. Consequently, these papers not only offer unique perspectives on methods for carbon emission analysis and data comparison, but they also contribute to dashboard design by meticulously examining our data from various angles. These insights are invaluable for crafting visual representations, with a particular emphasis on findings across all sectors, notably the educational sector.

2.2 Importance of Data Science in Carbon Emission Analysis

In the previous section, the significance of carbon studies for environmental sustainability has been highlighted, prompting an exploration into the role of data science in the sustainability domain. Various studies have been conducted in this regard, and we delve into a selection of them here. The study in reference [37], focuses on identifying feature variables applicable to CO₂ analysis emphasising the significance of certain variables, including energy consumption, income, industrialisation, and economic factors. Researchers employed Gaussian regression to overcome over-fitting challenges that are commonly encountered with conventional linear regression, and this resulted in a significant improvement in accuracy. While another study in [38], provides an extensive analysis of traditional machine learning techniques for forecasting energy consumption data. The analysis employs data analytics platforms like RapidMiner Studio and IBM SPSS (Statistical Package for Social Sciences) Modeler. The study compares various models,

including Artificial Neural Networks (ANN), Support Vector Machine (SVM), Classification and Regression Trees, Linear Regression, and SARIMA. Among these, ANN and SARIMA demonstrate a better performance. The research underscores the significance of feature variables in prediction accuracy and emphasizes that model selection should align with specific objectives, data size and type, and the nature of feature variables. Furthermore, it stresses the importance of iterative testing, validation, and evaluation to assess the model's fit with the data. [39] emphasize the insights into the design criteria for visualizing specific energy data, offers an in-depth analysis of both functional and non-functional criteria, encompassing modes, techniques, displayed information, and considerations related to hardware and software.

Additionally, in [41] an in-depth analysis of various anomaly detection techniques using Artificial Intelligence is conducted. This study explores how data sources, such as appliance-specific parameters like the presence of individuals, appliance operation and standby times, or ambient conditions, influence anomalies in energy consumption. This approach goes beyond traditional anomaly detection, which solely relies on consumption data, by emphasizing the importance of analysing data at the appliance level rather than aggregating it. Also, the research assesses the contributions of multiple anomaly detection methods to the same dataset. It's worth highlighting that this approach not only enhances anomaly detection but also contributes to the development of predictive models. However, a limitation is that only consumption data is available for analysis, without access to other relevant features or factors responsible for consumption patterns. This limitation confines the research to some extent, as only traditional methods can be applied to the specific dataset. Nevertheless, this limitation serves as a suggestion for future advancements in the current project. Furthermore, in [42] a thorough examination of various predictive modelling techniques, including SARIMA, SARIMAX, Auto-ARIMA, and LSTM (Long Short Term Memory), has been conducted. The paper emphasizes that the nature of the data, the presence of external variables, and the specific characteristics of the data significantly impact forecasting outcomes. This is demonstrated by the comprehensive testing of each model using diverse datasets, highlighting the need to consider these factors when selecting an appropriate forecasting approach.

In summary, several recommendations and findings exist for selecting suitable models and visualization techniques, accompanied by critical considerations when choosing tools or models. Given the diverse nature of datasets, especially time series data like energy consumption, the selection of technology tailored to specific data characteristics is pivotal.

3 Methodology

The overarching goal of this project is to gain comprehensive insights into energy usage and associated carbon emissions at the University of Stirling. The methodology for this is created to effectively address the objectives by combining the strength of PowerBI visualisation and machine learning techniques. It comprises two distinct yet interlinked phases that work together to help us comprehend trends and patterns holistically.

3.1 Interactive Dashboard

This stage has been focused on the Power BI dashboard-enabled visualisation of historical time series data comprising energy consumption and carbon emissions. The procedure involves collecting the data, preparing it, integrating it, and then designing the dashboard. Data visualisation enables a thorough understanding of the dynamics of energy usage and how each of us contributes to it. This awareness also gives us the information we need to make well-informed choices that will reduce our energy consumption. A notable aspect of the dataset is that it contains timestamps, numerous locations, a variety of buildings, and a range of energy kinds, which adds levels of complexity. As a result, the decision to use Power BI visualisation is appropriate as it gives end users a thorough and quick solution to identify existing trends and patterns.

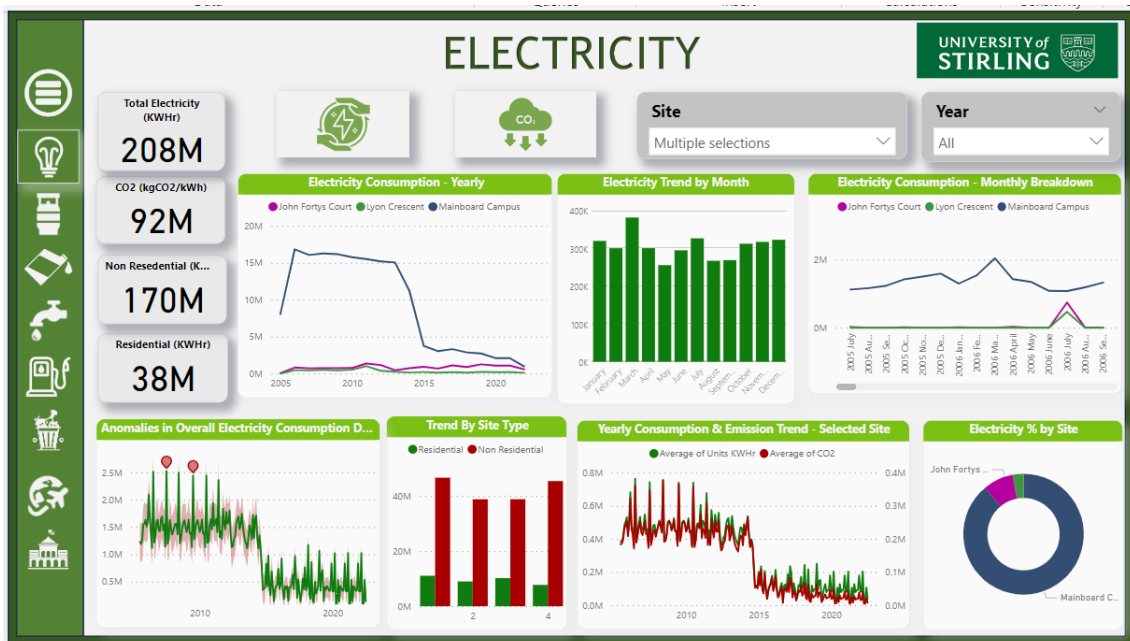


Figure 3.1 Visualization Sample on the Energy Dashboard

3.1.1 Intra-Campus Variability

The primary objective of the dashboard is to provide a comprehensive platform for analytics and visualisation of energy use and its associated carbon emissions. Specifically, the focus is on key CHP components, together with a detailed analysis of trends in data on electricity, gas, oil, water, fleet, waste, and travel. This detailed investigation covers a wide range of campus buildings, including both residential and non-residential ones. The dashboard serves as an essential tool, offering a thorough view of consumption and emission patterns.

3.1.2 CO2 Emission Benchmarking among Scottish Universities

The dashboard broadens its application by offering a thorough examination of trends that have been exhibited by various Scottish universities. This aspect makes it possible to evaluate the University of Stirling with other academic institutions. This section, which focuses mostly on CO2 emissions, also explains the sources of emissions and the full range of their potential effects; however, the data spans the time frame of 2021-2022. A similar comparison analysis enables the University of Stirling to determine its relative place within the greater academic landscape and permits well-informed judgements regarding emission reduction methods.

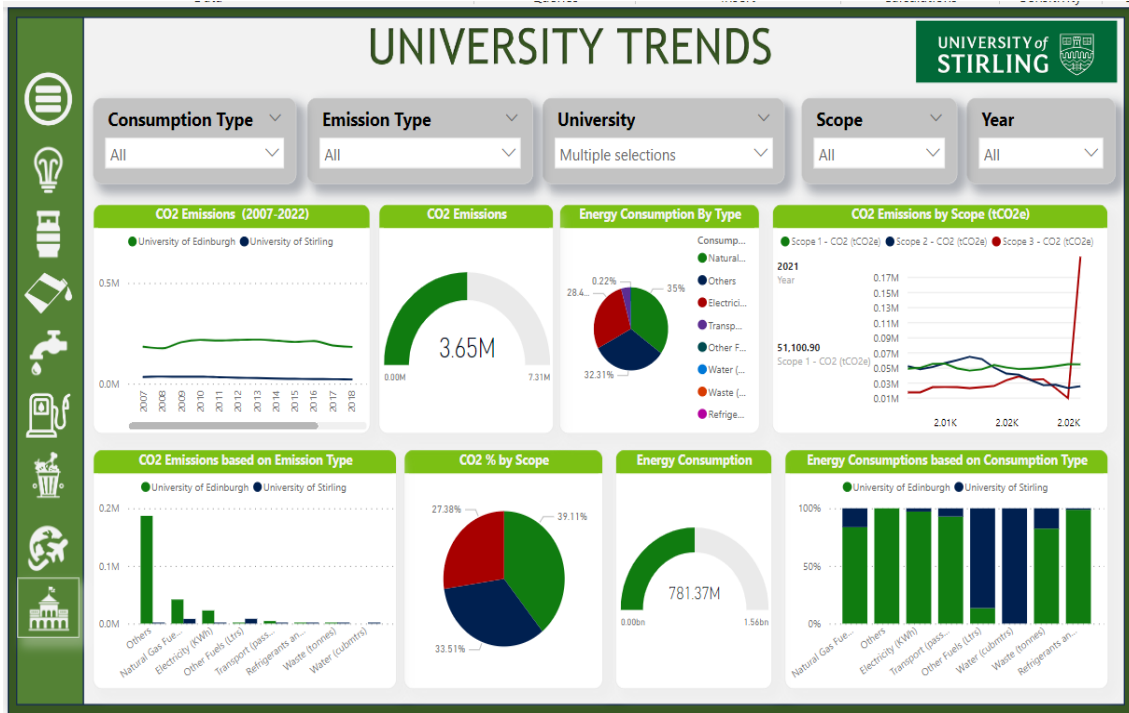


Figure 3.2 Dashboard showing trends across Scottish universities

3.2 Energy Consumption Forecast

The study advances by incorporating machine learning approaches to estimate energy use, which consists of the deployment of sophisticated computational algorithms for predictive analytics, following the study of data patterns through visualization. This attempt seeks to forecast energy usage patterns over the upcoming five-year interval. The dataset's relatively modest size is a significant constraint due to its limited scope. It spans the years 2005 to 2022 and provides monthly consumption data comprising 205 rows. This limits its ability to incorporate the diversity and temporal dynamics needed for reliable energy consumption trend analysis. As a result, a wide range of time series analysis approaches, including ARIMA, SARIMAX, Prophet, and exponential smoothing, have been tested to attain the best accuracy and selectivity.

3.3 Process Workflow

The diagram below illustrates the progression of stages carried out throughout the project detailed in Section 5 and 6.

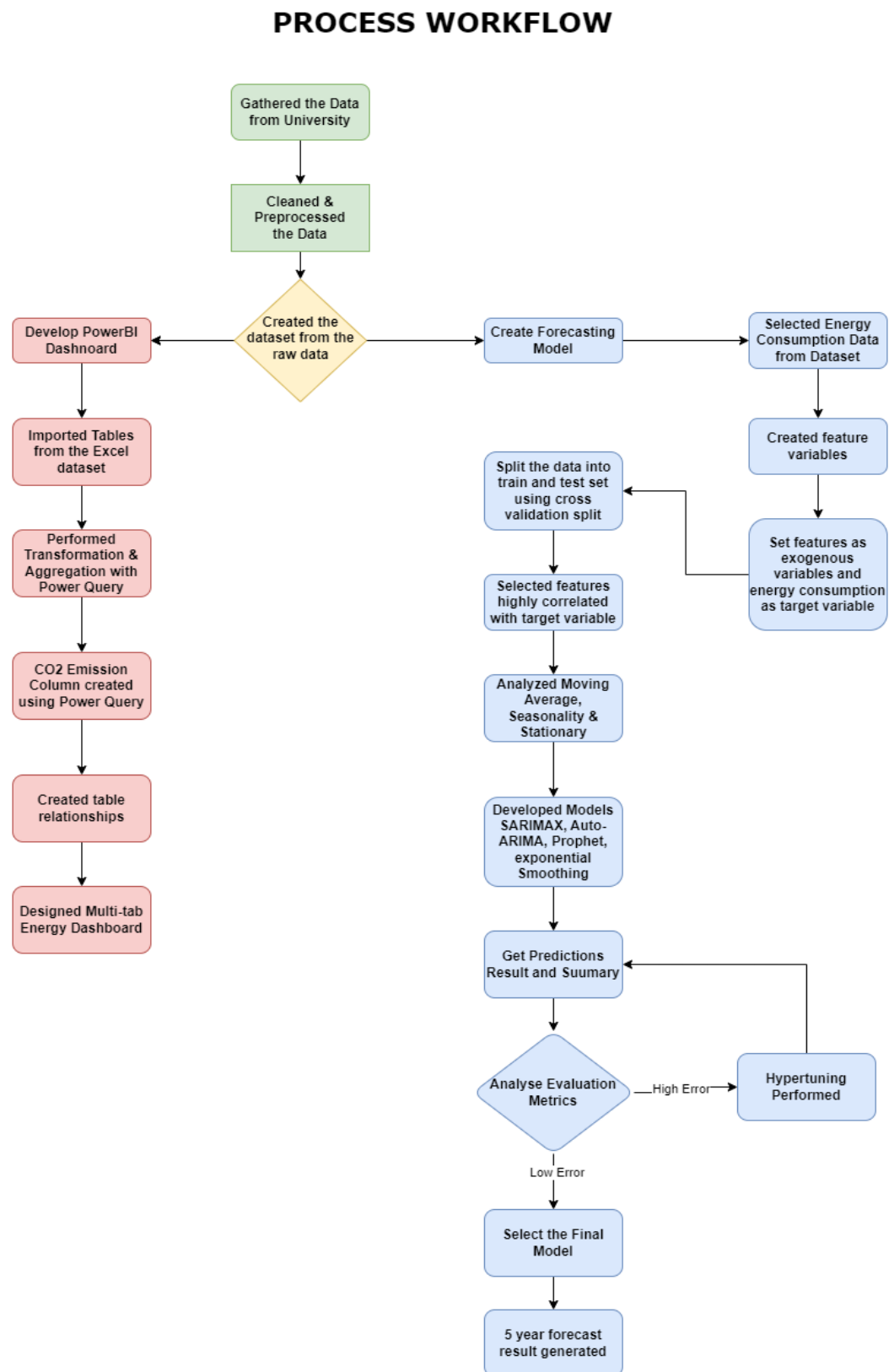


Figure 3.3 The process workflow of the project

4 Data Preparation & Integration

This section delves into the methods employed for structuring, refining, and aligning the data, ensuring its relevance and quality for subsequent analyses. It navigates through data scope, collection, preprocessing, and integration, establishing a solid groundwork.

4.1 Scope and Context of Data

The primary focus of the data analysis pertains to the educational institution University of Stirling, and data has been specifically sourced from the University campus, encompassing a temporal range spanning from 2005 to 2022. However, the raw data has been collected from two main channels: one of the datasets from the University's utility records provided by the Safety, Environment, Security, and Continuity team, containing monthly consumption data spanning a decade for analysing consumption trends at the University of Stirling. Each year's dataset covers monthly measurements of both direct and indirect emissions over this time frame (Table 4.1), where direct emissions are produced on-site from university-controlled activities and indirect emissions are generated from sources that are either purchased or not under university control [5]. These measures cover a range of buildings on the university campus, both residential and non-residential.









Name	Type	Date modified	Compressed size
 Business Travel Figures Submitted 2023 f...	Microsoft Excel Worksheet	20/06/2023 12:18	25 KB
 CHP for Period 2021-22.xlsx	Microsoft Excel Worksheet	20/06/2023 12:18	39 KB
 Electricity for Period 2021-22.xlsx	Microsoft Excel Worksheet	20/06/2023 12:18	114 KB
 EV energy from Charge Place Scotland - ...	Microsoft Excel Worksheet	20/06/2023 12:18	336 KB
 Fleet Transport 2021-22 Figures.xlsx	Microsoft Excel Worksheet	20/06/2023 12:18	29 KB
 Gas for period 2020-21.xls	Microsoft Excel 97-2003 Wor...	20/06/2023 12:18	355 KB
 Gas for Period 2021-22.xlsx	Microsoft Excel Worksheet	20/06/2023 12:18	31 KB
 ghg-conversion-factors-2022-full-set.xls	Microsoft Excel 97-2003 Wor...	20/06/2023 12:18	3,232 KB

Figure 4.1: First set of raw data from the university's utility records

In addition, the second set of data has been collected from the open data sourced from the Sustainable Scotland Network [8]. The latter serves as a benchmark for carbon emissions assessment among Scottish universities, with a specific focus on data from the year 2022.







 Edinburgh Napier University PBDR 2022.xlsx	18/07/2023 12:42	Microsoft Excel Work...	8,018 KB
 Glasgow Caledonian University PBCCD 2022.xlsx	18/07/2023 12:42	Microsoft Excel Work...	7,946 KB
 Heriot Watts University PBCCD 2021_2022.xlsx	18/07/2023 12:43	Microsoft Excel Work...	7,996 KB
 Stirling University PBCCD 2022.xlsx	18/07/2023 12:41	Microsoft Excel Work...	7,968 KB
 Strathclyde University PBCCD Report 2022.xlsx	18/07/2023 12:44	Microsoft Excel Work...	8,002 KB
 University of Aberdeen PBCCD 2022.xlsx	18/07/2023 12:43	Microsoft Excel Work...	7,978 KB

Figure 4.2 Second set of raw data collected from public sources

The data has been identified as time series data and the variables recorded are,

Scope (Type)	Emission Source
Scope 1 (Direct)	Fuels (Natural Gas, Burning Oil, Gas Oil, Diesel, Petrol)
Scope 2 (Indirect)	Electricity
Scope 3 (Indirect)	Water, Waste, Transport

Table 4.1: Breakdown of Variables in the dataset

In addition, the primary form of energy consumption, CHP, which includes electricity, gas, and oil, has been used to calculate carbon emissions by applying consumption figures to GHG conversion coefficients provided by the UK government [6].

UK Government GHG Conversion Factors for Company Reporting						
Fuels						
For more information refer to the 'Outside of scopes' tab for guidance.						
Activity	Fuel	Unit	Total kg CO ₂ e per unit	kg CO ₂ e of CO ₂ per unit	kg CO ₂ e of CH ₄ per unit	kg CO ₂ e of N ₂ O per unit
Gaseous fuels	Butane	tonnes	3033.32	3029.26	2.25	1.80
		litres	1.75	1.74	0.00	0.00
		kWh (Net CV)	0.24	0.24	0.00	0.00
		kWh (Gross CV)	0.22	0.22	0.00	0.00
	CNG	tonnes	2539.25	2534.47	3.44	1.34
		litres	0.44	0.44353	0.00060	0.00023
		kWh (Net CV)	0.20	0.20188	0.00028	0.00011
		kWh (Gross CV)	0.18	0.18219	0.00025	0.00010
	LNG	tonnes	2559.17	2554.39	3.44	1.34
		litres	1.16	1.15583	0.00156	0.00061
		kWh (Net CV)	0.20	0.20347	0.00028	0.00011
		kWh (Gross CV)	0.18	0.18362	0.00025	0.00010
	LPG	tonnes	2939.29	2935.18	2.28	1.83
		litres	1.56	1.55491	0.00121	0.00097
		kWh (Net CV)	0.23	0.22999	0.00018	0.00014
		kWh (Gross CV)	0.21	0.21419	0.00017	0.00013
		tonnes	2539.25	2534.47	3.44	1.34

Figure 4.3: A sample of GHG Conversion Factor from the year 2022

4.2 Data Collection & Integration

The approaches employed in collecting data are outlined in this section, followed by the preprocessing stages. In addition, the raw data is split up into many Excel files, with each file containing the monthly energy source data for an academic year, as shown in Figure 4.4.

Invoice Apportioned Electricity Usage and Cost per Data Set Accrual from Aug 2021 for 12 month(s)									
Name	Code	Reference	To Date kWh	Accrual kWh	Final kWh	To Date Cost(£)	Accrual Cost(£)	Final Cost(£)	Last Date
1 Airthrey Castle Yard (Principal's House)	ACY1	6689410000	20,944	1,346	22,290	3,437.90	254.62	3,692.52	30/06/2022
Airthrey Castle	ACAS	0733510000	68,713	0	68,713	13,845.70	0.00	13,845.70	31/07/2022
Airthrey Cottage	ACOT	6582310000	1,004	118	1,121	274.57	37.43	311.99	27/06/2022
Alangrange	ALANG	4928410000	69,645	0	69,645	11,486.67	0.00	11,486.67	01/08/2022
Buckieburn Fish Farm		0001057639	82,240	14,140	96,380	15,645.15	2,950.30	18,595.45	02/06/2022
Centro House		7038487970	7,510	0	7,510	788.92	0.00	788.92	06/05/2022
Friarscroft	FRIAR	0561510000	18,699	1,866	20,565	3,106.44	346.86	3,453.30	27/06/2022
Gardens and Grounds	GMA	3244510000	88,920	0	88,920	17,492.96	0.00	17,492.96	01/08/2022
John Forty's Court	JFC	2171510000	2,418	172	2,590	416.74	41.11	457.85	27/06/2022
John Forty's Court	JFC	8671510000	4,776	421	5,197	729.24	76.78	806.03	27/06/2022
John Forty's Court	JFC	7371510000	21,113	2,977	24,091	3,535.13	542.06	4,077.19	27/06/2022

Figure 4.4: Sample raw dataset with Electricity Consumption Data (2021-2022)

Additionally, data integration includes combining annual consumption figures for each form of energy into a single time series dataset. Additionally, a secondary phase of this process includes the inclusion of key CHP data. This aggregated data is then used as the input for a predictive forecasting method. Additionally, consistent timestamps have been ensured to attain temporal synchronisation, which helps with the interpretation and comparison of our time series data.

4.3 Data Preparation & Preprocessing

This section details the initial data cleaning process executed to merge multiple datasets based on their respective energy sources. Each file comprises the monthly consumption dataset, as illustrated in Figure 4.1.

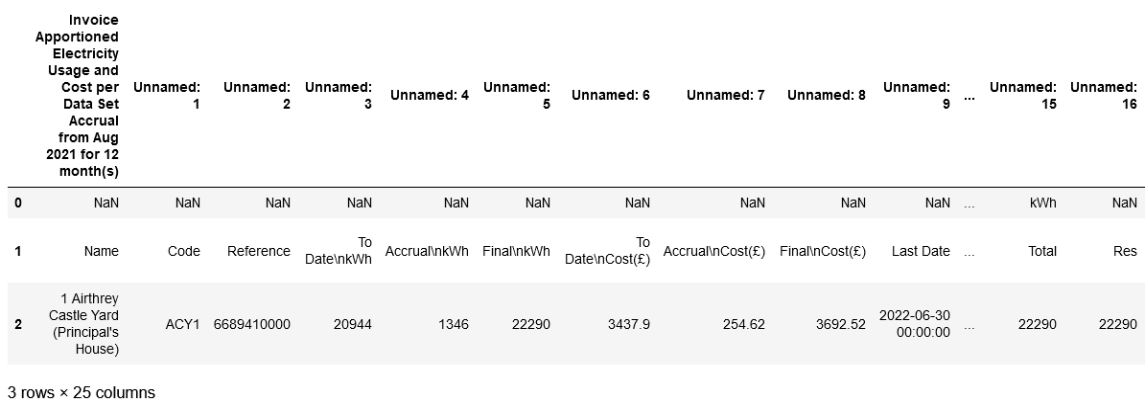


Figure 4.5: Raw data to pandas data frame - before cleaning

Notably, the files present a slightly distinctive format in terms of data arrangement. The process of data preprocessing was conducted using Jupyter Notebook, the Python programming language, and the Excel application.

The data cleaning has been carried out as given in Table 4.1.

No:	Steps
1	Libraries Pandas and NumPy were imported into Jupyter Notebook
2	Converted Excel sheet into a data frame, as presented in Figure 4.5
3	Only the columns <p><i>['Invoice Apportioned Electricity Usage and Cost per Data Set Accrual from Aug 2021 for 12 month(s)', 'Unnamed: 9', 'Unnamed: 15']</i></p> <p>were retained, while the remaining columns were deleted.</p>
4	Renamed the columns as below, <p><i>{'Invoice Apportioned Electricity Usage and Cost per Data Set Accrual from Aug 2021 for 12 month(s)': 'site_name', 'Unnamed: 15': 'KWH', 'Unnamed: 9': 'date_time'}</i></p>

5	Converted the <i>date_time</i> column to the appropriate date format using the pandas <i>.to_datetime()</i> function and extracted the <i>month and year</i> using <i>.dt.to_period('M')</i> function to a new column named <i>month_year</i> illustrated in Figure 4.6.
6	Total KWH values were aggregated for each site as shown in Figure 4.6 based on ['site_name', 'month_year'] columns using <i>groupby()</i> and <i>sum()</i> functions. <pre>df.groupby(['site_name', 'month_year'], as_index=False)['KWH'].sum()</pre> <p><i>An example of this is the first row with a summation of all the KWH values for site '1 Airthrey Castle Yard (Principal's House)' during the month of 'June 2022' (Figure 4.6)</i></p>
7	These steps were executed across all files, with slight adjustments made to accommodate changes in the data format. Nevertheless, this serves as an example of the data-cleaning approach.
8	Exported the dataframe (Figure 4.6) into .xlsx file using <i>df.to_excel()</i> function to further process in Excel as mentioned in Step 9.
9	Performed Excel-based consolidation and transposition of the data on each file generated in step 8 to produce distinct sets of files for Electricity, Gas, Oil, and so on. (as in Figure 4.7)
10	In the subsequent processing stage, Python was employed to perform various tasks, including filling in missing values, eliminating outliers, renaming columns, and scaling the data. The primary objective was to improve the data quality for each energy type generated in Step 9, and this process is elaborated upon in section 6.1.

Table 4.2: Initial Data cleaning stages to create a dataset for each energy source

An example of the cleaned and extracted data, shown in Figure 4.7, shows electricity consumption statistics for the year 2021. Each year's dataset has been subjected to the same Python and Excel-based data cleaning process, as shown in Table 4.2. These distinct datasets are then integrated into an Excel spreadsheet, which serves as the structure for visualisation and modelling. It's worthwhile to note, though, that additional data transformations were carried out at subsequent stages to meet specific requirements.

	site_name	month_year	KWH
0	1 Airthrey Castle Yard (Principal's House)	2022-06	22290
1	Airthrey Castle	2022-07	68713
2	Airthrey Cottage	2022-06	1121
3	Alangrange	2022-08	69645
4	Buckieburn Fish Farm	2022-06	96380

Figure 4.6: Sample Annual Data (2021-2022) for Electricity after cleaning - Stage 1

	A	B	C	D	E	F
1	date_time	Airthrey Castle	Airthrey Cottage	1 Airthrey Castle	Alangrange	Friar
2	Jul-05	11622	1044	0	6278	4343
3	Aug-05	9750	0	0	5996	0
4	Sep-05	11800	0	0	5858	0
5	Oct-05	11820	1281	0	7351	4346
6	Nov-05	11300	0	0	9266	0
7	Dec-05	12728	0	0	10569	0
8	Jan-06	14604	1715	0	11391	0
9	Feb-06	14720	0	0	8398	0

Figure 4.7: Sample Combined Electricity data (2005-2022) after cleaning - Stage 2

Similar to the above, a separate dataset has been developed to analyse trends among Scottish Universities. This dataset was created employing emission data from 11 universities, including the University of Stirling, for the years 2021 to 2022 using publicly available data from SSN. The SSN public data emphasised the CO2 emissions provided by the organizations. As a result, the final dataset centralises CO2 emissions produced by various energy sources, as shown in Figure 4.8.

A	B	C	D	E	F	G	H
date_time	Type	UOS Consumpt	UOS Emissions	ENU Consumpt	ENU Emissions	GCU Consumpt	GCU Emissions
01/09/2021	Electricity (KWh)	6312242	1220.661358	12455048	1314.443	44117323	2109.816174
01/09/2021	Natural Gas Fuel (KWh)	44679330	8155.764898	10165963	1855.695	22600227	4125.445437
01/09/2021	Other Fuels (Ltrs)	1467667.8	8337.94482	7076.18	16.55115	2435.18	3.856959843
01/09/2021	Refrigerants and process (Kg)	3.61	14.15842	0	0	774.37	268.58337
01/09/2021	Transport (passenger km)	2233179	326.4956788	1421231.06	217.557	88990356.87	10683.53811
01/09/2021	Waste (tonnes)	764.007	16.25821702	198.64	4.426741	381.6115	13.90926005
01/09/2021	Water (cubmtrs)	384126	69.88542	35792.25	6.029618	36569.325	6.20936385
01/09/2021	Others	10561.68	109.8146923	0	0	758.862261	13790.03047

Figure 4.8 University Trend Dataset

- Using the 'Unpivot Columns' feature, data in Figure 4.7 has been flattened. These merged similar values into a single column, allowing for the creation of a matrix format consistent within all the tables. Subsequently, the headers have been renamed and reassigned site names, resulting in the configuration shown in Figure 5.4.

date_time	Airthrey Castle	Airthrey Cottage	Airthrey Castle Yard	AlanGrange	FriarsCroft & FriarsView
01/07/2025	11622	1044	0	6278	4343
01/08/2025	9750	0	0	5996	0
01/09/2025	11800	0	0	5858	0
01/10/2025	11820	1281	0	7351	4346
01/11/2025	11300	0	0	9266	0
01/12/2025	12728	0	0	10569	0

Figure 5.3 Sample of the data added to PowerBI

	date_time	Site	Units KWHr
1	01/07/2005	Airthrey Castle	11622
2	01/07/2005	Airthrey Cottage	1044
3	01/07/2005	Airthrey Castle Yard	0
4	01/07/2005	AlanGrange	6278

Figure 5.4 A sample after transforming using 'Unpivot Columns'

- The column 'Year' has been created from date_time column for connecting the tables.
- A common column named 'Merged' has been created with values in the format 'YYYY-EnergySource' to establish links between the tables for necessary aggregations.
- CO2 conversion rates have been separately added, and calculated carbon emissions using the formula below. This involved multiplying energy consumption by CO2 factors, using the common column 'Merged' as in like Figure 5.5 and value lookup has been referred from [53].

$$CO2 = 'elec'[Units KWHr] * LOOKUPVALUE('co2 rates'[co2 factor], 'co2 rates'[Merged], 'elec'[Merged])$$

date_time	Site	Units KWHr	Year	Merged	CO2
01 March 2022	Airthrey Castle Yard	1000	2022	2022electricity	193.38
01 March 2022	CentroHouse & ScionHouse	39053	2022	2022electricity	7552.069
01 April 2022	CentroHouse & ScionHouse	23720	2022	2022electricity	4586.973
01 May 2022	CentroHouse & ScionHouse	18519	2022	2022electricity	3581.204
01 June 2022	Airthrey Castle Yard	3506	2022	2022electricity	677.9902
01 June 2022	CentroHouse & ScionHouse	16352	2022	2022electricity	3162.149

Figure 5.5: After Data Final Transformation in PowerBI

- And also, created a separate table, for total energy consumption data by adding all the consumption values.

5.2 Design Principle

The Energy dashboard has been designed to provide the institution with a comprehensive insight into energy consumption patterns spanning various ranges such as monthly, yearly, and overall trends. The design prioritizes user-friendliness and includes visualizations, complemented by interactive features like slicers and toggles, which facilitate tailored exploration. For instance, it becomes possible to pinpoint peak usage periods, enabling informed decision-making based on the derived analysis. This process inherently contributes to scrutinizing and managing historical consumption trends, leading to an enhanced understanding of costs as well. Moreover, the selection of PowerBI as the visualization tool arises from the necessity to merge various datasets stemming from diverse energy sources into a unified dashboard. PowerBI possesses the capability to consolidate our data comprehensively, provide a centralized platform [7], and ensure that the end user can access and understand the complex data in one location.

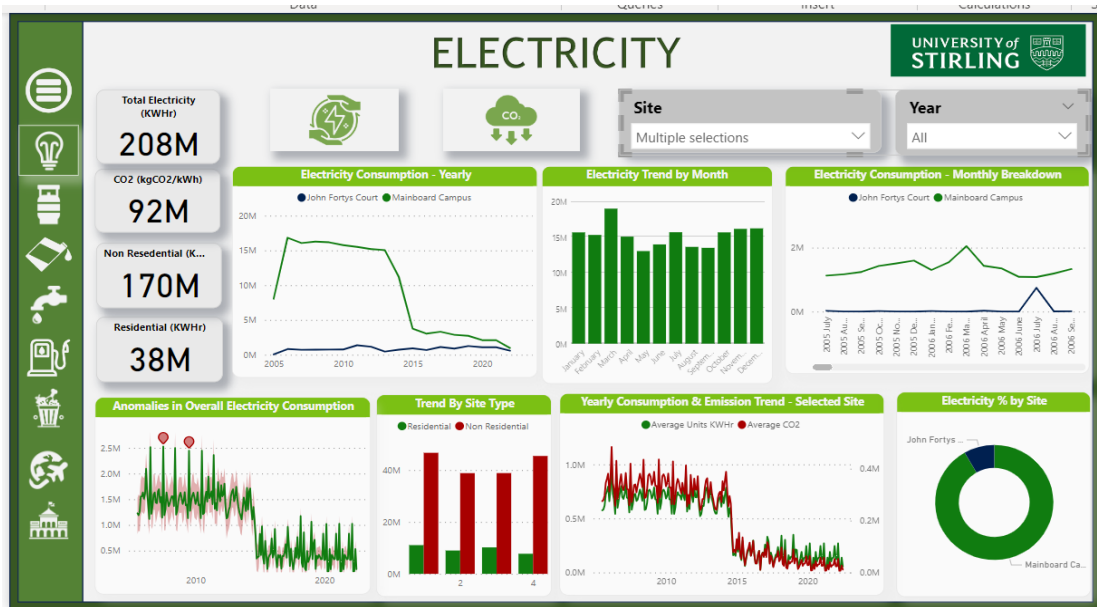


Figure 5.6 Layout of the Dashboard

The data has been imported from the Excel spreadsheet and transformed using Power Query Editor performing the transformation & aggregation operations listed in Section 5.1 Figure 5.5 to enhance the accuracy of the visual representation. Figure 5.6Figure 5.6 illustrates the general layout of the dashboard and various features that have been strategically used to align with different requirements is given below.

1. Tabs Available: The Overview tab for the dashboard is followed by the separate pages for Electricity, Gas, Oil, Water, Fleet, and University Trends. The Waste and Travel segments have also been added in response to feedback from end-users.
2. The dashboard also hosts a navigation bar on the left side, facilitating the selection of the desired page.



Figure 5.7 Sample of Navigation Pane

3. Slicers for 'Site', 'Year', or 'Consumption Type' are integrated to enable the customization of the visualization based on the chosen type.

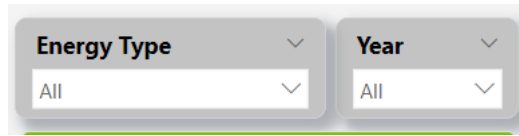


Figure 5.8 Sample of Slicers on Dashboard

4. The toggle button positioned at the top facilitates seamless switching between electricity consumption and its corresponding CO2 emissions.

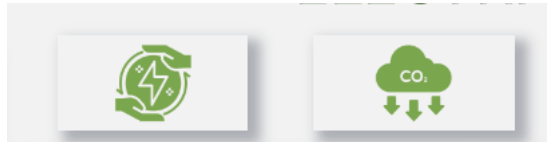


Figure 5.9 Sample of Toggle of Energy and CO2 page

5. The cards on the top/left display the comprehensive data, including total consumption and emissions, as well as consumption specific to residence/non-residential structures or the types of consumption.

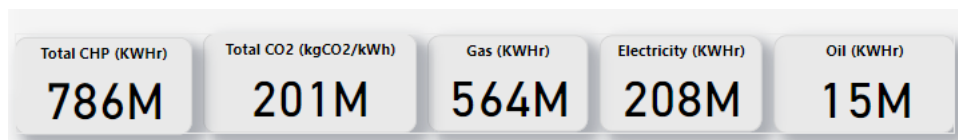


Figure 5.10 Sample of Display cards on the dashboard

6. Line Charts have been chosen to show the trend lines, and how energy usage has changed over time enabling us to find anomalies in the pattern easily.
7. Bar graphs and pie charts have been used to visualize the comparison between the energy usage of different categories and periods, which will also help us find the outliers.
8. Scatter plots have been used to show the relationship between different variables in the data.
9. The dashboard also provides insights into yearly and monthly consumption trends, usage categorized by building type, and the months with heightened electricity consumption.
10. Additionally, the utilization of PowerBI's analytics capabilities enables the detection of anomalies in the data, thereby exposing uncommon usage patterns within certain years and also a projection of CO2 emissions has been generated, enhancing the insights provided by the dashboard.

5.3 Case Study: Dashboard Analysis

The section presents a case study centred on the dashboard, providing a comprehensive investigation into key trends related to CHP comprising Electricity, Gas, and Oil, as it encompasses the most prominent emission data within the entire dataset.

5.3.1 Analysing Monthly Trends

It is evident that June, July, and August exhibit the lowest CHP consumption (Figure 5.11) throughout the year, in contrast to January and December, which manifest the highest consumption. Notably, the electricity (Figure 5.12) consumption reaches its peak in March, closely followed by July. While gas (Figure 5.13) follows a similar trajectory to the overarching CHP pattern, oil (Figure 5.14) displays a distinct characteristic, peaking in January with a substantial disparity of 3 million units compared to other months.

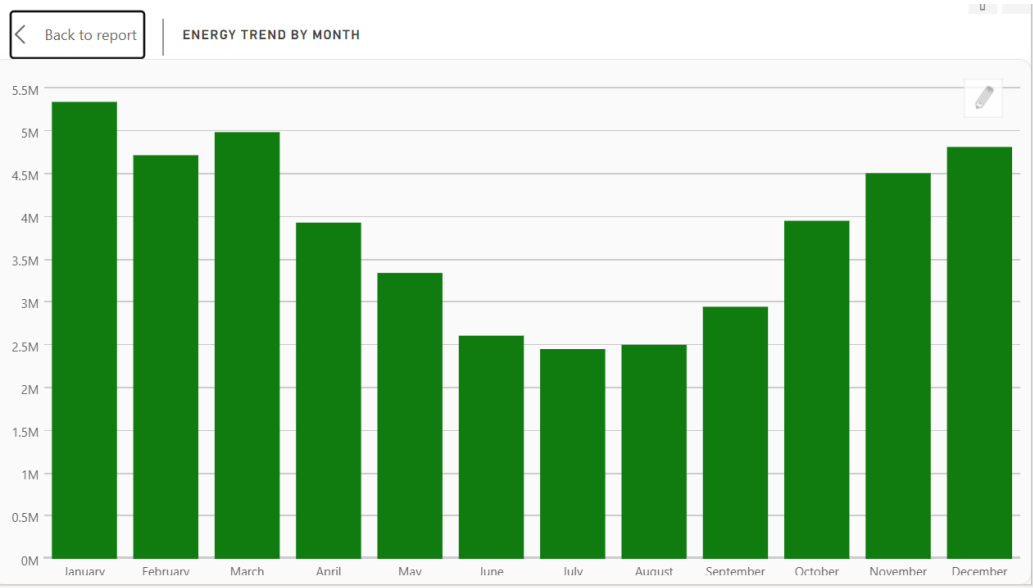


Figure 5.11 CHP Monthly Trend

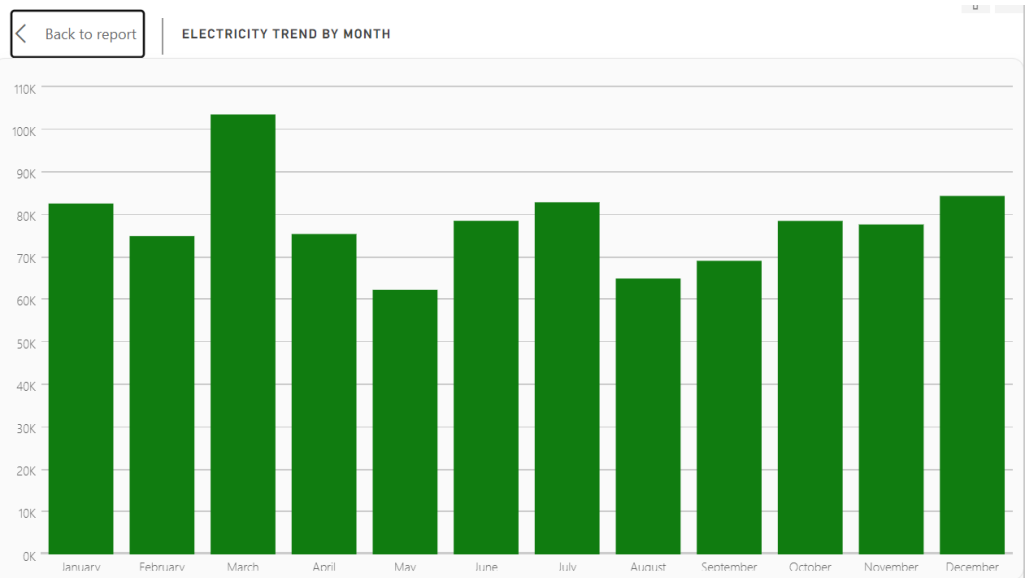


Figure 5.12 Electricity Monthly Trend

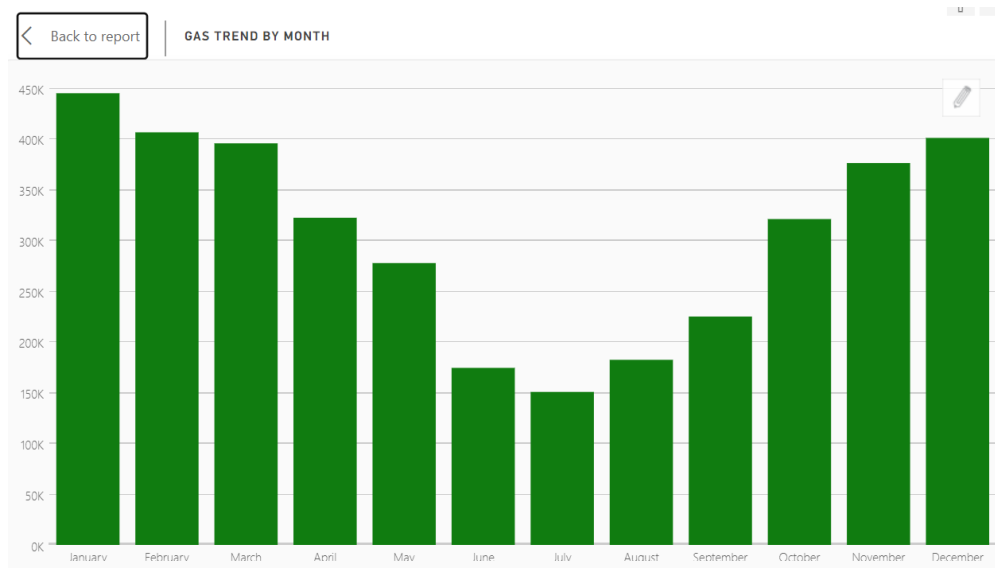


Figure 5.13 Gas Monthly Trend

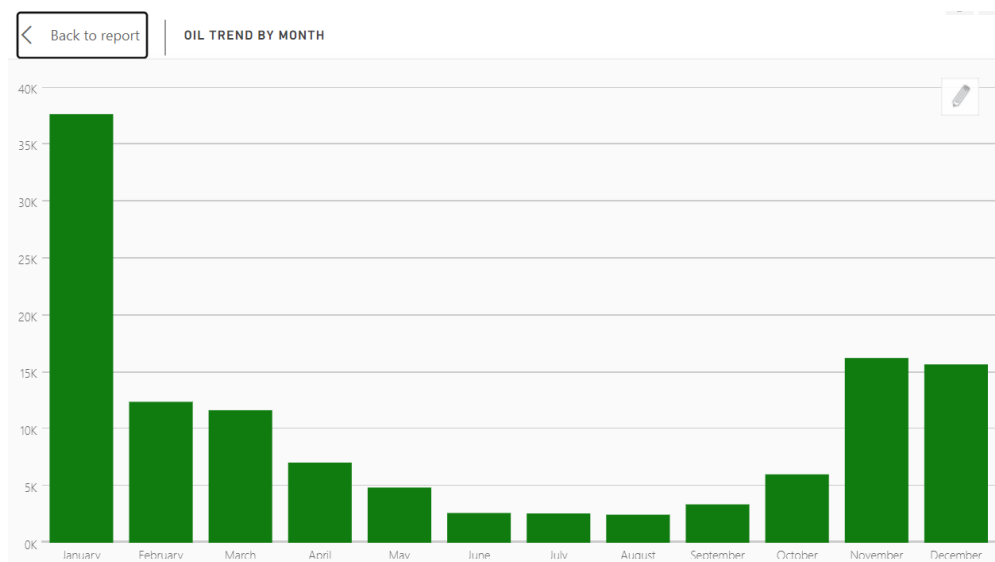


Figure 5.14 Oil Monthly Trend

5.3.2 An Example Based on Building

The possible reduction in total consumption from June to August can be attributed to the summer season and holidays, accompanied by a comparable trend in gas usage, likely due to the impact of summer. The distinct electricity trend may be attributed to the behaviour of residence halls. A comparison of the residence hall against the main board data reveals dissimilarities. While the main board electricity aligns with the CHP trend, residence buildings diverge. This deviation is rationalized by the continuous electricity consumption within accommodations such as lighting, kitchen appliances, cookers, and electric showers contribute to year-round electricity use. This disparity is shown in Figure 5.15 and Figure 5.16.

In response to this situation, potential measures such as the installation of renewable energy sources or energy-efficient appliances can be involved. Alternatively, orchestrating an awareness campaign to foster responsible appliance usage, or deploying visual reminders, such as posters, to encourage prudent electricity utilization among residents, can be considered. Additionally, the implementation of regulations to govern electricity usage may yield some efficacy. The introduction of sensor-activated lighting in communal spaces, triggered by human presence, stands as another feasible solution. It is worth noting that these practices should extend to the

Main Campus, where the consumption trend bears similarities. The substantial consumption value of approximately 900K underscores the importance of applying these practices across the board.

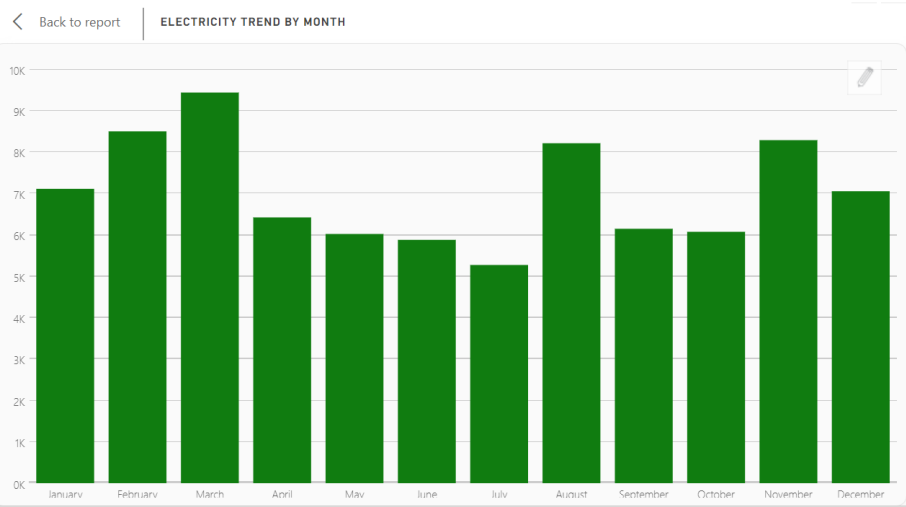


Figure 5.15 Electricity Consumption in Alan Grange

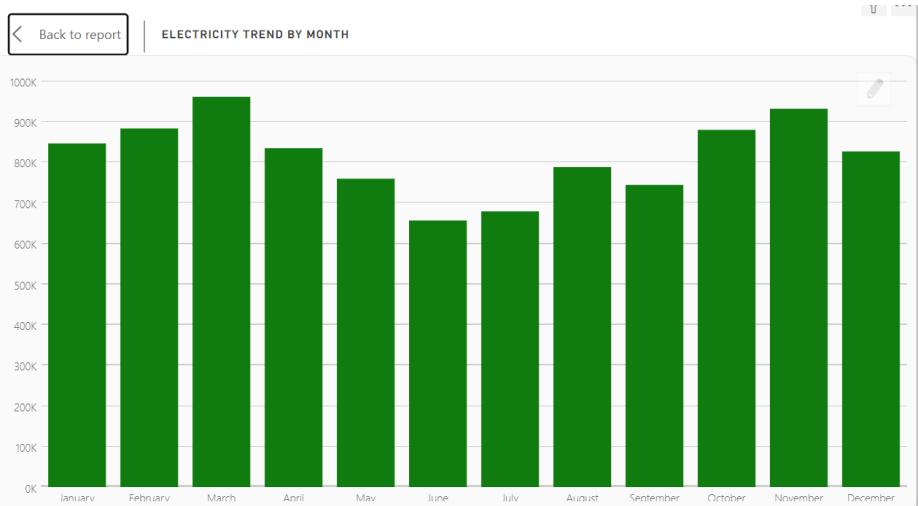


Figure 5.16 Electricity Consumption in Campus Mainboard

5.3.3 Analysing Yearly Trend

The examination of CHP consumption across different years presents in Figure 5.17 and Figure 5.18 a significant area for analysis. Notably, a substantial surge in the gas trend is observed from 2015, followed by a gradual decline. Conversely, electricity consumption remains lower than gas, with emissions from electricity displaying a peak during the initial years, subsequently exhibiting a substantial reduction in 2015 followed by a gradual decline. Furthermore, recent years show gas consumption as the highest among the three energy sources, despite its gradual reduction. Therefore, investigating potential wastage points for gas is advisable, with a focus on ventilation systems. Regular monitoring of consumption data is recommended, along with the installation of sensor-equipped heating systems that automatically deactivate heaters when rooms are unoccupied.

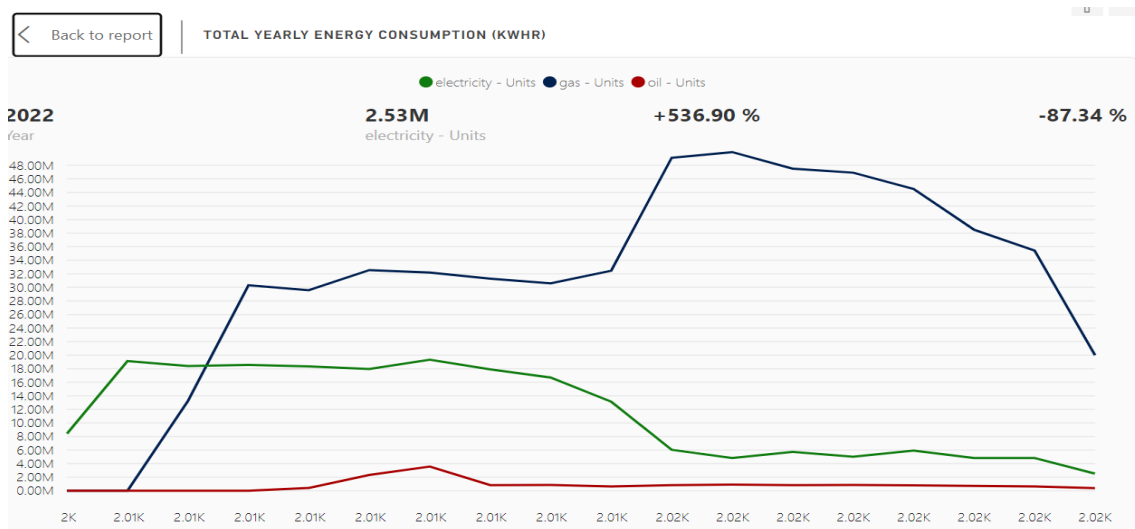


Figure 5.17 Yearly CHP Consumption Trend

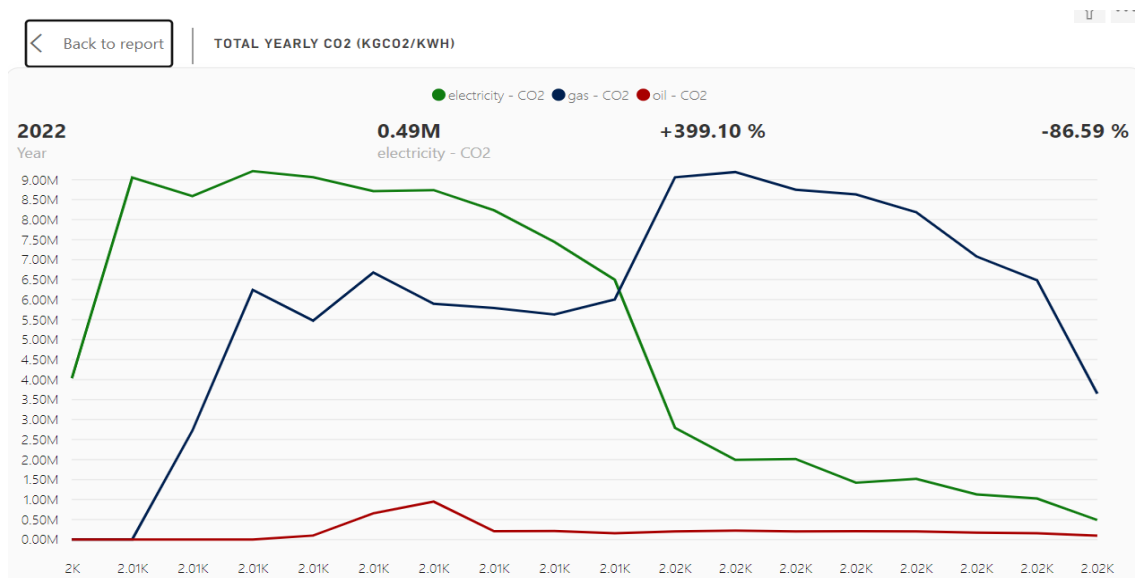


Figure 5.18 Yearly CO2 Emissions from CHP

5.3.4 Conclusion from the Case Study

In conclusion, it is anticipated that this brief analysis has helped to shed light on the notable trends in electricity and gas consumption, revealing instances of high usage in both categories. These findings emphasize the significance of vigilant energy management and the need for effective strategies to mitigate excessive consumption. The insights obtained from this study are obtainable from other energy source pages on the dashboard, making the dashboard a versatile tool for comprehensive energy analysis. Thus, the case study, therefore, aspires to serve as a guiding example, showcasing how data-driven decisions can be obtained towards steering sustainable energy practices.

However, it's important to highlight that if an in-depth analysis of a specific source, such as electricity consumption, is required, there is a need to explore additional data beyond the dashboard. This includes factors like the count of appliances, their usage timings, average usage duration, and the patterns of appliance utilization influenced by weather conditions or holidays, as mentioned in Section 2.2 and reference [41]. This limitation represents an area where the case study could be enhanced.

5.4 Challenges with PowerBI Dashboard

PowerBI exhibits limitations when handling high-frequency data. However, in our specific scenario, where data points occur less frequently, the dashboard provides a seamless user experience. Furthermore, PowerBI is not optimized for processing live streaming data, making it more suitable to work alongside complementary tools like Azure Stream Analytics for analyzing data streams from diverse sources or data streaming platforms such as Apache Spark Streaming or custom APIs. Additionally, while developing the dashboard, PowerBI offers certain built-in visualizations. Nevertheless, it may face constraints in adapting to the unique characteristics of the data or fulfilling specific requirements, such as displaying indicators or presenting specific subsets of data. These limitations often necessitate the inclusion of separate graphs and tables, which can add complexity to the dashboard. Our current calculations primarily focus on CO2 emission calculations. However, when testing the integration of future data, we encountered the need to update all tables with new values. To accommodate this, we introduced new tables and made amendments to the existing ones, maintaining the new tables as part of the dashboard architecture, which can contribute to its complexity. Moreover, if there is a requirement to update the dashboard from a different location other than the local computer, it necessitates modifying the data source in each file, as the dashboard hasn't been published online. However, for online publication and sharing of the dashboard, an upgrade to the current license to a pro version would be required, a consideration for potential future updates. In summary, the dashboard possesses its own set of advantages and disadvantages, contingent upon the specific needs and constraints of the end-user.

6 Timeseries Forecast Model using Machine Learning

The second stage of the project, denoted in this section, focuses on the prediction of energy consumption using historical data and machine learning techniques. The dataset primarily centres around historical records spanning the past decade. These records are sequential, relying on prior observations and influenced by recurring seasonal patterns like weather variations, holidays, pandemics, and social events. Additionally, the dataset is structured chronologically, making it exceptionally well-suited for conducting time series analysis. Despite covering ten years, the data is captured every month, resulting in a constrained dataset size. Following meticulous analysis and experimentation with various time series models, it was found that SARIMAX produced comparatively more favourable forecasting outcomes when contrasted with other models. A comprehensive discussion of these findings will be presented in the upcoming sections. The high parameter counts in LSTM models [9] can potentially lead to overfitting issues, rendering regularization a challenge, especially within the confines of a limited dataset. Subsequent sections delve into the details of feature engineering, model selection, hyperparameter tuning, and evaluation. A visual representation of the subsequent process stages has also been provided in Section 3.3.

6.1 Feature Engineering & Selection

The steps below outline the data pre-processing stages performed in Python.

1. As an initial step, **consolidated the dataset** created in Section 4.3 which comprised various energy types and building categories, into a Pandas data frame. In the initial step, we summed all the columns to obtain the total consumption values for electricity, gas, and oil. Then, in the subsequent step, we introduced three columns: total electricity, total gas, and total oil, ultimately yielding the 'Overall Energy Consumption' value as shown in Figure 6.1.
2. The data **rows starting from the year 2009** have been extracted, as all 3 sources of energy (CHP) are effectively available only starting from 2009.

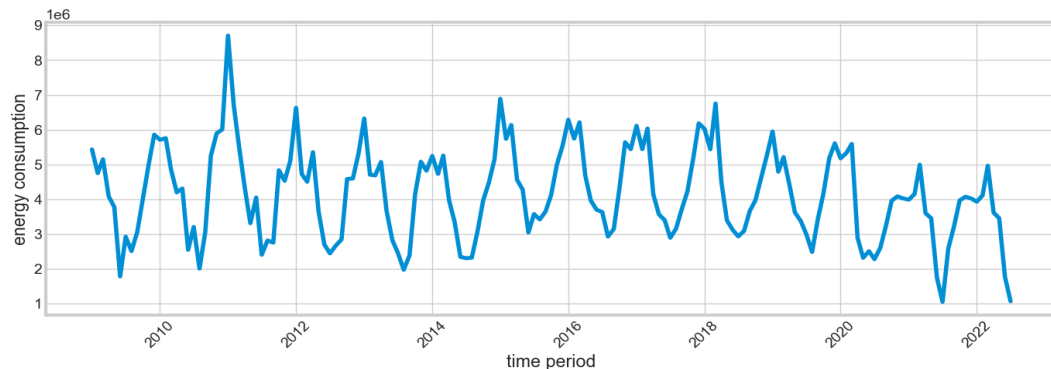


Figure 6.1 CHP Consumption Data

3. The date column has been **set as index** to facilitate calculation of time series metrics such as moving average and seasonality.

consumption	
date_time	
2009-01-01	5.434464e+06
2009-02-01	4.752303e+06

Figure 6.2 Date Indexing Sample

- Various attributes including **month**, **year**, **day of the week**, **quarter**, and **lag_1** were generated using the date index column. The creation of features was based on the assumption that they would effectively capture underlying trends and non-linear relationships, thereby enhancing model performance and interpretability.

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 163 entries, 2009-01-01 to 2022-07-01
Data columns (total 8 columns):
#   Column      Non-Null Count  Dtype
---  -
0   total ene    163 non-null    float64
1   total co2    163 non-null    float64
2   day_of_week  163 non-null    int64
3   month        163 non-null    int64
4   year         163 non-null    int64
5   quarter      163 non-null    int64
6   day_of_year  163 non-null    int64
7   lag_1        162 non-null    float64
dtypes: float64(3), int64(5)
memory usage: 11.5 KB
```

Figure 6.3 Target & Feature Columns in the data

- Few best correlated features - '**day_of_week**', '**year**', '**month**', '**quarter**' has been selected as feature variables based on correlation test and manual tuning.

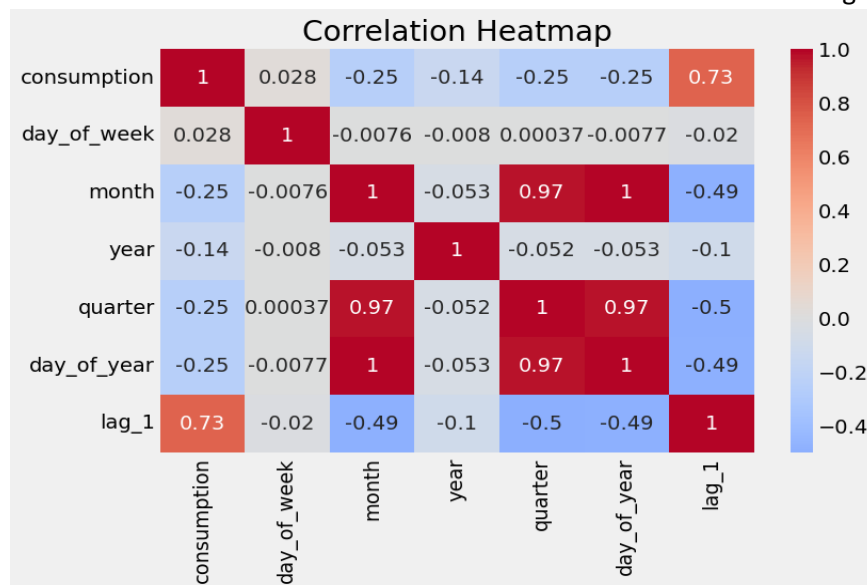


Figure 6.4 Correlation with Target Variable

- The consumption value column has been set as the target variable (y) and the created features as feature or exogenous variables (X) to further train the model. Timeseries split has been done on both target and feature variables using **TimeSeriesSplit()** cross-validation technique with 5 splits in the **ratio 80:20**, resulted in creating X_train, y_train, X_test and y_test datasets.

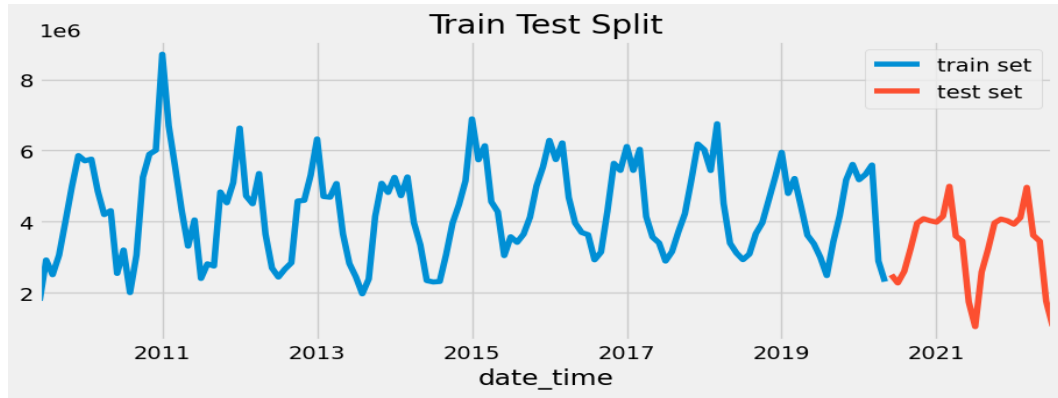


Figure 6.5 Train Test Split

7. Further, extreme anomalies have been detected in the training data by employing the **IsolationForest model** [55] with specific hyperparameters (**contamination=0.05, random_state=42**). Subsequently, these anomalies were excluded from the training dataset. We accomplished this by fitting the training data to the anomaly model to pinpoint the anomalies, followed by their removal through substitution with 'NaN' values. Finally, we addressed the 'NaN' values by filling them using the 'ffill' method [54], which replaces them with preceding values. The figure below illustrates the locations of these anomalies.

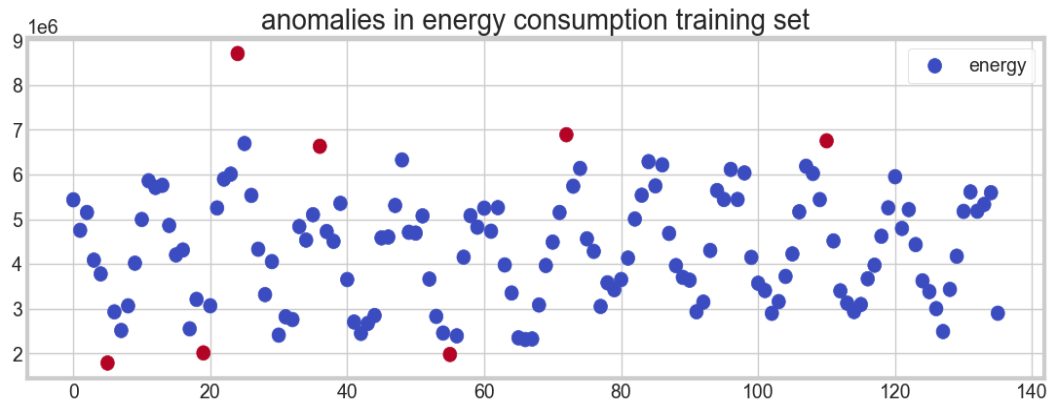


Figure 6.6 Anomalies (red labels) found in the training data

8. Computation of time series features - moving average, rolling standard deviation, identification of seasonality, residuals, trends, and autocorrelation within the data has been analysed on training data detailed in the next section.
9. Additionally, verification for stationarity has been conducted, which has been detailed in subsequent section.
10. Scaling and de-seasonalization of the data has not been performed as it tends to remove the seasonality and trend of the data resulting in poor performance of the model, explained in sections 6.2.2 and 6.2.3.
11. And data has not been made to stationary at this stage; instead, it was performed within the models. All the trained models are equipped to perform stationarity on the data when provided with the appropriate hyperparameters and through hyperparameter tuning.
12. The final training data has been prepared as in below figure.

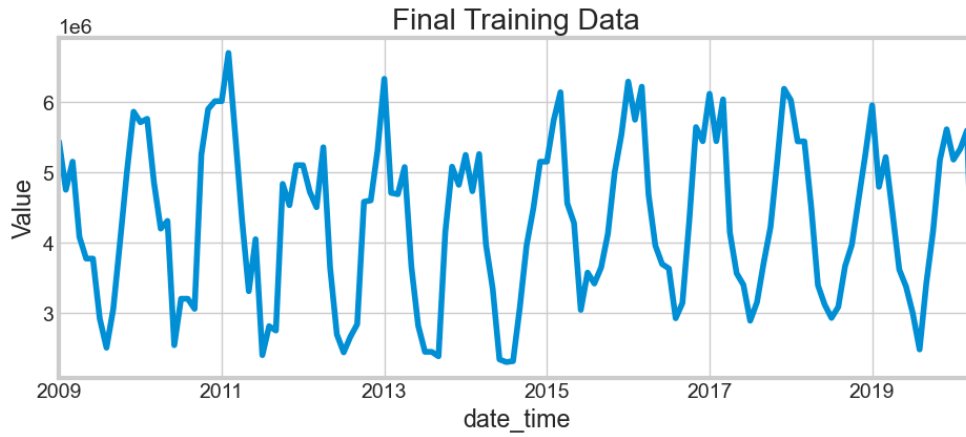


Figure 6.7 Final Pre-processed Training Data

6.2 Interpretation of Timeseries Features

6.2.1 Moving Average

A rolling 12-month average has been calculated from the training dataset, due to its capacity to approximate the underlying trend that the model is expected to generalize., where the calculation encompasses the mean value of the earliest 12 data points, succeeded by the subsequent 12 data points. This computation has been employed to ascertain the prevailing trend direction. By mitigating the influence of transient fluctuations, the rolling average provides a generalized trajectory of the data [10]. Furthermore, the moving average has been computed and visualized using the Python `'.rolling()'` function referred from [11], which has been provided in the below figure.

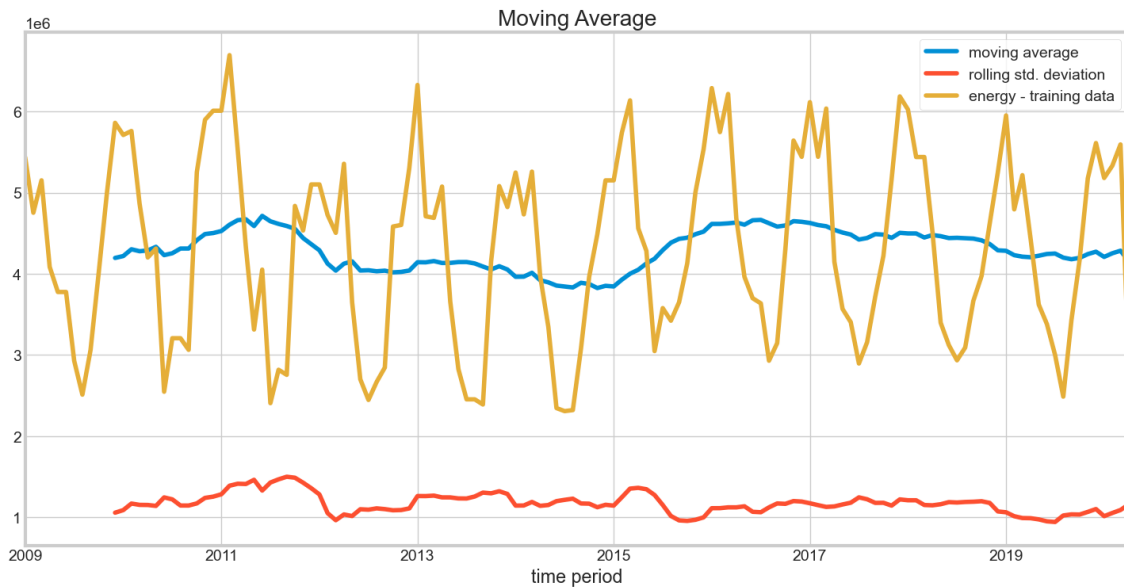


Figure 6.8 Moving Average and Standard Deviation of Train Data

By observing Figure 6.8 and given the span of electricity data starting in 2005, gas data from 2007, and oil data from 2009, it is noteworthy that the original data from these sources was effectively available only starting from 2009. This has led to a pronounced uptrend between 2006 and 2009. Subsequently, from 2009 to 2014, a slightly modest uptrend was observed, indicative

of a gradual increase in energy usage with minor oscillations likely attributed to seasonal influences. Notably, a downward trend is evident between 2018 and 2019, signifying a marginal decline in consumption.

6.2.2 Seasonal Decomposition

The seasonal decomposition technique facilitates the straightforward identification of both the linear and seasonal patterns within our dataset, along with the residual components. The illustration of these patterns is available in Figure 6.9, and the decomposition has been checked on the training dataset. The seasonal decomposition has been identified using the `seasonal_decompose()` function referred from [12] and a multiplicative model has been used to decompose as the dataset has fluctuations based on period. The trend is similar to the moving average mentioned in Section 6.2.1; however, it provides the overall long-term underlying direction of the data. The trend shows that the consumption has been going downtrend from the year 2011 to 2014 and its increasing and uptrend from 2015 to 2018 as well as from 2019 there is a slight decrease in the energy consumption. On the other hand, seasonality provides repeating patterns in the data, where we can see that at the start and end of the year, energy consumption is at its peak and the mid-year has the lowest consumption as well as emission. In addition, the irregularities have been shown using the residual plot, where some fluctuations are showing up in the years 2011 to 2013.

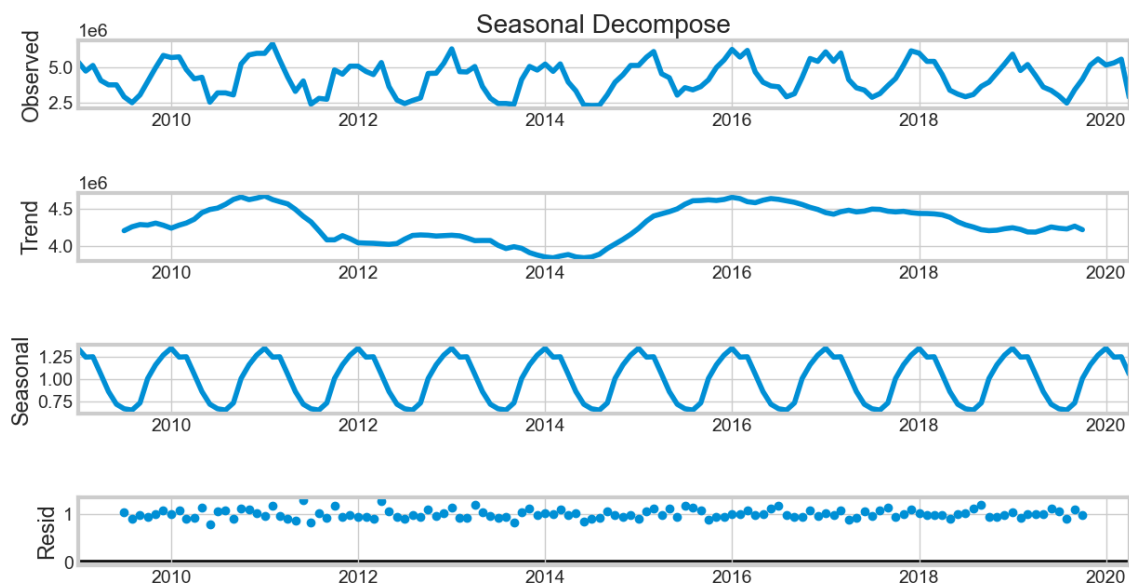


Figure 6.9 Seasonal Decomposition of the train data

This procedure of de-seasonalization aids in diminishing noises and fluctuations within the data. Nevertheless, it's noteworthy that SARIMAX models are inherently equipped to accommodate seasonal fluctuations, rendering the necessity of de-seasonalization optional. As a result, both scenarios—with and without de-seasonalization—were assessed using the model and decided not to proceed with de-seasonalization.

6.2.3 Stationarity of the Data

The assessment of data stationarity was conducted through the Augmented Dickey-Fuller (ADF) test. This test states the null hypothesis that the coefficient α (associated with the first lag on y) is equal to 1. The resulting p-value, calculated at 0.4, exceeds the significance level of 0.05. This outcome indicates the non-rejection of the null hypothesis, thereby confirming the data is non-stationary. Despite the data's inherent non-stationarity, the ARIMA and SARIMAX models are

supported by the integrated component (I), which facilitates the conversion of the data into a stationary state [13], a process elaborated upon in Section 6.3.

Test Statistic	-1.949773
p-value	0.309020
#lags used	12.000000
number of observations used	123.000000
critical value (1%)	-3.484667
critical value (5%)	-2.885340
critical value (10%)	-2.579463
dtype: float64	
Data is not stationary	

Figure 6.10 Results of the ADF Stationarity Test

6.2.4 Auto Correlation

The Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) were employed to investigate the lag between data points by comparing each current data point with the previous one. The ACF and PACF of the training data are depicted in the below figures.

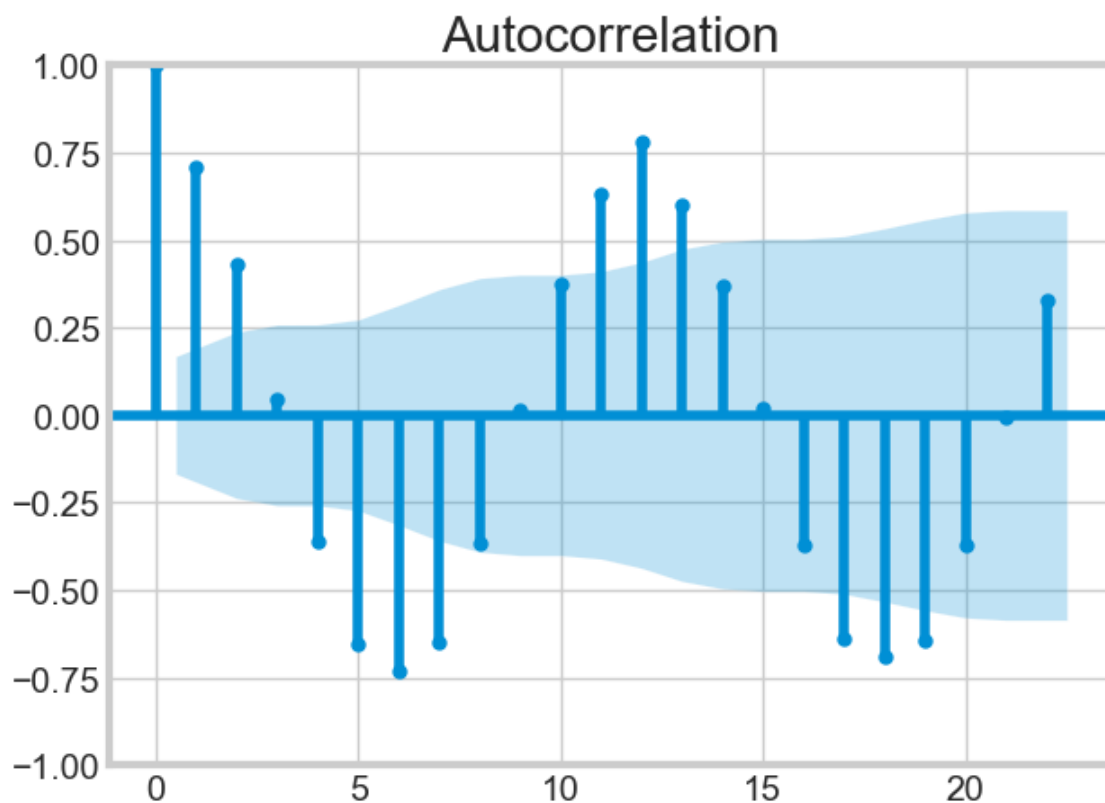


Figure 6.11 Auto-Correlation Plot

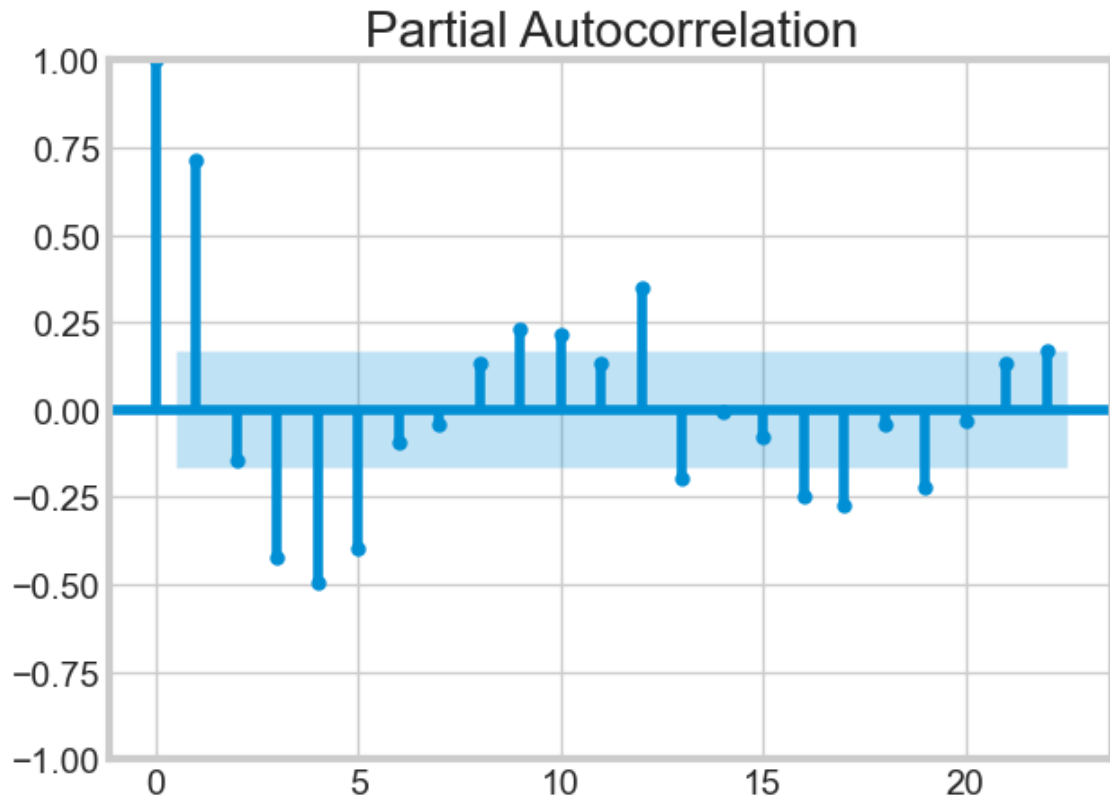


Figure 6.12 Partial Auto-Correlation Plot

The observed pattern reveals a consistent alternation of higher and lower trends, indicative of the presence of a recurring seasonal pattern. Additionally, the lag points that extend beyond the confidence interval indicate the existence of underlying correlations between the data points. Conversely, the lag points falling within the confidence interval exhibit correlations that are not statistically significant, likely attributed to noise within the data.

6.3 Timeseries Model Training

The trained models for predicting CHP consumption (referred to [14]) are,

- Auto-ARIMA (referred [44])
- SARIMAX (referred [16])
- Prophet (referred [43])
- Exponential Smoothing (referred [15])

Among these, Prophet, SARIMAX, and Exponential Smoothing have yielded comparatively superior outcomes. Moreover, LSTM and Gradient Boosting models have been tested and not proceeded with, as both exhibited less favourable results, possibly due to the complexity of the model & overfitting, considering the dataset's presumed less complex.

6.3.1 Model Assumptions & Hyperparameters

The models chosen for predictions are SARIMAX, Auto-ARIMA, Prophet, and Exponential Smoothing. The selection of the SARIMAX model was underpinned by the prominent presence of robust seasonal patterns within the data (Section 6.2.4), coupled with the model's adaptable

customization. Auto-ARIMA, Prophet, and Exponential Smoothing are also chosen due to their capacity to automatically determine hyperparameters, leveraging insights from seasonal trends and variations. Moreover, Prophet exhibits the capability to handle multiple variables, including seasonal influences, holidays, and patterns. The specific hyperparameters of all the models have been detailed in the subsections.

6.3.1.1 SARIMAX

The hyperparameters of SARIMAX (Seasonal Autoregressive Integrated Moving Average Exogenous) that have been tuned are provided in Table 6.1. Within these parameters, the specific set of (p, d, q) and (P, D, Q, S) parameters were identified as the optimal configuration for the final model. This decision was guided by the fact that these parameter specifications yielded the lowest error rates.

Parameters: The order=(p, d, q) represents the ARIMA order, where

p = auto-regressive order represents the lagged value of dependant variable

q = differencing order represents the number of times differencing done for making data stationary

r = moving-average order represents the lag of error in the forecast

Similarly, P , D and Q is same as above ARIMA order but addresses the seasonal components, with S as the seasonal period (S=12 means a 12-month cycle)

Assumptions: Furthermore, when utilizing the model, it is assumed that stationarity will be attained by applying differencing (d), the relationship between the data and previous observations is considered linear, the linearity of exogenous variables is also assumed, there is low correlation among the exogenous variables, the coefficients remain constant, and the residuals follow a normal distribution.

Hyperparameters:

Hyperparameters	SMAPE	MAPE	MAE	RMSE
order=(0, 0, 1), seasonal_order=(0, 1, 0, 12)	29.83	38.87	969214.95	1104421.85
order=(0, 1, 0), seasonal_order=(0, 1, 0, 12)	72.97	41.46	1410327.53	1714807.43
order=(0, 1, 1), seasonal_order=(0, 1, 1, 12)	21.63	19.72	592910.00	753745.22
order=(3, 1, 0), seasonal_order=(3, 1, 0, 12)	23.96	21.77	638526.52	797168.83
order=(1, 0, 2), seasonal_order=(1, 0, 2, 12)	21.96	30.91	697816.29	917675.24
order=(0, 0, 3), seasonal_order=(0, 0, 3, 12)	200.00	5650243186.93	187104942881770.19	324264007752388.56
order=(0, 0, 2), seasonal_order=(2, 0, 2, 12)	25.41	36.17	833928.54	1034770.47
order=(1, 1, 2), seasonal_order=(1, 1, 2, 12)	21.27	22.71	602987.91	690947.54

Table 6.1 Hyperparameters used in SARIMAX

The hyperparameter values selected are **order=(0, 1, 1), seasonal_order=(0, 1, 1, 12)**

Summary Statistics & Residual Plot:

The Ljung-Box Q-test results stand at 0.13, implying a slight degree of autocorrelation within the residuals. Further, the standardized residual plot indicates a central alignment of residuals around zero, implying that the model is capturing the data well in both models. The histogram demonstrates a distribution approximating normality, affirming that the residuals are nearly normally distributed. The quantile-quantile (Q-Q) plot further underscores this by illustrating residuals that are clustered around the diagonal line confirming residuals are normally distributed. Additionally, the correlogram explains that the residuals are in proximity to zero, signifying there is no significant correlation among them.

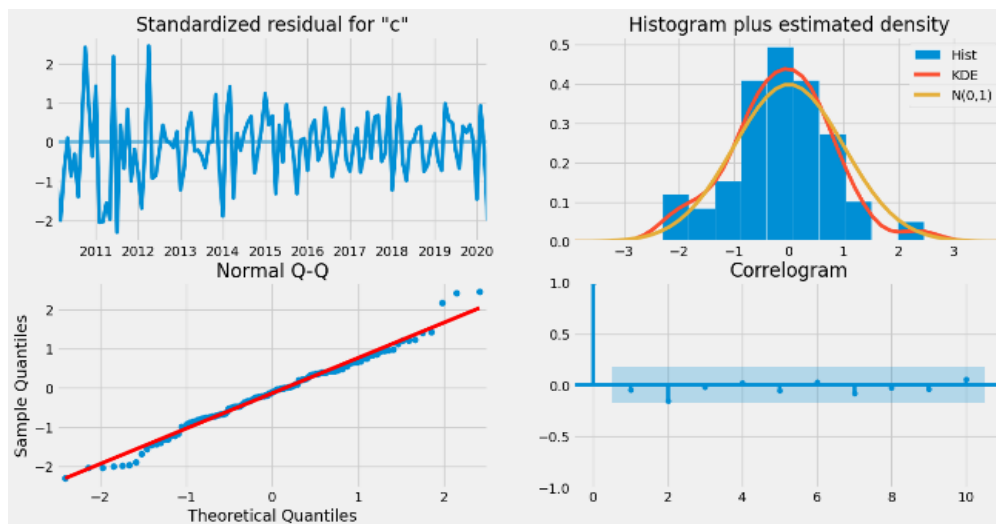


Figure 6.13 SARIMAX Residual Plot

```
1 sarimax_results2.summary()
```

SARIMAX Results

Dep. Variable:	consumption	No. Observations:	136
Model:	SARIMAX(0, 1, 1)x(0, 1, 1, 12)	Log Likelihood	-1821.982
Date:	Sun, 27 Aug 2023	AIC	3657.965
Time:	23:22:53	BIC	3677.650
Sample:	01-01-2009	HQIC	3665.961
	- 04-01-2020		

Covariance Type:	opg					
	coef	std err	z	P> z	[0.025	0.975]
day_of_week	-1.818e+04	2.92e+04	-0.623	0.533	-7.54e+04	3.9e+04
year	0.0004	2235.271	1.83e-07	1.000	-4381.049	4381.050
month	2.497e-06	4.59e+05	5.44e-12	1.000	-9e+05	9e+05
quarter	1.01e-06	9.4e+05	1.07e-12	1.000	-1.84e+06	1.84e+06
ma.L1	-0.4049	0.091	-4.472	0.000	-0.582	-0.227
ma.S.L12	-0.5865	0.100	-5.865	0.000	-0.782	-0.391
sigma2	5.013e+11	3.039	1.65e+11	0.000	5.01e+11	5.01e+11

Ljung-Box (L1) (Q):	0.26	Jarque-Bera (JB):	0.82
Prob(Q):	0.61	Prob(JB):	0.66
Heteroskedasticity (H):	0.35	Skew:	0.04
Prob(H) (two-sided):	0.00	Kurtosis:	3.39

Figure 6.14 SARIMAX Summary Statistics

6.3.1.2 Auto-ARIMA

The Auto-ARIMA (Automatic Auto-Regressive Integrated Moving Average) model has been trained with the features provided below,

(exog=X_train, seasonal=True, stepwise=True, suppress_warnings=True)

Assumptions: The Auto-ARIMA model can accommodate non-stationary data and automatically determines the differencing order required to render the data stationary. It also presumes a linear association between the present values of the time series and the autoregressive and moving-average terms. Additionally, it operates under the assumption that the data does not contain outliers or anomalies. The 'exog' parameter corresponds to the training set of feature variables, while 'seasonal' is enabled to account for seasonal patterns. When 'stepwise' is set to 'True,' the model performs an automated stepwise search to identify the optimal (p, d, q) orders, automatically selecting the best ARIMA configuration. These (p, d, q) values serve the same role as described in previous section. Additionally, 'suppress_warnings' is configured as 'True' to prevent warning messages during the fitting process, resulting in a more streamlined and tidy output.

Summary Statistics & Residual Plot: Notably, the Auto-ARIMA model has automatically determined the optimal time series order as (5, 0, 0). The Ljung-Box Q-test results stand at 0.11, respectively, implying a slight degree of autocorrelation within the residuals. In the Auto-ARIMA model, the intercept coefficient and auto-regressive order lags 1 to 4 exhibit a positive influence on the model, while lag 5 exerts a negative impact.

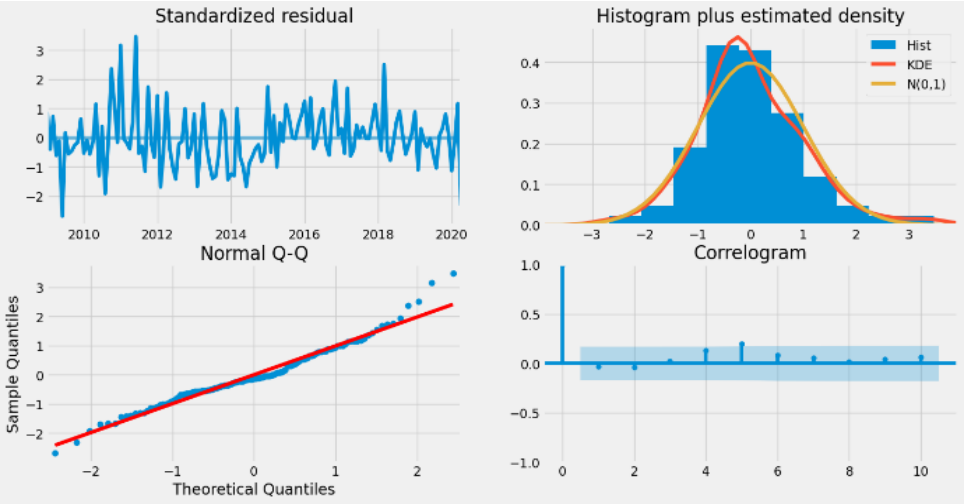


Figure 6.15 Auto-ARIMA Residual Plot

1 autoarima_results3.summary()

SARIMAX Results						
Dep. Variable:		y	No. Observations:		132	
Model:		SARIMAX(5, 0, 0)	Log Likelihood:		-1958.689	
Date:		Sun, 27 Aug 2023	AIC		3931.378	
Time:		19:36:07	BIC		3951.557	
Sample:		06-01-2009	HQIC		3939.578	
		- 05-01-2020				
Covariance Type:		opg				
	coef	std err	z	P> z	[0.025	0.975]
intercept	4.074e+06	1.62e-08	2.52e+14	0.000	4.07e+06	4.07e+06
ar.L1	0.4227	0.066	6.446	0.000	0.294	0.551
ar.L2	0.2038	0.105	1.944	0.052	-0.002	0.409
ar.L3	0.0774	0.092	0.845	0.398	-0.102	0.257
ar.L4	-0.2841	0.095	-2.979	0.003	-0.471	-0.097
ar.L5	-0.3636	0.064	-5.703	0.000	-0.489	-0.239
sigma2	4.373e+11	1.14e-13	3.85e+24	0.000	4.37e+11	4.37e+11
Ljung-Box (L1) (Q):		0.11	Jarque-Bera (JB):		13.54	
Prob(Q):		0.74	Prob(JB):		0.00	
Heteroskedasticity (H):		0.55	Skew:		0.62	
Prob(H) (two-sided):		0.05	Kurtosis:		3.97	

Figure 6.16 Auto-ARIMA Summary Statistics

6.3.1.3 Exponential Smoothing

The hyperparameters selected to train Exponential Smoothing (ETS) is,

(seasonal='mul', seasonal_periods=12, initialzition_method='heuristic')

Assumptions: The 'seasonal' parameter, configured as 'mul,' signifies multiplicative seasonality, which means the seasonal impact of a given month is expressed as a percentage of the overall series. Specifying a 'seasonal period' of 12 indicates a yearly cycle, and the choice of the

'heuristic' initialization method means that the (p, d, q) parameters are determined using rule-of-thumb heuristics.

Summary Statistics: The absence of a trend component in the model is indicated as 'None,' signifying that the model did not identify any discernible trend within the dataset. The AIC and BIC values serve as metrics for assessing the goodness of fit, with lower values indicating better fit. Among all the models, SARIMAX and Prophet produced the lowest AIC and BIC values, implying better fit. It's worth noting that no Box-Cox transformation was applied. Normally, such a transformation is employed to stabilize variance and normalize the data.

```
1 exp_smooth_results5.summary()
```

ExponentialSmoothing Model Results

Dep. Variable:	consumption	No. Observations:	132
Model:	ExponentialSmoothing	SSE	43493091048285.406
Optimized:	True	AIC	3528.748
Trend:	None	BIC	3569.108
Seasonal:	Multiplicative	AICC	3533.479
Seasonal Periods:	12	Date:	Sun, 27 Aug 2023
Box-Cox:	False	Time:	19:36:08
Box-Cox Coeff.:	None		

	coeff	code	optimized
smoothing_level	0.1935714	alpha	True
smoothing_seasonal	0.2758835	gamma	True
initial_level	4.0439e+06	l.0	True
initial_seasons.0	0.6996458	s.0	True
initial_seasons.1	0.6110471	s.1	True
initial_seasons.2	0.5551426	s.2	True
initial_seasons.3	0.6459089	s.3	True
initial_seasons.4	1.0994229	s.4	True
initial_seasons.5	1.1797074	s.5	True
initial_seasons.6	1.2830640	s.6	True
initial_seasons.7	1.5652350	s.7	True
initial_seasons.8	1.2551654	s.8	True
initial_seasons.9	1.1312634	s.9	True
initial_seasons.10	1.1063919	s.10	True
initial_seasons.11	0.8680056	s.11	True

Figure 6.17 Exponential Smoothing Summary Statistics

6.3.1.4 Prophet

Hyperparameters: The hyperparameters tested on the model are,

Hyperparameters	SMAPE	MAPE
seasonality_mode='multiplicative', changepoint_prior_scale=0.5, holidays_prior_scale=0.1, n_changepoints=200	19.77	25.13
seasonality_mode='additive', changepoint_prior_scale=0.1, holidays_prior_scale=0.1, n_changepoints=100	24.41	32.70
seasonality_mode='multiplicative', changepoint_prior_scale=0.1, holidays_prior_scale=0.1, n_changepoints=100	22.89	30.69
seasonality_mode='multiplicative', changepoint_prior_scale=0.3, holidays_prior_scale=0.3, n_changepoints=200	20.14%	26.01
seasonality_mode='multiplicative', changepoint_prior_scale=0.01, holidays_prior_scale=0.01, n_changepoints=100	24.67	33.54
seasonality_mode='multiplicative', changepoint_prior_scale=0.5, holidays_prior_scale=0.3, n_changepoints=100	19.66	24.93

Table 6.2 Hyperparameters used in Prophet

The parameters [**change-point_prior_scale=0.5**, **holidays_prior_scale=0.3**, **n_change-points=100**] has been selected as it exhibits the least error rates compared to others.

Assumptions: The model can handle time series data that is not stationary, and it independently addresses the trend, seasonality, and holiday elements, automatically capturing these components. This adaptability makes the model well-suited for forecasting over short to medium time horizons. The '**seasonality_mode**' parameter specifies the nature of seasonality within the model. When set to 'additive,' the seasonal component is added to the trend components, whereas in the 'multiplicative' mode, the seasonal component is multiplied by the trend components. Both modes were tested in the model, with the 'multiplicative' mode yielding better performance. The '**changepoint_prior_scale**' parameter influences the model's responsiveness to detecting significant changes in the data. A value of 0.5 reduces sensitivity to change points, providing a more stable model.

Regarding '**holidays_prior_scale**,' it determines the influence of holidays on the model. A value of 0.3 suggests a moderate impact of holidays on forecasts. Lastly, '**n_changepoints**' specifies the maximum number of change points considered by the model, set at 100 in this case, to capture significant shifts in the data."

Cross-Validation: The results of performance metrics following the implementation of cross-validation using the specific parameters are presented below.

```
from prophet.diagnostics import cross_validation
df_cv = cross_validation(model, initial='500 days', period='180 days', horizon='50 days')
# print(df_cv.head())
```

	horizon	mse	rmse	mae	mape	mdape	smape	coverage
0	5 days	3.752612e+11	6.125857e+05	4.265875e+05	0.097465	0.035513	0.088577	0.333333
1	6 days	4.866911e+11	6.976325e+05	5.669221e+05	0.119177	0.100649	0.112274	0.333333
2	8 days	5.145553e+11	7.173251e+05	6.490980e+05	0.131030	0.100649	0.124324	0.166667
3	10 days	4.772636e+11	6.908426e+05	6.863900e+05	0.147212	0.164000	0.150575	0.000000
4	11 days	1.811956e+12	1.346089e+06	1.247538e+06	0.235606	0.176987	0.272314	0.000000

Figure 6.18 Cross-Validation Results from Prophet Model

While analysing the cross-validation results, it is observed that the SMAPE error rose significantly, going from 8% for a 5-day forecast to 27% for an 11-day forecast. Similarly, the MAPE also increased, climbing from 0.09 to 0.23, as the prediction horizon extended. Additionally, RMSE, MAE, and MSE scores showed upward trends with longer horizons. Furthermore, the model's coverage reached 0.00 at the maximum prediction horizon, indicating that it encounters difficulties when making longer-term predictions. In essence, these findings suggest that the model is not well-suited for forecasting over extended time periods.

Component Plot: The components plot of the prophet is given below. It is noteworthy, as depicted in Figure 6.19, that the forecast shows a downward trend in the forthcoming years, with January and December displaying the most pronounced consumption aligning well with the historical data even though the model shows as no trend has been captured.

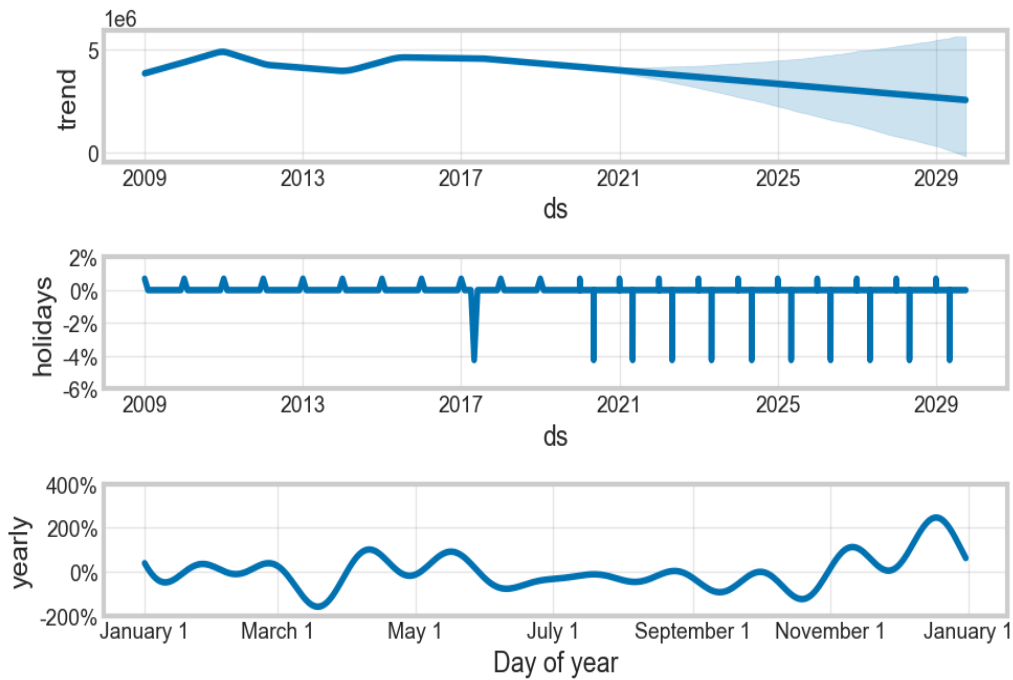


Figure 6.19 Component Plot of Prophet Model

6.3.2 Comparison of Predictions & Evaluation Metrics Across All Models

Figure 6.20 and Figure 6.21 provides the predictions generated by each of the models on the test data. The Auto-ARIMA predictions exhibit the highest error rate and notable disparities in prediction accuracy among all the models, followed by Prophet even though Prophet has managed to give the trend in the data. In contrast, the SARIMAX and Exponential models yield closely comparable outcomes. However, it's important to highlight that the SARIMAX model achieves the lowest MAPE and shows a moderate SMAPE error score, indicating strong overall performance. This suggests that the model's forecasts have a reasonably balanced percentage error, and both MAE and RMSE scores are also at a moderate level. On the contrary, Auto-ARIMA exhibits the highest MAPE score among all the models, signifying the highest forecast error rate. Furthermore, Prophet delivers promising results with the lowest SMAPE and the second lowest MAPE score, suggesting its effectiveness in providing accurate forecasts. However, it's worth noting that, as observed in the cross-validation results in section 6.3.1.4, Prophet may not excel in making long-term predictions compared to other models, despite its low error rates. This factor should be considered when evaluating its suitability for specific forecasting needs.

In summary, SARIMAX stands out with favourable error rates and the ability to fine-tune hyperparameters to meet specific requirements, making it a strong performer in this context.

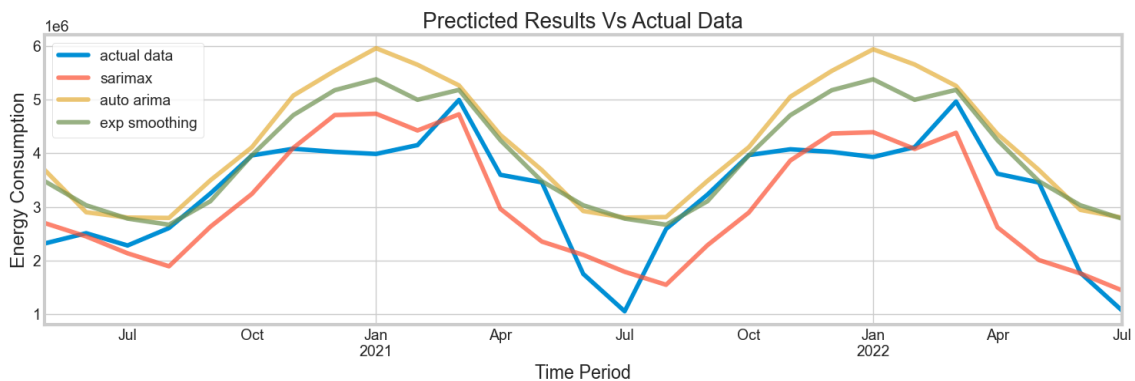


Figure 6.20 Predictions Comparison on Test Data

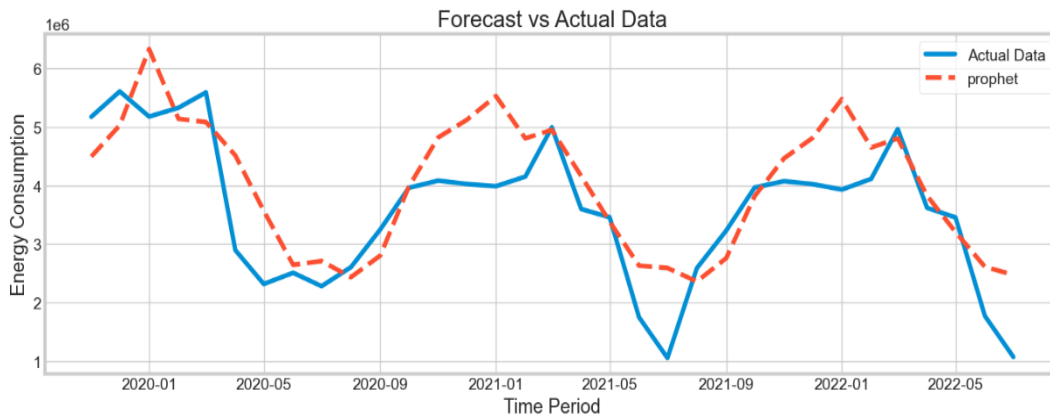


Figure 6.21 Prediction Comparison of Prophet on Test Data

MODEL	MAPE	SMAPE	MAE	RMSE
Sarimax	0.19	19.96	555085.86	670298.49
Auto-Arima	0.36	26.24	883735.66	1085635.71
Exp Smoothing	0.30	22.11	683166.61	880906.59
Prophet	0.25	19.79	644873.69	809373.66

Figure 6.22 Comparison of Evaluation Metrics Results

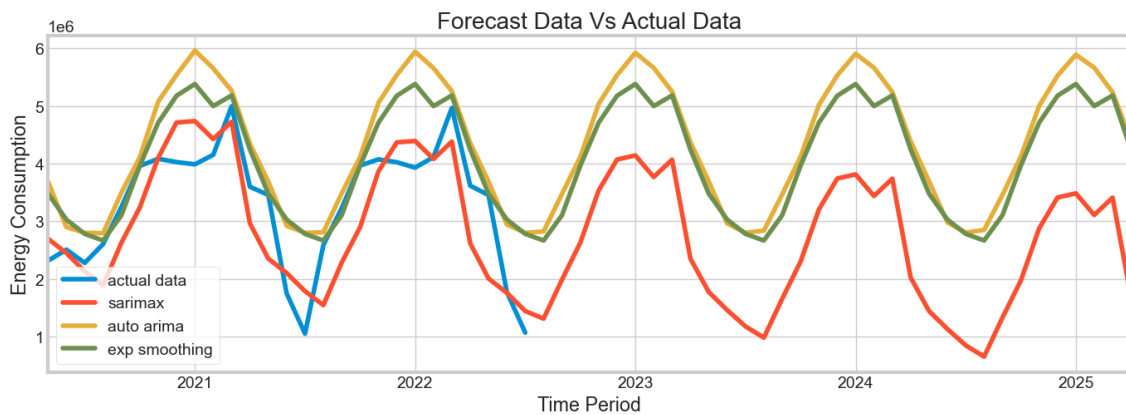


Figure 6.23 Comparison of Forecasted Results with Test Data

6.4 Final Timeseries Model

The final model selected for forecasting the consumption data is the SARIMAX model with specific hyperparameters [`order=(0, 1, 1)`, `seasonal_order=(0, 1, 1, 12)`, `enforce_stationarity=True`, `enforce_invertibility=True`].

6.4.1 Final Model Interpretation

As outlined in section **Error! Reference source not found.**, both Prophet and SARIMAX yielded relatively better results, with Prophet showing strong performance. However, it's worth noting that Prophet has been found to be less capable when it comes to making long-term forecasts. Hence the SARIMAX model has been preferred as it combines Seasonal ARIMA with Exogenous

features and this framework allows for hyperparameter tuning of time series orders, seasonal orders, and stationarity enforcement. Additionally, it offers insights into coefficient estimates, enhancing the model's customizability and interpretability, ultimately leading to improved accuracy. The reference [17] has been consulted for a comprehensive analysis of time series properties and methodologies. Notably, the autoregressive (p) and seasonal autoregressive (P) orders are both set to 0, indicating that the current values are not influenced by past values. Integration orders (d and D) are set to 1, ensuring stationarity is achieved. The moving average orders (q and Q) are both set to 1, capturing recent residuals to account for dependencies. Finally, the seasonal pattern (s) is defined as 12 to capture a 12-month seasonal cycle. For a deeper understanding of time series and seasonal orders, [18] has been referred to. Furthermore, the error metrics are shown in Figure 6.22, wherein, MAPE indicates a 19% disparity between the projected and observed values, signifying an accuracy of 81%. Meanwhile, SMAPE records an error of 19.96%, encompassing both overestimation (predictions exceeding actual values) and underestimation (predictions falling short of actual values). [19] has been referred for the same. The forecast results obtained from the model are presented below.

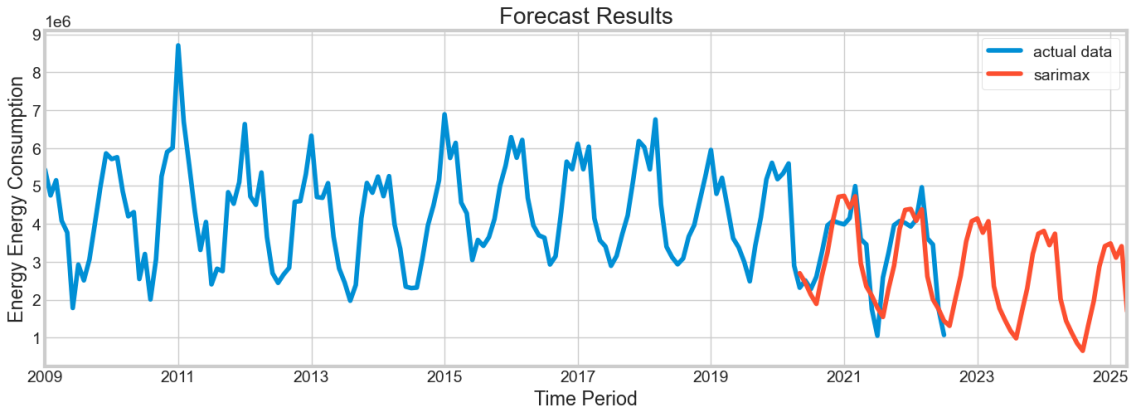


Figure 6.24 Forecast Result on Final Model

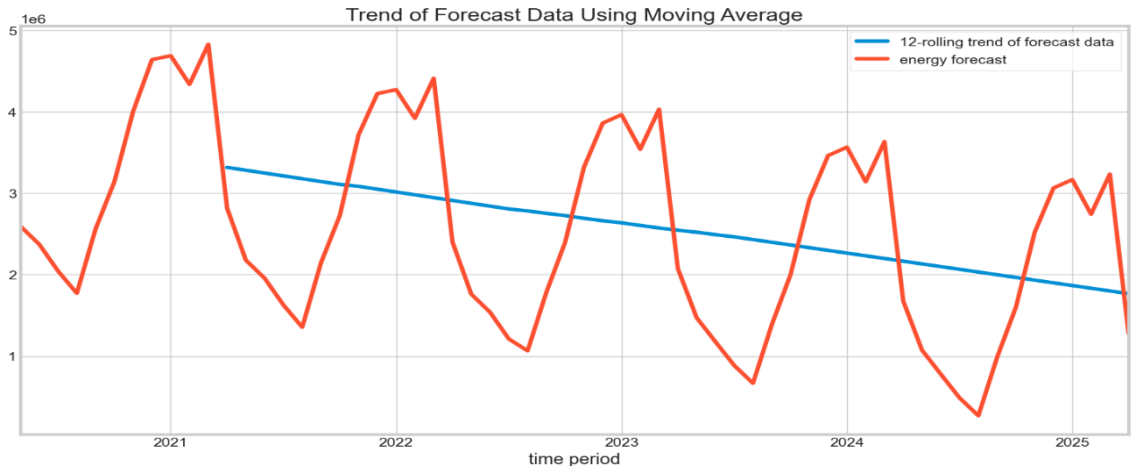


Figure 6.25 Trend of Forecasted Result

6.4.2 Final Model Results

The forecasts were generated for a 5-year period spanning from 2020 to 2025 (Figure 6.24), using the `.predicted_mean` and `.get_forecast()` functions. The outcomes reveal a consistent pattern of seasonality, characterized by peaks in November and December, followed by declines in February and March. August and July emerge as the months with the lowest forecasted values. In general, the data exhibits a declining trend (Figure 6.25), with peak consumption dropping from 4.641138×10^6 in December 2020 to 3.064570×10^6 in December 2024. Additionally, the model

has effectively recognized the underlying patterns of the unseen test data, with minimal discrepancies. Furthermore, SARIMAX model presents as a computationally efficient choice, particularly because the dataset exhibits a pronounced seasonal trend, the dataset size is moderate, and the forecasting frequency is monthly, which is considered moderate.

6.5 Challenges in Model Training

The challenges faced throughout the process stem from the constrained size of the dataset, comprising a total of only 205 data points. It has restricted both the effectiveness of analysis and the capacity to apprehend patterns, especially when it comes to applying a time series model. The model's ability to identify trends is hindered by the limited data available, a limitation that is less significant in datasets of larger scale. Moreover, the dataset's size hampers the appropriateness of sophisticated models like neural networks, reducing their feasibility. Notably, the presence of outliers has exhibited minimal influence on the model's ability to discern underlying trends; its performance remains consistent whether outliers are present or removed. Despite these challenges, the model has achieved an accuracy of 81%, competently capturing the trend in a commendable manner.

7 Conclusion

This section outlines the summary of the project, critical evaluation, and recommendations for future work on the project.

7.1 Summary

The purpose of the project was to conduct an in-depth analysis of energy usage and the associated carbon emissions at the University of Stirling. The main aim was to construct a dashboard that visually represents energy consumption. This undertaking was divided into two distinct phases: the first involved designing the dashboard, while the second focused on predicting energy consumption for the upcoming five years. At the outset, the data was spread across multiple files, each containing a year's worth of data for a specific energy source. Notably, these files exhibited distinct tabular formats and dissimilar naming conventions and this complexity of data structure necessitated additional efforts for data processing. Furthermore, a notable data gap was present spanning from 2014 to 2015. To address this temporal inconsistency within the time series data, random values were inserted. It is worth noting that while electricity data extended back to 2005, gas data was available from 2007, and oil data was accessible from 2009. Furthermore, travel data was limited to the most recent three years. This incongruity in data availability led to the selective use of specific components, such as CHP data from 2009, for forecasting purposes. Additionally, a study of carbon management systems within educational institutions has highlighted challenges arising from data reliability issues & associated costs, in carbon management and analysis (Section 2.1.1); therefore, maintaining data storage & integrity emerges as a pivotal factor for conducting comprehensive future analyses. And in a subsequent step, CO₂ emissions were derived through the application of GHG conversion factors. Consequently, the data has been consolidated from these sources to form a dataset that could be employed for this purpose.

In the initial stage, the dataset, previously derived from the raw data as described in the preceding paragraph, underwent a series of transformations and preprocessing steps in PowerBI to attain a format suitable for integration into the dashboard. The dashboard itself was composed of seven distinct pages: an overarching overview page featuring CHP consumption, inclusive of emission-critical components, five dedicated pages providing consumption breakdowns for each energy source - Electricity, Gas, Oil, Water, and Fleet, and a final page encompassing a comparative analysis of Scottish universities. This dashboard facilitated an in-depth analysis of energy consumption and CO₂ emissions. It offers granular insights through monthly breakdowns and annual analyses, all organized by categories such as year, building, and consumption type. Moreover, the dashboard incorporated features such as anomaly detection and CO₂ forecasting. Notably, the dashboard's scope expanded to encompass Waste and Travel data consumption and a comprehensive documentation was generated, outlining guidelines for future data modifications and updates (Appendix 2) to accommodate end-user requirements.

The second phase involves the implementation of predictive modelling for energy usage using time series data, facilitated through the utilization of the Python programming language. The modelling process specifically incorporates CHP data spanning from 2009 to 2022. To achieve this, a cross-validation technique with an 80:20 ratio time-series split is adopted, enabling a rolling basis cross-validation approach. The initial steps encompass data cleaning and preprocessing, crucial for data integrity. The chosen forecasting model is SARIMAX, selected after careful consideration. Performance assessment relies on evaluation metrics comprising MAPE, SMAPE, RMSE, and MAE. Notably, the model attains an accuracy rate of 81%. Manual hyperparameter tuning is executed to optimize the model's performance and the selection of optimal hyperparameters is guided by minimizing error rates. Subsequently, a five-year forecast is conducted, revealing a gradual and slight downward trend in the anticipated energy pattern over the

forthcoming years. Throughout this phase, acknowledged challenges include managing inconsistencies within the data, addressing missing values, and contending with the constraints imposed by a relatively limited dataset size. These limitations are recognized as factors that influenced the training and predictive capabilities of the model in the context of forecasting energy usage patterns.

The key findings derived from the analysis reveal that energy consumption has exhibited a gradual reduction throughout the years, displaying marginal discrepancies. Notably, electricity usage has experienced a noteworthy decline, whereas gas consumption has concurrently shown an increase. Surprisingly, it has been revealed that Natural Gas is the primary source of carbon emissions, accounting for an overall 105 million KgCO₂e, while electricity follows closely with 92 million KgCO₂e. Nevertheless, a previous study referred from [32] identified electricity as the main contributor to the emissions followed by transportation among UK Russell Group Universities discussed in Section 2.1.1; however, it is noteworthy that the approach of analysis is different in both cases. Conversely, oil, water, and waste data exhibit a consistent pattern with minor fluctuations. Hence, the purpose of both the dashboard and the forecasting is to equip end users with the means to enhance outcomes based on historical data. Section 5.3 presents a comprehensive case study example, showcasing monthly and yearly data analysis. This illustration serves as a practical demonstration of how end users might interact with and leverage the dashboard, thereby aiding them in making informed decisions towards a more environmentally sustainable future. As a result, the goals of this project, which involves creating an interactive tool for end users and forecasting, have been accomplished. This tool stands poised to assist the University in drawing future conclusions informed by historical data, supporting sustainable energy management practices.

Overall, the project contributes to the existing knowledge already available in the field of energy management and sustainability by tackling the challenges surrounding data integration, visualisation all in one location, and predictive modelling. It provides insightful information that might act as a compass point for future research projects.

7.2 Critical Evaluation

The broader objectives of the project have been divided into three distinct parts, and a critical evaluation of these components is detailed in the subsequent sections.

7.2.1 Data Integration

Addressing data integration faced challenges related to incomplete and inconsistent datasets, encompassing concerns about data quality and the complexity of collating data from multiple sources. Even though only the consumption data needs to be extracted from it, achieving data integrity presents its own set of complexities. To overcome these challenges, the strategy taken involved creating a combined dataset using Python libraries in conjunction with the Excel application. However, this was met with several hurdles, prominently stemming from the difficulties detailed in paragraph 2 of section 7.1. Additionally, the computation of CO₂ emissions was automated using the Power Query, ensuring the consistency of emission data. Furthermore, a documentation was created, providing instructions for future data management.

7.2.2 PowerBI Dashboard

Due to the challenges faced in the data integration while designing the dashboard, an emphasis was also placed on achieving user-friendliness and interactivity suited for a non-technical user base along with technical aspects. PowerBI seemed more of a practical solution than other applications such as Tableau or Flask due to the availability of licences and the ease of use and

future data management. Aside from accessibility, careful consideration was given to the dashboard design that was appropriately suited to user requirements, resulting in a design that achieves a balance between easy navigation and in-depth data exploration. The use of navigational panes, slicers, and toggles provides users with the flexibility needed for seamless exploration.

7.2.3 Timeseries Forecast Model

The performance of various models relevant to the specific dataset was assessed while developing the predictive model. The SARIMAX model was chosen since the dataset has a relatively lower level of complexity and the model's hyperparameters can be adjusted to meet its distinct requirements. Section 6.4 describes in detail on the challenges the model faced along the way, its performance metrics, and the comparison with other models. Additionally, the performance of the model may be affected by the impact of events like natural disasters, changes in regional or international energy policy, the adoption of energy-efficient practices or appliances, or even abnormal weather patterns or interruptions in supply chain operations. These elements may appear as anomalies that do not fit into the known patterns. Despite these complexity levels and dataset-related challenges, the model achieved an accuracy rate of 81%. On unseen test data, it effectively generated a five-year projection that shows a gradual downward trend that closely resembles the test data's characteristics.

7.3 Future Work

Considering the limitations and challenges addressed in sections 7.1 and 7.2, the solution can be improved and expanded as mentioned in the sub-sections below,

7.3.1 Efficient Data Storage & Management

The structuring of the data was the main area of challenge during the project. Several strategies may be employed to deal with this issue,

- Adopting a standardised table format with similar table structures and naming guidelines. This strategy will make processing future data easier.
- Implementing a version control system to assure data adherence to a predetermined format and track dataset revisions.

The following steps are recommended for improving the current storage system in future developments:

- Utilising a time series database, such as TimeScaleDB or Clickhouse (indicated in [20]) as a central storage system that can accommodate real-time data from smart sensors (described in section 7.3.2). Additionally, it offers the ability to integrate with visualisation tools and store & manage large amounts of data as well.
- Using automation scripts or real-time ETL technologies [21] to organise, clean, and pre-process data, which reduces work and potential errors.
- Besides addressing data management concerns, as elaborated in Section 2.1.1 and referenced in [36], the utilization of smart energy systems like smart grids for electricity and smart gas grids will enable us to oversee and monitor performance, while also utilizing the results for effective energy management.
- Managing and utilizing the data like the hourly consumption, count of appliances, their usage timings, average usage duration, and the patterns of appliance utilization

influenced by weather conditions or holidays will help an in-depth study of consumption and its optimization, in conjunction with the use of a dashboard.

7.3.2 Live Data and Dashboard

A future enhancement for the PowerBI Dashboard could involve the creation of a live-streaming dashboard that receives real-time data feeds. The source of this real-time data could be smart sensor networks, yielding a substantial volume of data as elaborated in [23]. To accommodate this data influx, a suitable database is necessary for storage, as outlined in section 7.3.1. The live data can then be fed into PowerBI through either Streaming Data or Push feature [22]. The visual components within the dashboard may remain consistent or undergo modifications; however, the primary difference lies in the method of data loading between the streaming and conventional dashboards. However, establishing a streaming dashboard necessitates several components, including the installation of smart sensors to capture real-time data, the design of a suitable database to store the data, a Power BI Pro license, and the configuration of a PowerBI desktop application. The advantages of the architecture include improved performance tracking and efficiency, diligent monitoring, and timely alerts once these elements are in place. Additionally, regular system maintenance is recommended for optimal performance over time.

7.3.3 Integration of Timeseries Model with Dashboard

Another potential research study entails integrating the predictive model into PowerBI. To achieve this, one approach is deploying the machine learning model through Azure Machine Learning services [24] and seamlessly integrating its projections with PowerBI [25]. On the other hand, if the current SARIMAX model is to be integrated, it is advisable to deploy an API using tools like Flask. This API can be hosted on a cloud platform to expose the forecasts. In this context, endpoints will need to be established, and these endpoints can then be employed within Power BI to acquire the relevant data [26]. The data retrieval will be done through the 'Get Data from Web' option provided by PowerBI.

References

- [1] <https://www.gov.uk/government/publications/net-zero-strategy>
- [2] J. Chontanawat, "Relationship between energy consumption, CO2 emission and economic growth in ASEAN: Cointegration and causality model," vol. 6, pp. 660–665, 2020, doi: 10.1016/j.egy.2019.09.046. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2352484719305517>
- [3] https://en.wikipedia.org/wiki/University_of_Stirling
- [4] https://www.stir.ac.uk/about/our_people/
- [5] <https://www.circularise.com/blogs/scope-1-2-3-emissions-explained>
- [6] <https://www.gov.uk/government/collections/government-conversion-factors-for-company-reporting>
- [7] R. Sousa, R. Miranda, A. Moreira, C. Alves, N. Lori, and J. Machado, "Software Tools for Conducting Real-Time Information Processing and Visualization in Industry: An Up-to-Date Review," vol. 11, no. 11, 2021, doi: 10.3390/app11114800.
- [8] <https://sustainablesotlandnetwork.org/reports>
- [9] https://www.tensorflow.org/api_docs/python/tf/keras/layers/LSTM
- [10] https://school.stockcharts.com/doku.php?id=technical_indicators:moving_averages
- [11] <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.rolling.html>
- [12] <https://timeseriesreasoning.com/contents/time-series-decomposition/>
- [13] <https://analyticsindiamag.com/quick-way-to-find-p-d-and-q-values-for-arima/>
- [14] A. Kumar Dubey, A. Kumar, V. García-Díaz, A. Kumar Sharma, and K. Kanhaiya, "Study and analysis of SARIMA and LSTM in forecasting time series data," vol. 47, p. 101474, 2021, doi: 10.1016/j.seta.2021.101474. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2213138821004847>
- [15] https://www.statsmodels.org/dev/examples/notebooks/generated/exponential_smoothing.html
- [16] <https://www.statsmodels.org/dev/statespace.html#seasonal-autoregressive-integrated-moving-average-with-exogenous-regressors-sarimax>
- [17] M. B. Shrestha and G. R. Bhatta, "Selecting appropriate methodological framework for time series data analysis," vol. 4, no. 2, pp. 71–89, 2018, doi: 10.1016/j.jfds.2017.11.001. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2405918817300405>
- [18] <https://towardsdatascience.com/understanding-the-seasonal-order-of-the-sarima-model-ebef613e40fa>
- [19] <https://medium.com/illumination/mape-vs-smape-when-to-choose-what-be51a170df16>
- [20] A. Struckov, S. Yufa, A. A. Visheratin, and D. Nasonov, "Evaluation of modern tools and techniques for storing time-series data," vol. 156, pp. 19–28, 2019, doi: 10.1016/j.procs.2019.08.125. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1877050919310439>
- [21] K. C. Mondal, N. Biswas, and S. Saha, "Role of Machine Learning in ETL Automation," presented at the Proceedings of the 21st International Conference on Distributed Computing and Networking, Kolkata, India, 2020, doi: 10.1145/3369740.3372778 [Online]. Available: <https://doi.org/10.1145/3369740.3372778>
- [22] <https://learn.microsoft.com/en-us/power-bi/connect-data/service-real-time-streaming>
- [23] M. Jaradat, M. Jarrah, A. Bousselham, Y. Jararweh, and M. Al-Ayyoub, "The Internet of Energy: Smart Sensor Networks and Big Data Management for Smart Grid," vol. 56, pp. 592–597, 2015, doi: 10.1016/j.procs.2015.07.250. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1877050915017317>
- [24] <https://learn.microsoft.com/en-us/azure/machine-learning/concept-automl-forecasting-methods?view=azureml-api-2>

- [25] <https://learn.microsoft.com/en-us/power-bi/connect-data/service-aml-integrate>
- [26] <https://towardsdatascience.com/build-live-updating-dashboards-with-dash-and-power-bi-b82edcc0566d>
- [27] C. Thomas, J. Rolls, and T. Tennant, The GHG indicator: UNEP guidelines for calculating greenhouse gas emissions for businesses and non-commercial organisations. UNEP Paris, 2000.
- [28] J. Yu, Y. Yu, and T. Jiang, "Structural factors influencing energy carbon emissions in China's service industry: an input-output perspective," vol. 29, no. 32, pp. 49361–49372, 2022, doi: 10.1007/s11356-022-19287-8. [Online]. Available: <https://doi.org/10.1007/s11356-022-19287-8>
- [29] R. K. Sinha and N. D. Chaturvedi, "A review on carbon emission reduction in industries and planning emission limits," vol. 114, p. 109304, 2019, doi: 10.1016/j.rser.2019.109304. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S136403211930512X>
- [30] L. Xie, H. Yan, S. Zhang, and C. Wei, "Does urbanization increase residential energy use? Evidence from the Chinese residential energy consumption survey 2012," vol. 59, p. 101374, 2020, doi: 10.1016/j.chieco.2019.101374. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1043951X1930135X>
- [31] Y. Fan, L.-C. Liu, G. Wu, and Y.-M. Wei, "Analyzing impact factors of CO2 emissions using the STIRPAT model," vol. 26, no. 4, pp. 377–395, 2006, doi: 10.1016/j.eiar.2005.11.007. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0195925506000059>
- [32] O. Robinson, S. Kemp, and I. Williams, "Carbon management at universities: a reality check," vol. 106, pp. 109–118, 2015, doi: 10.1016/j.jclepro.2014.06.095. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0959652614007082>
- [33] G. P. Hammond and J. B. Norman, "Decomposition analysis of energy-related carbon emissions from UK manufacturing," vol. 41, no. 1, pp. 220–227, 2012, doi: 10.1016/j.energy.2011.06.035. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S036054421100421X>
- [34] <https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions>
- [35] O. J. Robinson, A. Tewkesbury, S. Kemp, and I. D. Williams, "Towards a universal carbon footprint standard: A case study of carbon management at universities," vol. 172, pp. 4435–4455, 2018, doi: 10.1016/j.jclepro.2017.02.147. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0959652617303736>
- [36] V. Kourgiouzou, A. Commin, M. Dowson, D. Rovas, and D. Mumovic, "Scalable pathways to net zero carbon in the UK higher education sector: A systematic review of smart energy systems in university campuses," vol. 147, p. 111234, 2021, doi: 10.1016/j.rser.2021.111234. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1364032121005219>
- [37] N. Ma, W. Y. Shum, T. Han, and F. Lai, "Can Machine Learning be Applied to Carbon Emissions Analysis: An Application to the CO2 Emissions Analysis Using Gaussian Process Regression," vol. 9, 2021, doi: 10.3389/fenrg.2021.756311. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fenrg.2021.756311>
- [38] J.-S. Chou and D.-S. Tran, "Forecasting energy consumption time series using machine learning techniques based on usage patterns of residential householders," vol. 165, pp. 709–726, 2018, doi: 10.1016/j.energy.2018.09.144. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0360544218319145>
- [39] L. K. Murugesan, R. Hoda, and Z. Salcić, "Design criteria for visualization of energy consumption: A systematic literature review," vol. 18, pp. 1–12, 2015, doi: 10.1016/j.scs.2015.04.009. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2210670715000499>

- [40] W. Reitler, M. Rudolph, and H. Schaefer, "Analysis of the factors influencing energy consumption in industry: A revised method," vol. 9, no. 3, pp. 145–148, 1987, doi: 10.1016/0140-9883(87)90019-3. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/0140988387900193>
- [41] Y. Himeur, K. Ghanem, A. Alsalemi, F. Bensaali, and A. Amira, "Artificial intelligence based anomaly detection of energy consumption in buildings: A review, current trends and new perspectives," vol. 287, p. 116601, 2021, doi: 10.1016/j.apenergy.2021.116601. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0306261921001409>
- [42] T. Falatouri, F. Darbanian, P. Brandtner, and C. Udokwu, "Predictive Analytics for Demand Forecasting – A Comparison of SARIMA and LSTM in Retail SCM," vol. 200, pp. 993–1003, 2022, doi: 10.1016/j.procs.2022.01.298. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1877050922003076>
- [43] https://facebook.github.io/prophet/docs/quick_start.html
- [44] <https://towardsdatascience.com/time-series-forecasting-using-auto-arma-in-python-bb83e49210cd>
- [45] <https://online.stat.psu.edu/stat510/lesson/4/4.1>
- [46] <https://towardsdatascience.com/time-series-forecast-in-python-using-sarimax-and-prophet-c970e6056b5b>
- [47] <https://towardsai.net/p/data-visualization/electricity-production-forecasting-using-arma-model-in-python>
- [48] <https://blog.paperspace.com/anomaly-detection-isolation-forest/>
- [49] <https://www.datatechnotes.com/2020/04/anomaly-detection-with-one-class-svm.html>
- [50] <https://stephenallwright.com/rmse-vs-mape/>
- [51] https://www.statsmodels.org/dev/examples/notebooks/generated/exponential_smoothing.html
- [52] <https://www.datacamp.com/tutorial/power-bi-dashboard-tutorial>
- [53] <https://community.fabric.microsoft.com/t5/Desktop/lookup-value-in-another-table-which-matches-value-in-another/td-p/1537454>
- [54] <https://www.section.io/engineering-education/missing-values-in-time-series/>
- [55] <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html>

Appendix 1 - Python Libraries Used

Library	Version
datetime	3.9.13
matplotlib	3.5.2
NumPy	1.21.5
pandas	1.4.4
seaborn	0.11.2
math	3.9.13
statsmodels	0.13.2
sklearn	1.0.2
pmdarima	2.0.3

Table 7.1 Used Python Library Versions

Appendix 2 – User Guide to Append Data in PowerBI Tables

To append the new table

- For reference, the existing 'Electricity' table is shown below and will be used throughout,
In Excel:

	A	B	C	D	E	F	G	H	I
1	date_time	air castle	air cottage	acy	alang	friar	gma	jfc	lyon
2	Jul-05	11622	1044	0	6278	4343	2754	23819	0
3	Aug-05	9750	0	0	5996	0	2658	2619	0
4	Sep-05	11800	0	0	5858	0	3529	0	0
5	Oct-05	11820	1281	0	7351	4346	4123	15439	0
6	Nov-05	11300	0	0	9266	0	4389	2964	0
7	Dec-05	12728	0	0	10569	0	4152	0	0

In PowerBI:

date_time	Site	Units KWHr	Year	Merged	CO2
01 July 2005	CentroHouse & ScionHouse	0	2005	2005electricity	0
01 August 2005	CentroHouse & ScionHouse	0	2005	2005electricity	0
01 September 2005	CentroHouse & ScionHouse	0	2005	2005electricity	0
01 October 2005	CentroHouse & ScionHouse	0	2005	2005electricity	0
01 November 2005	CentroHouse & ScionHouse	0	2005	2005electricity	0
01 December 2005	CentroHouse & ScionHouse	0	2005	2005electricity	0

[Note** The used dataset has been provided with the PowerBI file for future reference.]

- Here is a sample of the new file that is going to be added.
Please ensure that the header contains the correct **Site Name, such as 'Airthrey Castle,' to avoid issues in the data addition in subsequent steps. You can refer to the PowerBI table above for reference.

New Table:

	A	B	C	D	E	F	G	H
1	date_time	Airthrey C	Airthrey C	Airthrey C	AlanGrang	FriarsCroft	Gardens a	John Forty
2	Jul-25	11622	1044	0	6278	4343	2754	23819
3	Aug-25	9750	0	0	5996	0	2658	2619
4	Sep-25	11800	0	0	5858	0	3529	0
5	Oct-25	11820	1281	0	7351	4346	4123	15439
6	Nov-25	11300	0	0	9266	0	4389	2964

- Open the PowerBI desktop app -> Click on File -> Open Report -> Open the "Dashboard" file.
- Go to the **Home** tab -> In Data field -> **Excel Workbook** -> select the Excel file and the spreadsheet (here, Sheet1) to be added.
- Then click on **Transform Data**
- The **Power Query Editor** will be opened.
- Then **select** the "date_time" column -> Go to the "**Transform**" tab -> Click **Unpivot Columns**.
- And click on "**Unpivot Other Columns**" (**Please make sure to select the date_time column before doing it)
- Double Click on Column Header to change the column name as below,
Attribute to 'Site'
Value to 'Units KWHr' -> to match the existing Electricity table in the dashboard.
- Then **Close and Apply** in Home tab to add the sheet to existing data.
- Once the Excel Sheet is added -> Click on **Transform Data** in Queries
- Choose the table** you want to append the data to. In this case, I've opted for the 'Electricity' table, as I'm adding the new data to it.
- On the **Electricity table** -> **Home** tab -> In **Combine** field -> Click **Append Queries** -> Again **Append Queries**.
- In the Append tab -> Select Two Tables -> Select the new table (Sheet1) that need to be appended -> Click Ok.

15. And click **Close & Apply** to go back to the dashboard or table.

16. The **CO2 will be automatically calculated** by the query.

****However, please make sure to add the CO2 conversion rates in the “co2 rates” table in order for the query to run.**

To check and verify the data is appended correctly please go to the **Data View** option in the dashboard page as shown below,

17. Finally **save the work** in PowerBI file.

To remove the new table from the model

1. To remove the new file from the model -> Go to **Transform Data** on the **Home** tab -> In the Power Query Editor -> Select the **new table** (here, Sheet1) -> **Right click** and select **Enable Load** to uncheck & disable it.
2. Close & Apply -> It will be removed from the model
(** Please select an appropriate name for the new file, as I named it 'Sheet1' for testing purposes.)

To roll back the append table query (only if required)

If the table has not been added properly or the appending action needs to be rolled back in any case please follow the steps below,

1. Go to the **Home** tab -> **Transform Data**
2. In **Power Query Editor** -> **Select the table** 'Electricity' table in our case
3. In **Query Setting** on the right -> **Applied Steps** -> Delete the 'Append Query' by clicking on **X mark**.
4. **Close & Apply**

Appendix 3 – Installation Guide for PowerBI Desktop

1. Download Power BI from official website <https://powerbi.microsoft.com/en-us/desktop/>
2. Select the version (32-bit or 64-bit) based on your system.
3. Locate the downloaded installer file and install the application.
4. Read and accept the license terms by selecting "I accept the terms in the license agreement."
5. Once Power BI Desktop is installed launch it from the Start Menu or the desktop.
6. Run Power BI Desktop by signing in with the University email account to get the appropriate licence.
7. Finally, from File tab open the energy_dashboard.pbix file.

END OF DOCUMENT