Faculty of Engineering and Digital Technologies

MSc Engineering Management

COS7049-B Statistical Applications of Industrial Big Data

Title: Analysing Household Electricity Consumption in the UK: A Comprehensive Study Using the National Energy Efficiency Data-Framework (NEED) from 2005-2022

UB Number:23045085

Name: Kamomoe Annie Njambi

Year 2024

Table of Contents

List of Figures

List of Tables

Abstract

This report investigates the influence of building characteristics and geographical location on household electricity consumption in the UK, utilizing data from the National Energy Efficiency Data-Framework (NEED). Employing time series analysis, multiple linear regression, and machine learning techniques, the study reveals significant correlations between property features (such as number of bedrooms and age) and energy usage patterns. The research highlights both the strengths and limitations of current predictive models, proposing improvements in methodology and data analysis approaches.

Additionally, the study addresses the legal, social, and ethical considerations associated with energy consumption data analysis. The findings contribute to a more nuanced understanding of domestic energy use, offering insights for policymakers and energy efficiency initiatives while emphasizing the need for responsible data practices in the field of energy analytics.

This study provides new insights into regional disparities in energy use and proposes novel methods to address predictive modelling limitations.

Chapter One: Introduction

1.1 Background

The United Kingdom's transition towards smart and sustainable urban cities remains a critical focus in addressing climate change and improving energy efficiency. Building upon the analysis conducted in Coursework One, which examined broader energy trends from 1990 to 2022, this study narrows its focus to household energy consumption patterns in the UK, specifically examining data from 2005 to 2022.

The availability of large datasets like the NEED consumption data tables provides opportunities for advanced analytics. These datasets encompass millions of entries, allowing for granular analysis of trends such as property-level energy use. Through big data analytics, patterns that were previously obscured can now inform more targeted and effective policymaking.

Despite significant advancements, gaps remain in understanding how factors such as building characteristics and regional variations influence energy consumption patterns. By addressing these gaps, this study aims to provide actionable insights to support the UK's energy transition.

1.2 Summary of Coursework One

In the previous coursework one, my team ‘Data detectives’ explored three key datasets related to the UK's energy landscape which were energy consumption patterns, renewable energy by local authority and renewable energy and waste sources. The analysis revealed significant trends, such as:

- A gradual shift towards more energy-efficient housing.
- Regional disparities in renewable energy adoption, with regions like London having lower renewable energy contributions.
- An overall increase in renewable energy capacity across the UK.

These findings underscore the importance of understanding household energy use, particularly the role of building characteristics and regional variations, leading to the design of this study.

1.3 Focus of the Current Study

This study utilizes specific dataset from the National Energy Efficiency Data-Framework (NEED), a publicly available and reliable source provided by the UK government (Department for Energy Security and Net Zero 2024a). With a focus on the following data:

1. Electricity consumption by number of bedrooms (2005-2022)
2. Electricity consumption by property age (2005-2022)
3. Electricity consumption by country and region (2005-2022)

This selection allows for a comprehensive analysis of how building characteristics and geographical location influence energy use in UK households. This analysis applies time series techniques and machine learning using Python to uncover trends.

1.4 Research Objectives

The primary objectives of this study are to:

1. Analyse the relationship between property age and energy consumption trends over time.
2. Investigate how the number of bedrooms affects electricity use.
3. Examine regional variations in household energy consumption across the UK.

These objectives will be supported by integrating findings from relevant research papers, including studies on socio-economic factors influencing energy consumption.

1.5 Significance of the Study

While many studies explore socio-economic factors, fewer focus on granular building characteristics. Understanding household energy consumption patterns will be vital for several reasons:

- Making informed policy decisions on building regulations and energy efficiency standards.
- Help in developing targeted interventions for reducing energy consumption.
- Contributes to the broader goal of creating sustainable urban environments that align with UK's climate goal.
- Provide insights into the effectiveness of past energy efficiency initiatives.

This study will address key challenges such as data completeness and regional disparities in energy consumption, aiming to provide a detailed understanding of UK household energy use patterns.

1.6 Structure of the Report

The report is structured as follows:

- Abstract: Summary of the study's focus, methodology, and key findings
- Chapter 1: Introduction-Background on UK's energy transition and previous findings from Coursework One.
- Chapter 2: Literature Review- comparing NEED report findings with other studies, highlighting trends in energy efficiency and consumption patterns.
- Chapter 3: Methodology using a three-stage process: data preparation, exploratory data analysis, and statistical analysis/machine learning.
- Chapter 4: Analysis using time series models (ARIMA and SARIMAX) and multiple linear regression to examine relationships between property features and energy consumption. Discussion of strong points & Limitations
- Chapter 5: Conclusion and Recommendations

Chapter Two: Literature Review

This chapter critically analyses the main research paper and compares it with other relevant studies to provide a comprehensive overview of household energy consumption patterns in the UK.

2.1 Main Research Paper Analysis

The primary source for this study is the "National Energy Efficiency Data-Framework (NEED) report: Summary of analysis 2024" published by the UK government. This report uses regression analysis and trend analysis to provide insights into domestic energy consumption patterns (Department for Energy Security & Net Zero 2024b). Key important findings from the report include:

1. Significant year-on-year reductions in median domestic electricity consumption between 2021 and 2022, attributed to rising energy prices and the increasing cost of living pressures.
2. Larger properties and those with higher adult occupancy show higher median electricity consumption.
3. More energy-efficient properties (higher EPC ratings) have lower energy consumption overall.
4. Buildings built recently consume less gas, this can be seen in the new builds since 2010 showing progressively lower electricity consumption.

2.2 Comparative Analysis

To contextualize the findings of the NEED report, this section examines similar studies on household energy consumption patterns.

2.2.1 Property Characteristics and Energy Consumption

The NEED report's findings align with other studies. The House of Commons Library report confirms that newer homes have much higher energy efficiency ratings (Bolton 2024). This consistency strengthens the reliability of the NEED data. Quantitative data from Uswitch provides context: the average British household (2.4 people) uses 2,700 kWh of electricity and 11,500 kWh of gas per year. For a medium-sized house (3 bedrooms, 2-3 people), this consumption costs £1,769.46 annually (Gallizzi 2023).

2.2.2 Energy Efficiency Trends

The Office for National Statistics report reinforces the NEED findings, revealing that new dwellings in England and Wales achieved a median EPC score of 84 (Band B) in the five years leading to 2024, compared to scores of 82 in England and 81 in Wales during the five years leading to 2013 (Office for National Statistics 2022). Statista reports a general downward trend in UK electricity consumption, with domestic users now consuming the most electricity (Statista 2024). This trend aligns with the NEED report's observations on improving energy efficiency.

2.2.3 Regional Variations and Urban-Rural Divide

While the NEED report offers valuable national-level insights, it is essential to account for regional and urban-rural disparities. Research by the Rural Services Network highlights that energy costs in rural areas are, on average, 10% higher, increasing to 17% in Yorkshire and the Humber. This urban-rural divide is further underscored by the stark contrast in energy efficiency: nearly 20% of rural homes are classified as highly energy inefficient, compared to just 2.4% in urban areas (Ellard 2023). Addressing the energy inefficiency of older properties, particularly in rural areas, could involve targeted subsidies or incentives for retrofitting energy-efficient measures, informed by region-specific data analysis. However, subsidies for retrofitting rural homes or incentives for energy-efficient appliances could address regional disparities.

2.2.4 Behavioural Insights and Policy Interventions

The importance of behavioural insights in energy consumption is increasingly recognized. The Behavioural Insights Team (Londakova et al. 2023) highlights how occupant habits, such as thermostat use and appliance settings, significantly influence energy efficiency. While the NEED report focuses on structural characteristics, integrating such behavioural data could enhance predictive models for policy interventions. By incorporating behavioural insights into data-driven models, policymakers can design more effective interventions that align with occupant habits, further advancing energy efficiency goals. Machine learning models could also integrate thermostat use data to predict seasonal trends more effectively.

This comparative analysis reveals that while the NEED report provides comprehensive current data, integrating insights from other studies offers a more detailed understanding of UK household energy consumption. Key points are:

1. The significant impact of property characteristics on energy consumption is consistently observed across studies.

2. There's a clear urban-rural divide in energy efficiency and consumption patterns.
3. Behavioural insights and occupant habits play a crucial role in energy consumption, an aspect not fully captured by the NEED report.
4. The slowing rate of improvement in overall housing stock energy efficiency suggests that addressing older properties remains a significant challenge, particularly in rural areas.

2.3 Critical Analysis

Despite its limitations, the NEED report's use of regression analysis provides a robust framework for understanding key trends, offering valuable insights for policymakers. However, there are several areas that warrant further investigation:

1. **Methodology limitations:** The NEED report relies heavily on data from government schemes, particularly the Energy Company Obligation (ECO) schemes (Department for Energy Security & Net Zero 2024b). This may introduce bias towards households participating in these schemes.
2. **Behavioural aspects:** While the report provides data on energy efficiency measures, it could benefit from incorporating insights on occupant behaviour, as highlighted by the Share the Warmth initiative (Taylor 2023).

In conclusion the NEED report demonstrates the potential of big data analytics in identifying macro-level consumption patterns. However, its findings could be significantly enriched by leveraging machine learning algorithms to predict energy usage trends based on occupant behaviour and integrating real-time data sources such as smart meters. This literature review sets the foundation for our methodology and subsequent analysis, ensuring a comprehensive approach to understanding UK household energy consumption patterns while highlighting areas for further research and policy consideration.

Chapter Three: Methodology

3.1 Experiment Design, Description, and Evidence

The National Energy Efficiency Data-Framework (NEED) dataset provides valuable insights into factors affecting household energy consumption in the UK. The experiment process described involved three main stages:

1. **Data Preparation**

- Missing values were replaced with 0.
- Outliers were removed using the Interquartile Range method.
- Categorical variables were converted to numerical formats.
- Data reduction was performed by reducing the years from 2005-2022 to using 2011-2011.This ensured that the data used covered both England and Wales

2. **Exploratory Data Analysis (EDA)**

- Visualizations were created, including heatmaps, line plots, and bar graphs.
- These visualizations revealed patterns in energy consumption across different property types, ages, and number of bedrooms. A sample of the visualization is attached in the appendix.

3. **Statistical Analysis and Machine Learning**

- Multiple Linear Regression was used to analyse the impact of property age on energy consumption.
- Time Series Analysis was employed to identify trends in consumption from 2011-2022 based on number of bedrooms.
- This analysis accounted for seasonal variations and long-term shifts in consumption patterns.

This comprehensive approach provided a robust framework for understanding the complex relationships between building characteristics, geographical location, and household energy consumption in the UK.

Below is a detailed flowchart highlighting the experiment process and an example of pseudocode used in the experiment.

3.1.1 Flowchart

Here's a flowchart representing the methodology.

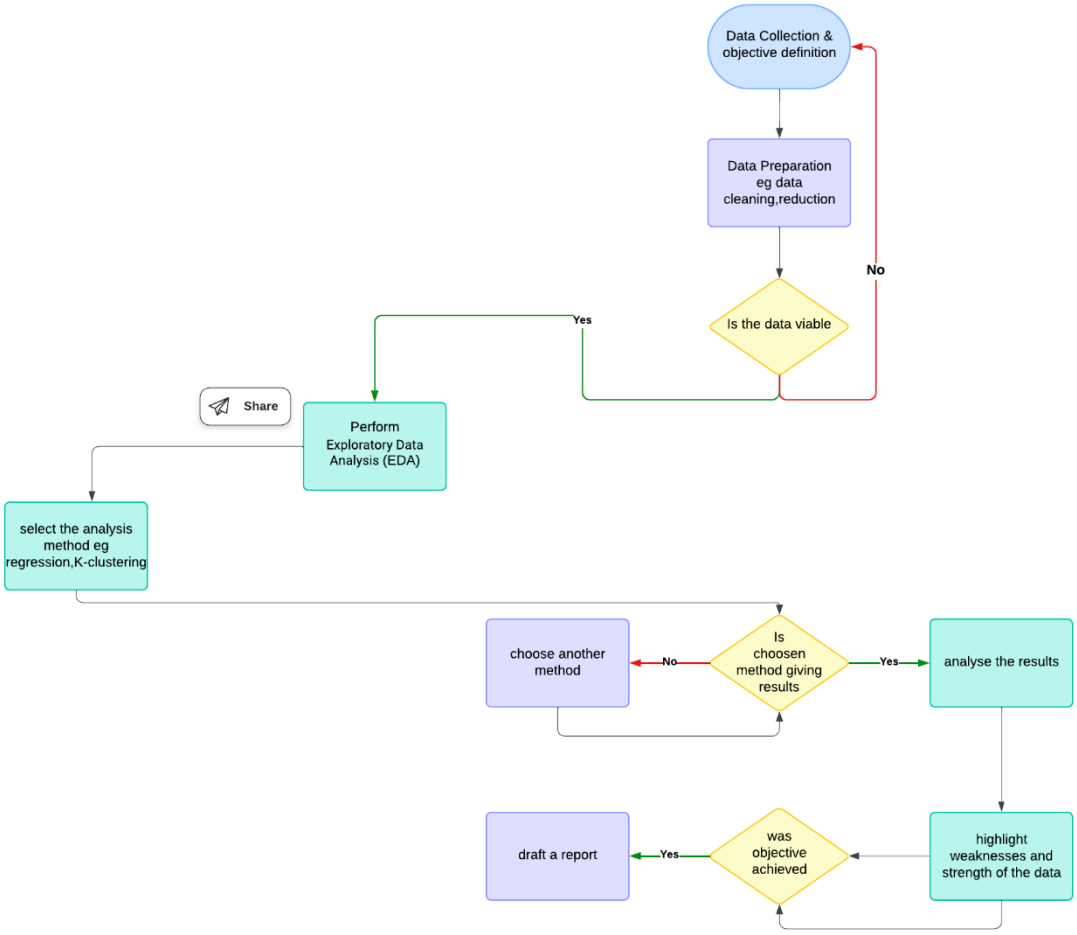


Figure 1:Flowchart.

3.1.2 Pseudo-Code for the data preparation and cleaning process

Load Consumption_headline_EW_2022 Excel file.

Select relevant data sheets (bedrooms, property age, region)

Extract tables (number of dwellings, mean consumption, median consumption)

Assess data quality.

Replace missing values with 0.

Reduce dataset to 2011-2022 (England and Wales data)

Save prepared tables to perform EDA.

3.2 Interpretation of the results

Time Series Analysis

Focuses on analysing data points collected over time to identify trends, seasonality, and patterns. This paper used both models ARIMA & SARIMAX to forecast future electricity consumption based on historical data.

Key Findings

- The models use only 12 observations, which is quite limited for a time series analysis.
- The models suggest a possible slight decrease in mean electricity consumption over time.
- Both models fit the data reasonably well, with no significant autocorrelation in residuals.
- The residuals appear to be normally distributed and show consistent variance over time.

Limitations

- The simple model structure may not capture complex patterns in the data.
- The small dataset limits the models' ability to identify long-term trends.

Multiple Linear Regression

It is a statistical method that estimates the relationship between one or more independent variables and a dependent variable. It provides interpretable coefficients that quantify the impact of each variable on electricity usage. Overall model is statistically significant. The F-statistic p-value is 4.62e-06.

Key Findings

- Strong overall relationship between number of bedrooms and electricity consumption However, individual bedroom coefficients are not statistically significant.
- High multicollinearity detected (VIF values > 10 for all predictors)

Limitations

- Possible positive autocorrelation in residuals (Durbin-Watson: 1.370)
- High multicollinearity may affect coefficient stability.

Chapter Four: Analysis

4.1 Strong Aspects & Achievements of the paper

Some of the strong aspects include:

1. **Novelty of the Work:**

The study's focus on how the number of bedrooms affects electricity use through Multiple Linear Regression offers a unique perspective on household energy consumption. This approach provides valuable insights into the relationship between specific household characteristics and energy usage patterns.

Evidence: While previous research has examined broader demographic factors influencing energy consumption, this work's specific analysis of bedroom count contributes novel insights that can inform targeted energy efficiency initiatives (Reddy et al. 2023).

2. **Innovative Solutions**

The implementation of Auto ARIMA for time series forecasting represents an innovative technique that automates the selection of optimal ARIMA model parameters. This approach enhances the robustness and efficiency of the time series analysis.

Evidence: Auto ARIMA operates by performing differencing tests to identify the optimal order of differencing (d) and then systematically fitting models within predefined ranges for parameters such as start_p, max_p, start_q, and max_q. This automated approach streamlines the model selection process, often enhancing forecasting accuracy when compared to manual parameter selection methods (SKTIME, 2019; Pulagam, 2020).

Comprehensive Evaluation Methodology

The study employs a robust methodology, including thorough evaluations of both Multiple Linear Regression and ARIMA models. The use of residual analysis, multicollinearity checks, and autocorrelation assessments ensures the validity and reliability of the models.

Evidence: The importance of comprehensive model evaluation is highlighted in recent studies. For instance, a study on electricity consumption prediction using machine learning emphasizes the need for thorough model assessment to ensure reliable forecasts (Reddy et al. 2023). Additionally, the use of ARIMA and SARIMA models for real-world time series forecasting underscores the importance of preprocessing and rigorous evaluation techniques.

These strengths demonstrate the paper's contribution to advancing the understanding and predictive modelling of energy consumption patterns, supported by established methodologies in the field of electricity consumption forecasting.

4.2 Limitations

Assumptions or Constraints Not Addressed: The limitation of linear regression models in capturing complex relationships in energy consumption data is supported by the study mentioned in (Phan et al. 2024). This study explores the application of linear regression analysis for predicting energy consumption and discusses the importance of considering non-linear approaches.

Limited Scope or Generalizability of Results: The importance of including a broader range of variables when modelling energy consumption is highlighted in (Ejigu Tefera 2024). This paper proposes a nonlinear ensemble deep learning model that incorporates multiple factors for residential energy consumption prediction, demonstrating the need for a more comprehensive approach.

Missing Aspects in Evaluation or Unexplored Dimensions: The significance of addressing autocorrelation and exploring temporal trends in energy consumption is supported by (Ejigu Tefera 2024). This study discusses the use of various deep learning models, including LSTM and BiLSTM, which are designed to capture temporal dependencies in time series data.

Additionally, the importance of considering multiple factors in energy consumption prediction is emphasized in (Bara et al. 2024), which discusses the use of multiple linear regression for predicting electrical energy consumption. This study also highlights the limitations of using a single approach and suggests exploring other algorithms for improved accuracy.

4.3 Comparative Evaluation of results

Below are some of the suggested Improvements and alternative approaches for the Statistical Methods used in this paper.

Incorporate Non-Linear Relationships: To improve the analysis, it would be beneficial to move beyond traditional Multiple Linear Regression, which assumes a straight-line relationship. Instead, consider using polynomial regression or generalized additive models (GAMs) that can capture more complex, non-linear relationships between the number of bedrooms and electricity consumption. Alternatively, machine learning techniques such as Random Forests can effectively model these non-linear relationships without needing explicit transformations.

Address Multicollinearity: If high Variance Inflation Factor (VIF) values indicate multicollinearity among the predictors, it is valid to use methods like ridge regression or principal component analysis (PCA). These techniques help reduce dimensionality and mitigate the negative effects of multicollinearity. Another option is Lasso regression, which not only addresses multicollinearity but also performs variable selection and regularization, enhancing both prediction accuracy and interpretability.

Robust Residual Analysis: A thorough residual analysis is crucial for validating our model. This should include tests for homoscedasticity (like the Breusch-Pagan test) and normality (such as the Shapiro-Wilk test) to ensure that our model assumptions hold true.

Feature Engineering: To create a more comprehensive model, we should enhance our feature set by including additional relevant variables that influence electricity consumption. Factors like household size, income level, and the presence of energy-efficient appliances could provide valuable insights. We can also leverage domain knowledge to create interaction terms or derived features that capture complex relationships between variables.

4.4 Legal, Social, Ethical, Professional, and Privacy Risks Related to the Analysis

4.4.1. Legal Risks

When handling personal data, it is essential to comply with data protection regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). These regulations impose strict rules for data collection, processing, and storage, with significant penalties for non-compliance. Some of the mitigation strategies are:

- **Clarifying Ownership**: When using third-party data, explicit agreements must be established to ensure proper usage rights and prevent unauthorized use. For instance, licensing agreements or data-sharing contracts should outline permissible uses and liabilities.
- **Mitigation**: Conducting Data Protection Impact Assessments (DPIAs) and regularly reviewing compliance practices can help identify and address potential regulatory violations.

4.4.2. Social Risks

Predictive models, if based on skewed or incomplete datasets, may unintentionally reinforce existing biases, leading to unequal treatment of certain demographic groups in policymaking or resource allocation.

For instance, if data disproportionately represent urban households, rural areas may receive inadequate energy efficiency support. Results must be communicated transparently to prevent misuse or misunderstanding, which could erode public trust. This is particularly critical when policy decisions or funding allocations are based on the model's outcomes. Employ fairness audits and ensure datasets are representative of all demographic groups. Additionally, invest in public education efforts to demystify model predictions and their limitations.

4.4.3. Ethical Risks

Transparency and consent are foundational ethical principles in data analysis. Individuals whose data is analysed must provide informed consent, explicitly agreeing to the data's intended use. Repurposing energy consumption data for unauthorized applications, such as surveillance or targeted marketing, breaches ethical standards.

To mitigate this risk, one should maintain a transparent data usage policy and obtain explicit consent from participants. Ethical review boards or committees should evaluate the analysis plan to ensure compliance with ethical guidelines.

4.4.4. Professional Risks

Data analysts and scientists bear significant responsibility for the integrity and reliability of their models and findings.

- **Model Performance**: Poorly performing models or misinterpretation of results can lead to flawed decisions, undermining professional credibility and stakeholder trust.
- **Bias Awareness**: Professionals must recognize and address their biases during model development and interpretation to avoid skewed results.
- **Mitigation**: Regular peer reviews, model validation, and documentation of assumptions and limitations can enhance the reliability of analysis outcomes. Professionals should also engage in ongoing training to remain aware of best practices and emerging biases in data science.

4.4.5. Privacy Risks

Protecting individual privacy is paramount, particularly when dealing with sensitive household data. Data must be anonymized to remove personally identifiable information (PII). However, anonymization techniques can sometimes be circumvented, posing re-identification risks. Robust cybersecurity measures are essential to protect data from breaches and unauthorized access.

One should employ advanced anonymization techniques, such as differential privacy, which adds noise to data to prevent re-identification while preserving overall utility. Implement strong encryption protocols for data storage and transfer and conduct regular security audits to identify vulnerabilities.

Chapter Five: Conclusion

This study has provided valuable insights into the factors influencing household electricity consumption in the UK, focusing on the impact of property characteristics and geographical location. Through a combination of time series analysis, multiple linear regression, and machine learning techniques, we have identified significant relationships between the number of bedrooms, property age, and regional factors on electricity usage. Our findings highlight the potential for targeted energy efficiency measures based on household characteristics. However, the analysis also revealed limitations in current modelling approaches, suggesting opportunities for more sophisticated techniques to capture non-linear relationships and incorporate a broader range of variables. Future research should focus on addressing these limitations and exploring the ethical implications of using household data for energy consumption prediction.

References

Bara, A., Bara, A., Bara, A., Palanga Eyouleki Tcheyi Gnadi, Palanga Eyouleki Tcheyi Gnadi, Palanga Eyouleki Tcheyi Gnadi, Yao, B., Yao, B., Yao, B., Kuevidjen Dosseh and Nomenyo Komla (2024) Multiple Linear Regression to Predict Electrical Energy Consumption Based on Meteorological Data: Application to Some Sites Supplied by the CEB in Togo. *American Journal of Applied Sciences* 21 (1), Science Publications15–27.

Bolton, P. (2024) *Energy Efficiency of UK Homes*. https://commonslibrary.parliament.uk/research-briefings/cbp-9889/ Accessed 24 December 2024.

Department for Energy Security & Net Zero (2024a) *National Energy Efficiency Data-Framework (NEED) report: Summary of Analysis 2024*. https://www.gov.uk/government/statistics/national-energy-efficiency-data-framework-need-consumption-data-tables-2024 Accessed 23 October 2024.

Department for Energy Security & Net Zero (2024b) *National Energy Efficiency Data-Framework (NEED): Summary of Analysis, Great Britain, 2024*. https://assets.publishing.service.gov.uk/media/66e0203ad65d5c23df086710/NEED-report-june-2024.pdf Accessed 23 October 2024.

Ejigu Tefera (2024) Nonlinear Ensemble Deep Learning Model for Energy Consumption Prediction with Bayesian Optimization. *Communications on Applied Nonlinear Analysis* 32 (1), 155–172.

Ellard, S. (2023) *Bridging the energy-efficiency gap: Addressing Rural Home Inequalities*. https://www.insidehousing.co.uk/comment/bridging-the-energy-efficiency-gap-addressing-rural-home-inequalities-82135 Accessed 2 December 2024.

Gallizzi, B. (2023) *100+ UK Energy Statistics 2023*. https://www.uswitch.com/gas-electricity/studies/energy-statistics/ Accessed 2 December 2024.

Londakova, K., Human, S., Gross, M., Park, T. and Chan, D.E. (2023) *New Survey Shows a UK energy-saving Campaign Is Much Needed*. https://www.bi.team/blogs/new-survey-shows-a-uk-energy-saving-campaign-is-much-needed/ Accessed 2 December 2024.

Office for National Statistics (2022) *Energy Efficiency of Housing in England and Wales - Office for National Statistics*. https://www.ons.gov.uk/peoplepopulationandcommunity/housing/articles/energyefficiencyofhousinginenglandandwales/2022 Accessed 2 December 2024.

Phan, Anh, M. and Hung, H. (2024) Using Linear Regression Analysis to Predict Energy Consumption. *Research Square (Research Square)* Research Square (United States).

Pulagam, S. (2020) *Time Series Forecasting Using Auto ARIMA in Python*. https://towardsdatascience.com/time-series-forecasting-using-auto-arima-in-python-bb83e49210cd Accessed December 2024.

Reddy, G.V., Aitha, L.J., Poojitha, Ch., Shreya, A.N., Reddy, D.K. and Meghana, G. Sai. SK Tummala, S Kosaraju, PB Bobba & SK Singh (editors), (2023) Electricity Consumption Prediction Using Machine Learning. *E3S Web of Conferences* 391, 01048.

SKTIME (2019) *AutoARIMA — Sktime Documentation*. https://www.sktime.net/en/latest/api_reference/auto_generated/sktime.forecasting.arima.AutoARIMA.html Accessed December 2024.

Statista (2024) *Monthly Electricity Consumption from All Electricity Suppliers in the United Kingdom (UK) from January 2014 to May 2024*. Statista Research Department https://www.statista.com/statistics/322996/monthly-electricity-consumption-from-all-electricity-suppliers-in-the-united-kingdom-uk/ Accessed 10 December 2024.

Taylor, L. (2023) *Share the Warmth: Using Behavioural Insights to Reduce Energy Usage among Homeowners*. Local Government Association https://www.local.gov.uk/case-studies/share-warmth-using-behavioural-insights-reduce-energy-usage-among-homeowners Accessed 2 December 2024.

Appendices

Appendix A: Work Plan

Start Date	End Date	Tasks
19th Nov 2024	22nd Nov 2024	• Select the main research paper to focus on • Identify 4-5 related papers. • Start reading and taking notes on all selected papers.
23rd Nov 2024	29th Nov 2024	• Complete literature review • Outline the report structure. • Begin drafting the introduction and literature review sections.
30th Nov 2024	3rd Dec 2024	• Experiment Design and Planning using flow charts
4th Dec 2024	10th Dec 2024	• Data Exploration and Analysis
11th Dec 2024	20th Dec 2024	• Implementation and Evaluation
21st Dec 2024	27th Dec 2024	• Discussion and Conclusions
28th Dec 2024	2nd Jan 2025	• Review and revise the entire report. • Finalize all sections, including abstract and conclusions. • Prepare appendices with code, data. • Submit the coursework

Table 1:Workplan

Appendix B: Theory Figures

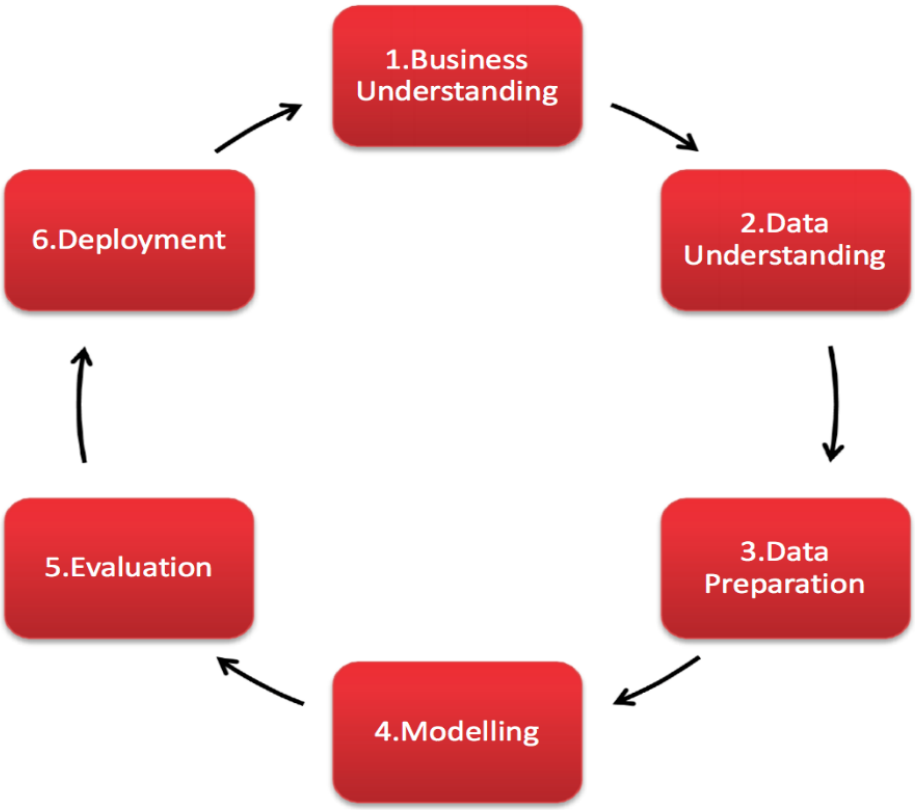


Figure 2:Crisp-DM Diagram

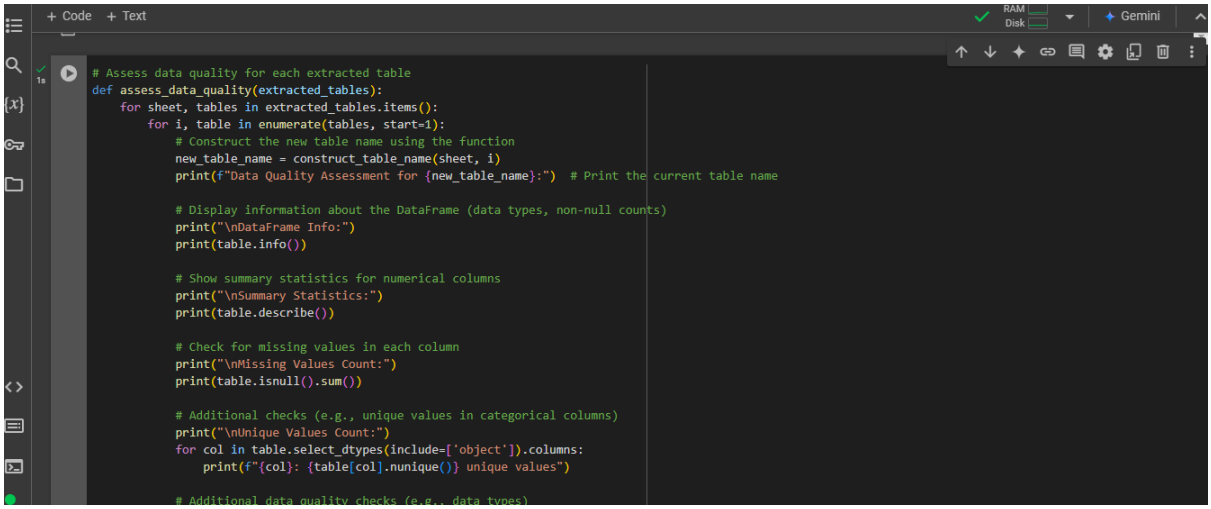
This model is a summary of what CW1 and CW2 entails.CW1 involved defining the role of the company, researching on appropriate data, understanding the data and lastly preparing it by selecting the data, cleaning it, formatting etc.

In CW2 the fourth and fifth step of the Crisp-DM model is implemented whereby the data is modelled to be analysed.

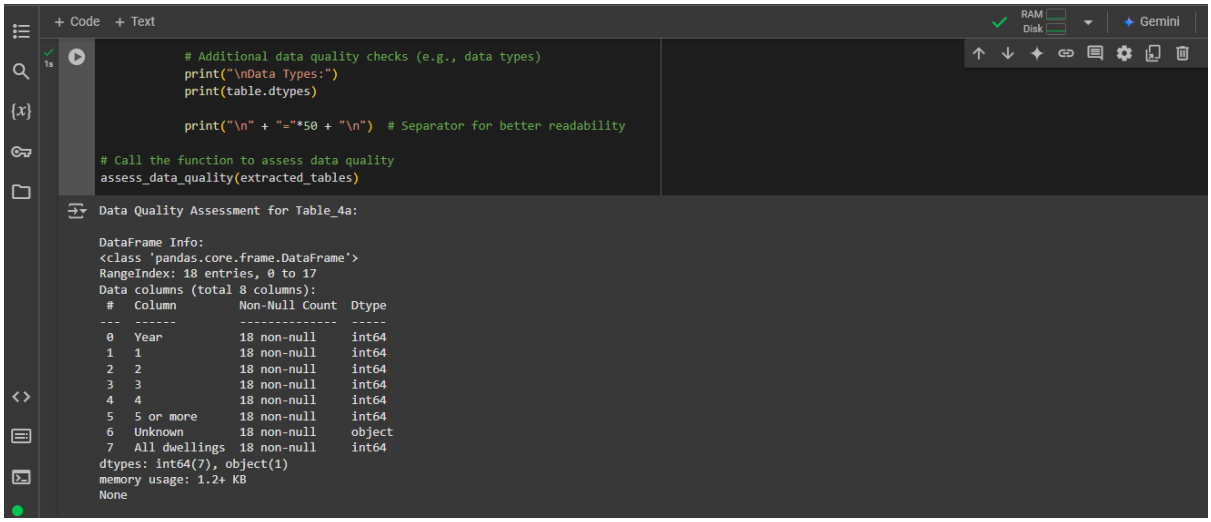
Appendix C: Data Preparation

C.1 Data Quality assessment code

The following function, assess_data_quality, evaluates the quality of data in extracted tables. It iterates over each table in the provided dataset, performs various checks such as identifying missing values, calculating summary statistics, and assessing data types. Below is the implementation and output sample.

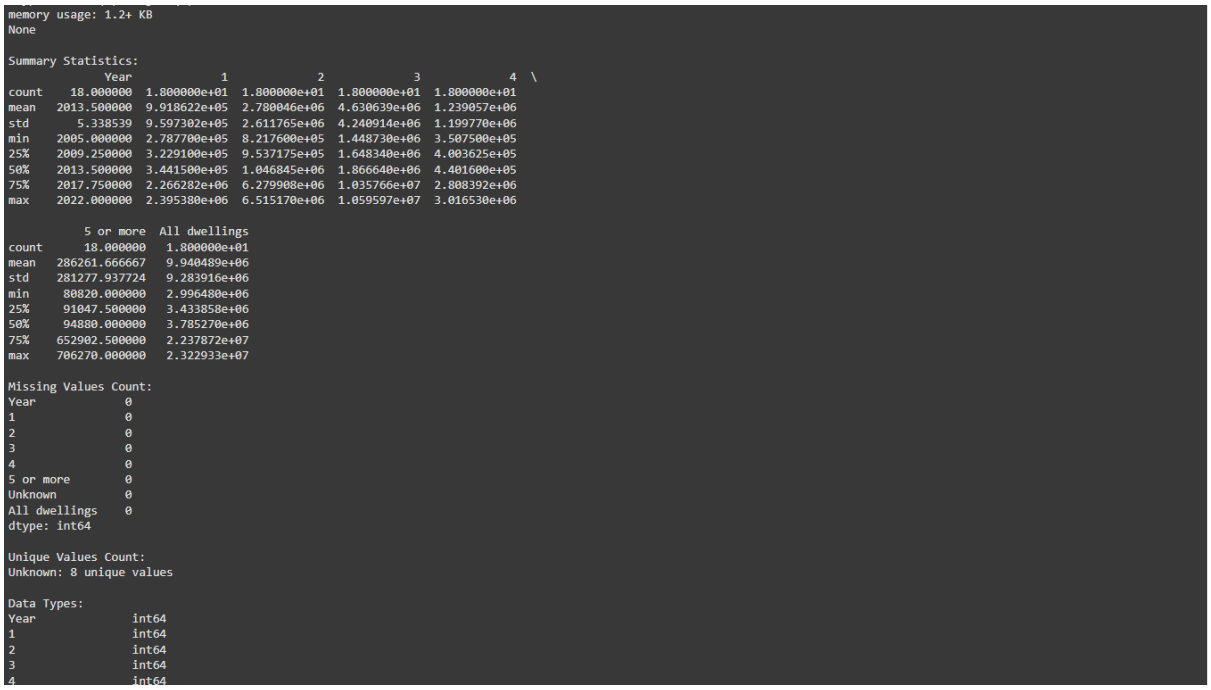


```
# Assess data quality for each extracted table
def assess_data_quality(extracted_tables):
    for sheet, tables in extracted_tables.items():
        for i, table in enumerate(tables, start=1):
            # Construct the new table name using the function
            new_table_name = construct_table_name(sheet, i)
            print(f"Data Quality Assessment for {new_table_name}:")  # Print the current table name

            # Display information about the DataFrame (data types, non-null counts)
            print("\nDataFrame Info:")
            print(table.info())

            # Show summary statistics for numerical columns
            print("\nSummary Statistics:")
            print(table.describe())

            # Check for missing values in each column
            print("\nMissing Values Count:")
            print(table.isnull().sum())

            # Additional checks (e.g., unique values in categorical columns)
            print("\nUnique Values Count:")
            for col in table.select_dtypes(include=['object']).columns:
                print(f"{col}: {table[col].nunique()} unique values")

            # Additional data quality checks (e.g., data types)
```



```
            # Additional data quality checks (e.g., data types)
            print("\nData Types:")
            print(table.dtypes)

            print("\n" + "="*50 + "\n")  # Separator for better readability

# Call the function to assess data quality
assess_data_quality(extracted_tables)
```

```
Data Quality Assessment for Table_4a:

DataFrame Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18 entries, 0 to 17
Data columns (total 8 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Year           18 non-null     int64
 1   1              18 non-null     int64
 2   2              18 non-null     int64
 3   3              18 non-null     int64
 4   4              18 non-null     int64
 5   5 or more      18 non-null     int64
 6   Unknown        18 non-null     object
 7   All dwellings  18 non-null     int64
dtypes: int64(7), object(1)
memory usage: 1.2+ KB
None
```



```
memory usage: 1.2+ KB
None

Summary Statistics:
              Year             1             2             3             4  \
count    18.000000  1.800000e+01  1.800000e+01  1.800000e+01  1.800000e+01
mean   2013.500000  9.918622e+05  2.780046e+06  4.630639e+06  1.239057e+06
std       5.338539  9.597302e+05  2.611765e+06  4.240914e+06  1.199770e+06
min    2005.000000  2.787700e+05  8.217600e+05  1.448730e+06  3.507500e+05
25%    2009.250000  3.229100e+05  9.537175e+05  1.648340e+06  4.003625e+05
50%    2013.500000  3.441500e+05  1.046845e+06  1.866640e+06  4.401600e+05
75%    2017.750000  2.266282e+06  6.279908e+06  1.035766e+07  2.808392e+06
max    2022.000000  2.395380e+06  6.515170e+06  1.059597e+07  3.016530e+06

           5 or more  All dwellings
count      18.000000   1.800000e+01
mean   286261.666667   9.940489e+06
std    281277.937724   9.283916e+06
min     80820.000000   2.996480e+06
25%     91047.500000   3.433858e+06
50%     94880.000000   3.785270e+06
75%    652902.500000   2.237872e+07
max    706270.000000   2.322933e+07

Missing Values Count:
Year            0
1               0
2               0
3               0
4               0
5 or more       0
Unknown         0
All dwellings   0
dtype: int64

Unique Values Count:
Unknown: 8 unique values

Data Types:
Year              int64
1                 int64
2                 int64
3                 int64
4                 int64
```

Figure 3:Code Snippet 1 on Data Quality Assessment Function

Appendix D: Exploratory Data Analysis (EDA)

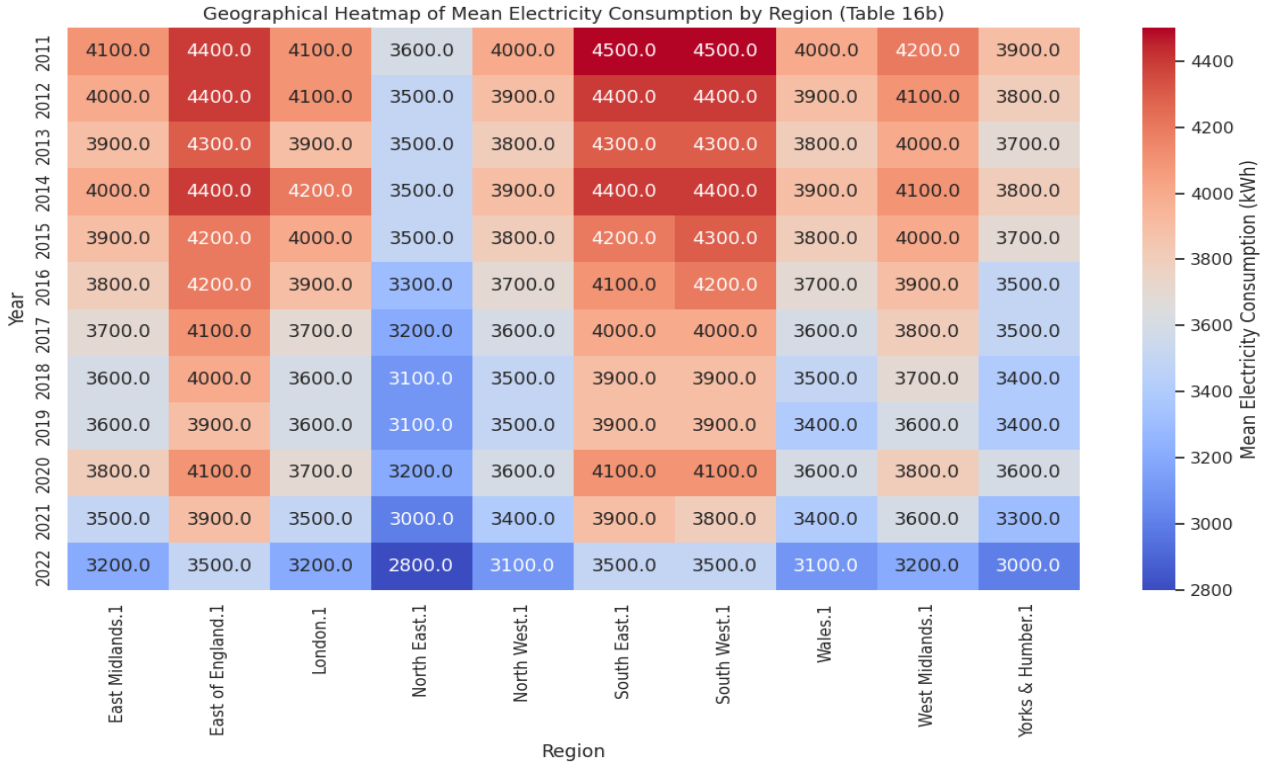


Figure 4:Geographical heatmap showing mean electricity consumption by region using Table 16b.

Appendix E: Link to GitHub Repository

- GitHubhttps://github.com/Anniemuthoni6/Big-Data-Analytics-2024
- Google Colab:https://colab.research.google.com/gist/Anniemuthoni6/2cca301203a80827ca2d7476a3482089/cw2.ipynb
- Data Frame: National Energy Efficiency Data-Framework (NEED): consumption data tables 2024 - GOV.UK