A Data-Driven Approach to Understanding User Behaviour for Energy Management System

Athmika Hebbar

Master of Science
School of Informatics
University of Edinburgh
2023

Abstract

With the growing use of technology and increasing demand for energy, understanding electricity usage patterns has become crucial for effective energy management. This project aims to focus on the comprehensive analysis of the IDEAL household energy dataset combined with external data sources to uncover intricate household electricity consumption patterns and derive meaningful insights. These insights are then further used to construct predictive models capable of forecasting the average monthly electricity consumption and its associated costs for households. Through meticulous data exploration, integration, and advanced modeling techniques, this initiative significantly contributes to advancing our understanding of energy usage dynamics. The outcomes of this research empower households with valuable tools to proactively manage their electricity consumption and make informed decisions for a more sustainable future.

Declaration

I hereby declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Acknowledgements

I would like to express my gratitude to my supervisor, Nigel Goddard, who provided me with much mentorship and technical insights throughout my time working on this project; to Jonathan Kilgour and and Lynda Webb who assisted me during certain tasks very relevant and significant to the project and to the staff members of the MSc Computer Science course, who equipped me with many of the skills that proved to be very helpful in my project.

Table of Contents

Chapter 1

Introduction

1.1 Household Energy Use and Management

The rapid pace of urbanization and growing concerns about environmental sustainability have intensified the focus on energy consumption patterns in residential households. Electricity usage is a crucial component of modern society, with almost every aspect of our daily lives relying on it. However, as technology continues to advance and households become more complex with the use of smart home devices, electric vehicles, solar energy, etc., the need to understand the factors that influence household electricity usage and costs is rapidly increasing. In this context, the availability of rich historical datasets like the IDEAL dataset[10] and REFIT dataset[21] provides an unprecedented opportunity to delve into the intricate dynamics of energy use, unraveling hidden relationships that hold the key to informed decision-making and sustainable living.

The broad foundations of this study are centered around the various factors that influence energy consumption and the challenges of managing electricity demand. Energy consumption is influenced by a wide range of factors such as user behavior, energy tariffs, carbon emissions etc. Understanding the influence of these factors on energy consumption is critical to developing effective energy management strategies. By understanding the various factors that influence energy consumption and the challenges of managing electricity demand, this project embarks on a comprehensive exploration of the IDEAL historical dataset[10], a valuable repository of household energy-related information, integrates it with tariffs and carbon intensity data with the aim of extracting meaningful insights. Additionally, it also seeks to develop a predictive machine learning tool to predict a household's average monthly consumption and the cost of electricity using current tariffs to further get insights on the features that drive the average consumption of a household and to find out what has changed in the past few years, especially with

the advent and increasing use of smart home automation, electric vehicles and other advancements.

The proposed project seeks to achieve this goal by gathering user constraints and leveraging the IDEAL energy dataset[10] to identify patterns and insights that can help users manage their energy consumption more effectively. The project's scope includes gathering user requirements and constraints, leveraging relevant historical data from the IDEAL dataset, analyzing the data through data mining techniques, integrating the carbon intensity API[14] with it and visualizing the results in a user-friendly way through different plots and additionally training a model that uses these features to predict consumption. The machine learning tool as well as the visualisations can be viewed through a user interface in a web application.

1.2 Rationale and Significance

The significance of the study lies in its innovative approach to understanding electricity consumption patterns and its potential to provide valuable insights to users which can be further used to give recommendations to users on how to save energy, reduce costs, minimize carbon emissions or a combination of these based on their energy goals. The proposed research builds on current state-of-the-art data mining techniques[21][7], integrating external data sources with user requirements to generate customised insights that are both informative and actionable. The project also highlights the importance of gathering user constraints and requirements, making it a user-centric approach to energy savings.

The outcomes of this project can have immediate applications for individuals and businesses looking to leverage the insights and patterns identified and also to reduce their energy consumption and costs, as well as for policymakers interested in promoting sustainable energy practices[3]. The insights gained from this study can help explore the impact of different factors and household practices on energy consumption and the effectiveness of different energy-saving strategies.

Chapter 2

Background

The use of data-driven approaches for energy management has gained significant attention in recent years. Previous studies have shown that energy consumption patterns can be learned from historical data and usage patterns. For example, studies by Gajowniczek et al.(2015)[9] and Ashouri et al.(2018)[1] used a combination of data mining techniques and user feedback to identify energy-saving opportunities in households. Similarly, a study by Tureczek et al. (2018)[9] used data from smart meters to analyze energy consumption patterns and identify opportunities for energy conservation. Another study by Silva et al.[3] proposed a demand response mechanism that takes into account user preferences and real-time electricity prices to optimize energy consumption.

Energy management tools have been developed to help individuals and organizations monitor and reduce their energy consumption. For example, the Energy Star Portfolio Manager[5] is a tool developed mainly for Canadian buildings that allows building owners and managers to track and manage their energy usage. Data mining techniques have also been applied to energy consumption data to identify patterns and develop predictive models[9]. In particular, time series analysis[12] and clustering algorithms[18] have been used to identify recurring patterns in energy consumption data. These techniques have been used to develop energy forecasting models and to identify energy-saving opportunities.

Some studies have used the IDEAL dataset to investigate different aspects of energy consumption and demand management. For example, some studies used the dataset to develop a model for predicting household electricity consumption based on weather conditions and household characteristics[8][15][16]. Another study used the dataset for similar customer identification based on metadata of household background[15]. In addition, the IDEAL dataset has also been used in studies related to demand response, energy forecasting, and appliance usage analysis. These studies demonstrate the potential of the IDEAL dataset in providing valuable

insights into household energy consumption patterns and supporting the development of advanced energy management systems.

Also, several advanced models have been proposed which use deep learning techniques for electricity prediction. For instance, the study by Kim et. al.[16] used the Electricity Load Diagrams Dataset from UCI Machine Learning repository[28] which has data from 370 clients and some of them belong to years as old as 2011. The current project, although makes use of simple machine learning algorithms but is based on the IDEAL dataset which has more recent data and also has metadata related to the households, rooms, appliances, tariffs. Thus, this serves as one of the useful primary studies that employ machine learning to build predictive models using the IDEAL dataset and is tailored mainly for understanding the energy consumption patterns in Scotland although it can also be used to make comparisons with other models and other studies which are based on data collected from other geographical locations. Our proposed study plans to build on some of these work by gathering survey data from users to know more about what insights will be useful to them. Most of these studies assume generic requirements and energy-goals such as reducing energy costs. However, different classes of users might want insights based on particular needs like appliance distribution analysis, patterns in carbon emissions, monthly costs, electrcity usage trends based on family size (number of residents), etc. The IDEAL dataset that will be used in this study is unique in itself and there aren't too many researches that have leveraged this dataset. This study not only uses the IDEAL dataset but also incorporates additional data from user constraints and external factors so to perform data mining tasks to give insights to different classes of users and for companies and researchers working in this field who can leverage these insights and patterns.

Chapter 3

Data and User Requirements

3.1 Historical Data - IDEAL dataset

The IDEAL dataset serves as a valuable resource for energy management research, offering a comprehensive collection of historical data related to household energy consumption. Accessed through the IdealDataInterface and IdealMetaDataInterface API[11], this dataset incorporates structured data stored in CSV files, encompassing various aspects of energy usage, home characteristics, meter readings, tariff information, sensor readings, room data, etc. As an extensive compilation of anonymised data from diverse households in Midlothian, Fife and East Lothian, the IDEAL dataset provides a rich and diverse source for in-depth historical data analysis. In this project, the relevant data that has been used includes home characteristics, meter readings which has multiple meter readings for a single household taken during different dates, tariff information which specifies the daily standing charge pence and the unit charge per kWh usage of electricity, appliance data, and room data.

The dataset's richness lies in the diverse and abundant information it offers. One of the key strengths of the IDEAL Energy dataset is its real-world relevance. The data is derived from actual households, making it highly representative of real-life energy consumption behavior. This authenticity ensures that the insights obtained from the dataset are applicable and actionable in practical energy management scenarios.

It includes data not just on household consumption through sensor readings but also metadata like appliance usage, household demographics like family size and income band, electricity meter readings along with the date they were taken, and tariff details. This wide range of data enables researchers to analyze multiple dimensions of energy usage patterns and identify significant factors influencing energy consumption in households. Due to its historical nature, the IDEAL Energy dataset supports longitudinal analysis over extended periods. Researchers

can track energy consumption trends over time, identify seasonal variations, and assess the impact of external factors on energy usage. Researchers can draw meaningful conclusions and make practical recommendations based on the dataset's authentic and diverse samples. By mining the dataset, we can identify energy-intensive appliances and equipment in households. This information can be utilized to devise energy-saving recommendations and promote the adoption of energy-efficient technologies.

3.2 Gathering User Requirements through Survey

A survey was conducted targeting a small number of students and professionals living in and around Edinburgh, with diverse housing arrangements, to gather user requirements on data visualizations and metrics. The survey encompassed various aspects of energy usage and preferences to tailor the system to their specific needs. The survey questionnaires and responses have been included in Appendix A.

Participants were asked to rank the importance of reducing energy costs, increasing the use of renewable energy resources, and reducing carbon emissions. Both groups ranked using renewable energy resources as their top priority, followed by reducing energy costs and then reducing carbon emissions. This insight underscores the participants' shared commitment to sustainable practices while being mindful of cost-saving measures.

The survey also probed participants about their usage of electric vehicles, home energy storage systems, and solar panels. These responses provided valuable data to understand the potential integration of electric vehicle charging patterns, energy storage utilization, and solar energy generation into the energy management system.

To enhance the system's weather forecast-related metrics, participants were asked about their preferences for metrics to schedule appliances or manage electricity use better. Responses revealed their interest in real-time weather forecasts and how they can impact energy consumption patterns. Incorporating these metrics will enable users to make informed decisions about optimizing appliance use based on weather conditions.

Understanding the participants' priorities on energy usage trends, the survey asked them to select relevant variables and metrics. Seasonal energy consumption and total carbon emissions emerged as the most critical factors for users, indicating their concern for understanding their environmental impact and aligning their energy usage with seasonal variations.

Energy efficiency metrics and benchmarks were also explored. Participants expressed an interest in tracking energy usage per square foot and energy efficiency ratings, highlighting their focus on maximizing energy efficiency in their respective living spaces.

To ensure the energy management system caters to a wide range of users, the survey sought preferences on visualizations formats and tools. Participants favored bar charts, line charts, and scatter plots for analyzing energy usage trends data. Their interest in heat maps and histograms showcased the desire for in-depth insights and understanding energy consumption patterns from different angles.

Moreover, the survey invited participants to share any additional visualizations they find useful. Responses included energy consumption breakdowns by appliance type, energy usage comparison with similar households, and energy-saving tips tailored to their specific living arrangements.

Incorporating the insights from this comprehensive survey, the energy management system will be tailored to address users' specific preferences and needs. By offering a wide range of visualizations and metrics, and catering to factors like renewable energy usage, energy efficiency, and weather conditions, the system will empower users in Edinburgh to make informed decisions about their energy consumption, contribute to sustainability efforts, and optimize their energy costs.

3.3 Other Data - Carbon Intensity, Tariffs

One of the components in this project involves the utilization of the Carbon Intensity API[18] as a critical data source to compute the average monthly carbon emissions associated with household energy consumption. This integration serves as a dynamic bridge between the home data, encompassing crucial details about energy consumption patterns, and the real-time carbon intensity measurements provided by the API. The process involved making API calls with the specific dates as periods for which energy consumption was recorded so that the API returns the actual carbon intensity between the period specified. Thus, this offered a granular understanding of the environmental impact at different times. Leveraging this fine-grained carbon intensity data, the total carbon emissions was calculated for each household by multiplying the average monthly consumption with the respective carbon intensity, providing a comprehensive assessment of the carbon footprint associated with their energy use. By quantifying the carbon emissions for each household and translating them into an average monthly metric, users can gain a clear and tangible indicator of their environmental contribution.

Simultaneously, for computing the average monthly costs after predicting the consumption, up-to-date tariff information like the daily standing charge (in pence) and the unit charge per kWh (in pence) directly sourced from Octopus Energy's website[5] provided real-time economic context to the analysis.

In the initial dataset, crucial tariff information such as the daily standing charge in pence and the unit charge per kilowatt-hour (kWh) played a pivotal role in various analyses, including assessing the relationship between carbon emissions and cost. However, for the specific task of predicting current costs, it became evident that relying solely on these historical tariffs would not provide accurate estimations. This realization prompted the acquisition of up-to-date tariff data from the Octopus Agile website, ensuring that the predictions align with the current pricing structure.

By integrating the fresh tariff data from the Octopus Agile website with the consumption results generated by the predictive model, we were able to devise a comprehensive approach for predicting the current costs of electricity consumption. The consumption predictions, reflecting the energy usage patterns of households, were combined with the real-time tariff information, which included the latest daily standing charge and the unit charge per kWh. This dynamic integration enabled the accurate estimation of current costs, considering the precise pricing structure in place at the time of the prediction.

This approach ensures that the cost predictions are not anchored to outdated tariff values but are instead informed by the most current and relevant pricing information available.

Chapter 4

Data Mining and Generating Insights

4.1 Exploring the IDEAL dataset and APIs

The exploration of the IDEAL Dataset, coupled with its corresponding MetaDataInterface API, stands as a consequential pursuit to unveil intricate insights from the household energy consumption data. The IDEAL Dataset encapsulates a complex network of structured data, encompassing diverse dimensions, including appliances, homes, meter readings, tariffs, and other salient aspects. To navigate this multifaceted data effectively, the IdealMetaDataInterface API emerged very useful, facilitating seamless access to data distributed across multiple CSV files.

The dataset is rich in information that researchers, analysts, and developers working on energy management can use to understand trends, detect patterns, and make better decisions. The dataset covered many aspects of energy use, from individual appliances, sensors in rooms within the households, to overall household characteristics, giving a complete understanding of how energy is used.

The IdealMetaDataInterface API returned the data in the form of pandas' dataframes which are versatile data structures that facilitate effortless exploration, transformation, and analysis of the metadata. This not only streamlined the data pre-processing pipeline but also offered a flexible environment for conducting in-depth examinations of the metadata's intricacies.

The MetaData explored and analysed to generate the insights are tabulated in table 2.1.

MetaData	Description	Columns
Home	Contains data pertaining to the 255 households like number of residents, income band, etc	home_id, residents, income_band, energytpe
Tariff	Encompasses information about the pricing structure for home electricity and gas supply	unit_charge_pence_per_kwh, daily_standing_charge_pence, home_id
Appliance	Contains data about large, high-power and generally fixed or rarely moved appliances within the home	appliance_id, home_id, powertype, number
Meter Reading	Contains the meter readings used to measure energy consumption and related data for home electricity and gas meters.	home_id, date, provenancedetail, energytype, reading
Room	Contains all the essential information related to all the rooms within the households	room_id, home_id

Table 2.1

Each of these metadata was initially stored into separate dataframes before transforming and integrating into bigger dataframes that encompassed all of this data together without any missing data or other discrepancies.

4.2 Data transformation and Integration

4.2.1 Handling Missing Data

Handling missing data is a critical aspect of data analysis, and in the IDEAL Energy dataset, missing values were addressed systematically across various metadata. For houses where monthly consumption could not be calculated due to unavailable meter readings or other reasons, a robust imputation approach was adopted. The home data was grouped based on income bands, and the missing values were replaced by the average consumption for the specific income band. This method ensured that the imputed values were consistent with the general

consumption patterns within each income group, minimizing bias and preserving the overall data integrity.

Similarly, missing values were observed for tariff-related information, specifically the daily standing charge in pence and the unit charge in pence per kilowatt-hour (kWh). To handle these missing values, the decision was made to replace them with overall average values. By utilizing the overall average, the imputed values remained representative of the dataset's tariff characteristics without introducing undue influence from specific outliers.

However, handling missing values for carbon intensities presented a unique challenge. In certain cases, carbon intensities could not be retrieved for specific dates, primarily due to data availability limitations within the API. These gaps were often observed for durations exceeding the allowed 14-day period. In such situations, it was determined that replacing the missing carbon intensities with average values could potentially introduce inaccuracies and skew the analysis. As a result, a more conservative approach was taken, and rows with missing carbon intensity data were dropped from the dataframe entirely. By removing these rows, the analysis remained focused on the actual carbon intensity values, ensuring precision in environmental impact assessments.

Overall, the approach to handling missing data in the IDEAL Energy dataset was thoughtful and tailored to the characteristics of each metadata category. The imputation techniques for monthly consumption based on income bands maintained a realistic representation of energy usage patterns for different income groups. Similarly, replacing missing tariff values with overall averages ensured the tariff data's continuity without distorting its overall distribution. However, when dealing with carbon intensity data, an exclusionary strategy was employed, as retaining actual values was deemed crucial for accurate environmental impact analysis. By implementing these techniques, the IDEAL Energy dataset remains a reliable and comprehensive resource for energy management research, facilitating precise insights and informed decision-making for a sustainable and efficient energy future.

4.2.2 Handling Categorical Variables

In the context of the IDEAL dataset, the categorical variable "income band" represents different income levels of households. Initially, the income band variable had 15 distinct categories, each corresponding to a specific income range. However, to simplify the analysis and reduce the dimensionality of the data, the income band variable was reduced to just 5 broader categories.

The reduction of income band categories was likely based on grouping households with similar income levels together. For example, the 15 original categories might have been merged

into broader ranges to create the following 5 categories: Low-Income, Lower-Middle-Income, Middle-Income, Upper-Middle-Income and High-Income.

Each category encompasses a range of income levels, making it easier to handle and interpret the data. The reduction to 5 categories allows for a more concise representation of income distribution in the dataset while still capturing the main income variations across households.

After reducing the income band variable to these 5 categories, ordinal encoding[15] was applied. Ordinal encoding is suitable when there is a natural order or hierarchy among the categories. In this case, the income band categories were assigned ordinal integer values from 1 to 5, where the lowest-income category (e.g., Low-Income) was assigned the value 1, and the highest-income category (e.g., High-Income) was assigned the value 5.

The ordinal encoding of the income band variable converts the categorical values into numerical representations while preserving the inherent order among the income categories[25]. This numerical representation allows machine learning algorithms to understand the relative relationships among the income levels during analysis and modeling[20].

By reducing the income band categories and applying ordinal encoding, the dataset became more manageable and suitable for various data analysis and machine learning tasks. The resulting ordinal representation of income band enables researchers to analyze the relationship between household income and other variables in the dataset while considering the natural hierarchy among income levels. It also facilitates statistical computations and helps in identifying trends or patterns related to income and energy usage in the IDEAL dataset.

4.3 Integrating different metadata's dataframes

Each of the data (other than the home data) had to be integrated with the home data through the home_id column present in each of the dataframes because home_id served as the unique attribute common in all which linked each of the metadata. Since each metadata served a different purpose, their integration with the home data was done differently depending on what kind of information was needed. They are explained in more detail in the following subsections -

4.3.1 Integrating tariff data with home data

Within the tariff metadata, two pivotal tariff components held significant importance: the daily standing charge and the unit charge per kilowatt-hour (kWh). However, a notable challenge emerged when dealing with certain households where multiple unit charge per kWh values were

recorded, potentially causing data redundancy. To address this, a rationalization process was undertaken, whereby the mean of all recorded unit charge per kWh values for such households was computed. This approach ensured that each household was represented by a single row, effectively consolidating the tariff information. Thereby, homeid column was used to integrate these two columns with the home data so that it had two additional columns now.

4.3.2 Integrating meter readings with home data

The meter reading was used to compute the monthly average consumption (in kWh) for the households. Thus, for households that had two or more meter readings recorded at different dates, the difference in meter readings as well as the number of days between the readings were computed and the average monthly consumption was computed as follows -

If reading_1 and reading_2 are two meter readings associated with a home dataframe and $date_1$ and $date_2$ which are the dates after conversion to the python's datetime library and are the dates on which these readings were recorded, then -

$$monthly\_consumption = |\frac{reading_2 - reading_1}{date_2 - date_1}| * 30$$

4.3.3 Integrating appliance data with home data

The count of total appliances owned and used per household was required. To find this out, initially appliances categorized with a "powertype" of "electric" were chosen from the appliance metadata. This careful filtering narrowed down the scope to appliances primarily utilizing electric power. Subsequently, a crucial aggregation step was implemented, wherein the selected electric appliances were grouped based on "home id." This grouping facilitated the accumulation of the number of electric appliances within each household. By summing up the electric appliances for each home, this methodology provided a comprehensive overview of the prevalence and distribution of electric-powered devices among households. The following python code was used to find the total count of all the appliances per household -

```
sum_appliances = appliance_df.groupby(['homeid'])["number"].sum()
```

4.3.4 Integrating room data with home data

In the room metadata, the essential information sought was the number of rooms in each household. To extract this vital detail, a strategic approach was adopted, involving a grouping operation on the room dataframe based on the "homeid," which serves as a unique identifier for each

household. By grouping the data using this identifier, it became possible to aggregate the room data within each household, resulting in the count of rooms for every individual household.

The following python code represents the grouping to count the number of rooms -

```
room_groupings = room_df.groupby(["homeid"])["homeid"].count()
```

As a result of this grouping and aggregation process, a new column aptly named ”rooms” was introduced within the dataframe. This ”rooms” column stored the calculated counts of rooms, effectively capturing the spatial dimension of each household

4.4 Finding patterns

In the process of generating insights from the IDEAL dataset, several key analyses and methodologies were employed:

1. Average Consumption per Family Size: The home data was grouped by family size into categories such as small, medium, and large families. By aggregating energy consumption data based on family size, researchers studied the average energy usage patterns for each group. This analysis aimed to identify any correlations between family size and energy consumption, helping to understand how household dynamics influence energy usage.

2. Usage of Households by Income Bands: The home data was further grouped by income bands, which were reduced to five categories using ordinal encoding. This analysis allowed researchers to study the energy usage patterns of households belonging to different income levels. Additionally, the number of households in each income band was studied to observe the distribution of energy usage across different income groups. The usage of smart automation within each income band was also examined, providing insights into the adoption of energy-efficient technologies by different income levels.

3. Cluster Analysis: Cluster analysis was conducted based on income bands, the number of residents in each household, and their monthly energy consumption. This methodology aimed to identify distinct groups or clusters of households with similar energy usage patterns and demographic characteristics. By clustering households, researchers can gain a deeper understanding of the energy consumption behaviors and demographic profiles associated with each group.

4. Relationship between Carbon Emissions and Monthly Consumption: Researchers analyzed the relationship between carbon emissions and monthly energy consumption. This analysis sought to understand how carbon emissions are influenced by the amount of energy consumed by households. It provides valuable insights into the environmental impact of energy usage and identifies potential areas for energy conservation and carbon footprint reduction.

5. Relationship between Carbon Emissions and Monthly Cost: Monthly cost was calculated using tariff data provided, including the daily standing charge in pence and the unit charge in pence per kilowatt-hour (kWh). This calculated monthly cost was then analyzed in relation to carbon emissions. The goal was to identify any correlations between the cost of energy usage and the resulting carbon emissions, offering insights into the environmental implications of varying energy consumption patterns.

By conducting these analyses, meaningful insights were uncovered regarding energy consumption trends, environmental impact, and the factors influencing energy usage in different household settings. The findings from these analyses can inform policymakers, energy companies, and individuals on energy management strategies, sustainability initiatives, and the importance of energy efficiency in different socioeconomic contexts. Additionally, the insights gained from these methodologies contribute to the development of data-driven energy management systems, promoting responsible and sustainable energy practices.

4.4.1 Consumption pattern analysis

The consumption analysis initially involved computing the monthly average consumption for each household that was explained in section 4.3.2. This data was subsequently dissected from two key perspectives: income bands and family size.

Income band	Salary range	Income group
1	£16,199 and below	Lower Income Group
2	£16,200 to £26,999	Lower Middle Income Group
3	£27,000 to £43,199	Middle Income Group
4	£43,200 to £65,999	Upper Middle Income Group
5	£66,000 and higher	High Income Group

Table 4.1

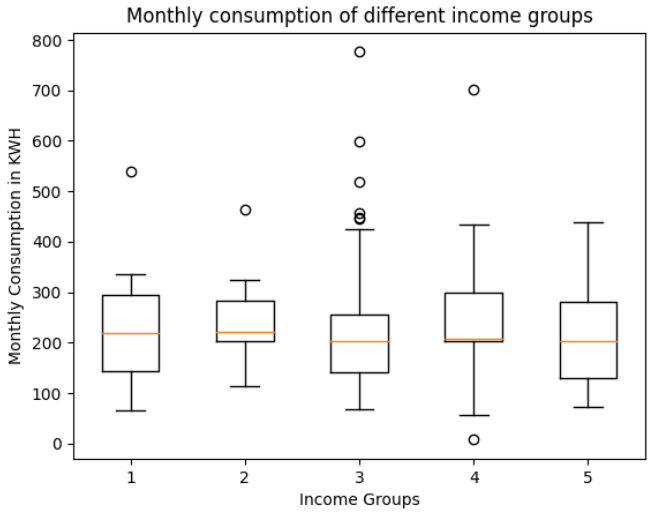


Figure 4.1: Consumption Analysis by income bands 1 - low income group, 2- lower middle, 3 - Middle Income, 4 - Upper Middle Income, 5 - high income

While examining consumption distribution by income bands, intriguing findings emerged. Table 4.1 shows what each of the income bands indicate. Despite the initial anticipation that income bands might be a dominant factor in consumption variation, Figure 4.1 reveals that the average consumption within each income band exhibited only marginal distinctions. This suggests that income band alone is not a defining factor in energy consumption; other variables such as family size, the number of appliances, and the size of the dwelling may have more significant influence.

In contrast, delving into consumption distribution based on family size yielded noteworthy insights. Figure 4.2 clearly demonstrates a direct relationship between family size and average consumption. As family size increased, so did the average energy consumption of the households. This finding underscores the importance of the number of residents as a pivotal feature in determining the energy consumption of a household.

These results indicate that while income bands have limited predictive power regarding energy consumption, family size emerges as a more influential factor. Understanding this relationship empowers households and energy management initiatives to tailor strategies to specific family compositions, potentially optimizing energy usage based on family needs. By incorporating the insights garnered from both the income bands and family size analyses, this project takes a holistic approach, paving the way for more precise energy management strategies that acknowledge the complex interplay of factors impacting consumption behavior.

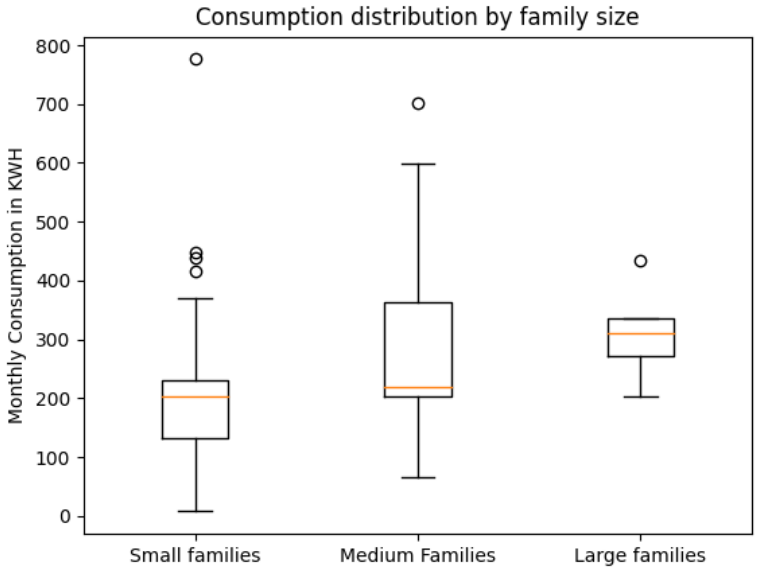


Figure 4.2: Consumption Distribution by Family size

4.4.2 Tariff Analysis

The three major columns useful for tariff analysis are -

1. Monthly cost (average) 2. Unit charge per kWh (in pence) 3. Daily standing charge pence

Among these, the daily standing charge pence was not too useful as it is a fixed component in the tariff which is set according to the tariff plan and does not depend on consumption patterns. Hence, the unit charge and monthly cost were studied. The figure 4.3 shows the relationship between monthly cost and unit charge.

The visual representation of monthly cost plotted against the unit charge per kilowatt-hour (kWh) in pence offers intriguing insights into the relationship between these two variables. The scatter plot reveals a concentration of data points within a specific region, primarily clustered around the range of 60 to 80 GBP for monthly cost and 12 to 15 pence for unit charge per kWh.

This concentration suggests that the majority of households fall within this cost range while experiencing a relatively consistent unit charge per kWh. The trend observed here indicates a potential correlation between the unit charge and monthly cost. As the unit charge per kWh increases, there is a tendency for the monthly cost to rise as well, highlighting the influence of energy consumption on the overall cost.

Of particular interest are the outliers in the plot, which are households exhibiting monthly costs exceeding 100 GBP. These outliers serve as notable deviations from the general trend and warrant further examination. Isolating and analyzing these outliers can provide valuable

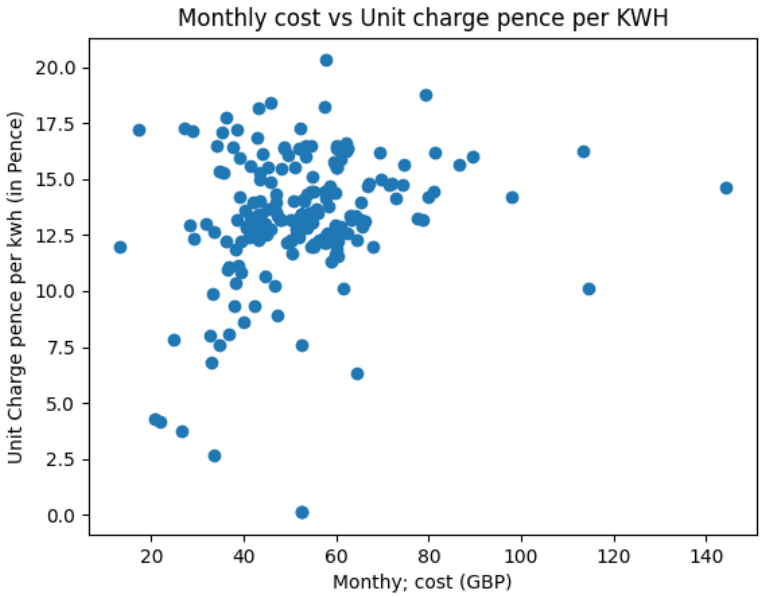


Figure 4.3: Cost vs Unit charge per kWh

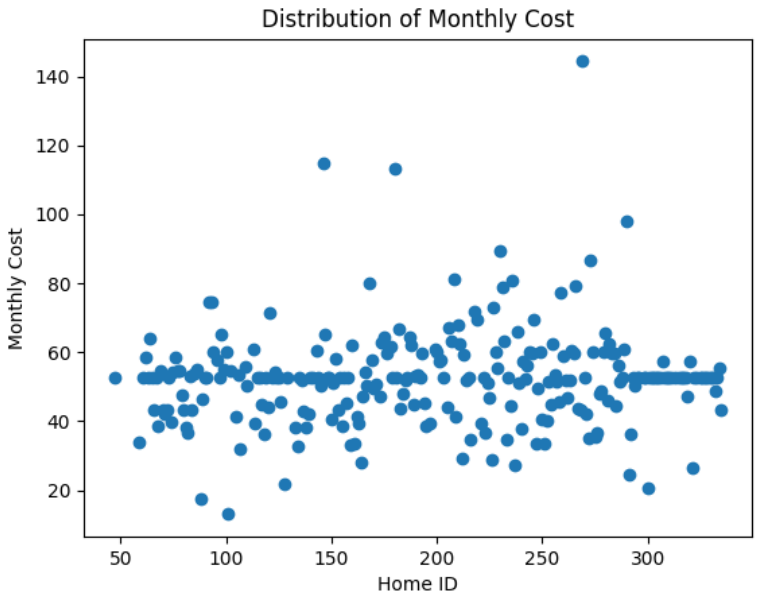


Figure 4.4: Monthly cost distribution of households

homeid	income_band	residents	unit_charge_pence_per_kwh	monthly_consumption	rooms	monthly_cost	appliances
180	3	3	16.250	599.36	13	113.40	13
146	4	4	10.105	701.85	14	114.64	16
269	4	2	14.635	776.36	12	144.38	25

Figure 4.5: Outlier Analysis of high cost households

insights into the unique factors contributing to their elevated energy costs.

Thereafter the outliers having monthly consumption greater than 100 were analysed as shown in Figure 4.7.

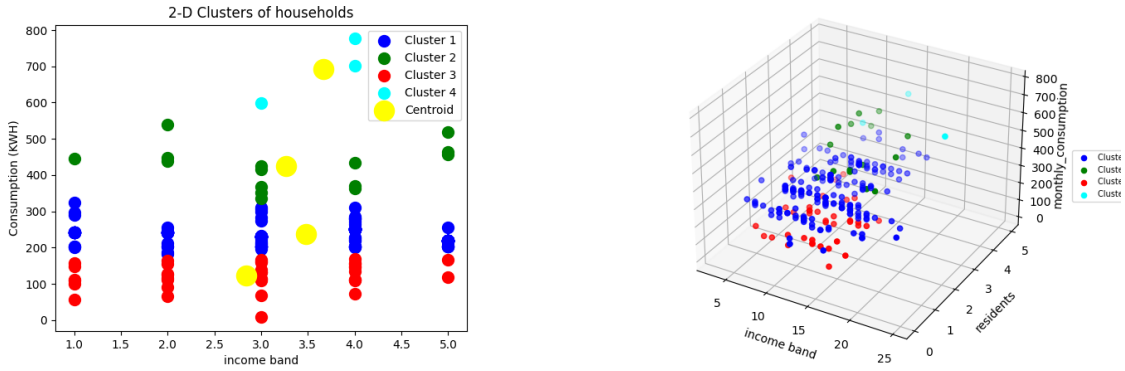
The household having the highest consumption seemed to have just two residents, however this household also has the highest number of appliances and thereby the highest consumption as well, thus this shows a direct correlation between the number of appliances and the monthly cost. The household with homeid 146 seems to have low unit charge but high consumption, 4 residents and more than average appliances(16). Thus this shows that it is important to consider multiple factors while deciding what are the important features while determining the consumption and cost which are done in the later sections in cluster analysis (section 4.4.2) as well as building the predictive model for predicting consumption and cost (chapter 5).

4.4.3 Cluster Analysis

An iterative clustering approach was adopted to identify clusters of households with each cluster representing a different energy use characteristics. Python's Scikit-learn library's K-means clustering algorithm[20] was applied where different columns were used as variables with the monthly consumption being the major variable that determined the cluster for a household among other variables.

There have been past attempts to perform clustering and cluster analysis to understand distinct household energy usage behaviours. A study by Tureczek et. al.[29] collected data from a sample of residential users equipped with smart meters, measuring their electricity consumption at fine-grained intervals. The patterns analysed were characterized by differences in peak load periods, overall consumption levels, and variations in energy usage throughout the day and week.

Initially the monthly consumption, income band and the number of residents were used as the feature variables for clustering. The clusters are represented in Figure 4.6. The median of each of these clusters, although belonging to different consumption levels, did not really



(a) (b)

Figure 4.6: Cluster Analysis

belong to different income levels. However, all of them were neither from the same income level too. So it is hard to say that income levels played a major role in the consumption pattern, even though they did to a small extent. The number of residents however indicated that higher residents always implied higher consumption but lower family size could have consumption values in a wide range.

In response to the above findings, the number of appliances owned by each household was introduced as an alternative variable to income levels. The 2-d cluster is shown in Figure 4.7. This alteration in feature variables provided more insightful results. Notably, the median consumption levels within each cluster were now more distinctly distributed and tied to the number of appliances. This change enhanced the clarity of the clusters' characteristics.

In detail, clusters 1 and 3 displayed households with relatively lower consumption levels, while clusters 2 and 4 represented households with higher consumption patterns. Cluster 4, in particular, stood out as it encapsulated households with the highest number of appliances, residents, and consumption levels. This shift in approach, replacing income levels with the number of appliances, resulted in more meaningful and coherent clustering outcomes.

The significance of this evolution in the clustering process lies in the identification of a more influential variable, the number of appliances, in determining consumption patterns. This adjustment highlighted the role of appliance usage in shaping households' energy consumption behaviors. By leveraging the number of appliances as a key feature in clustering, the project effectively unveiled deeper insights into the intricate relationship between appliance ownership, consumption levels, and potentially other underlying factors influencing household energy usage.

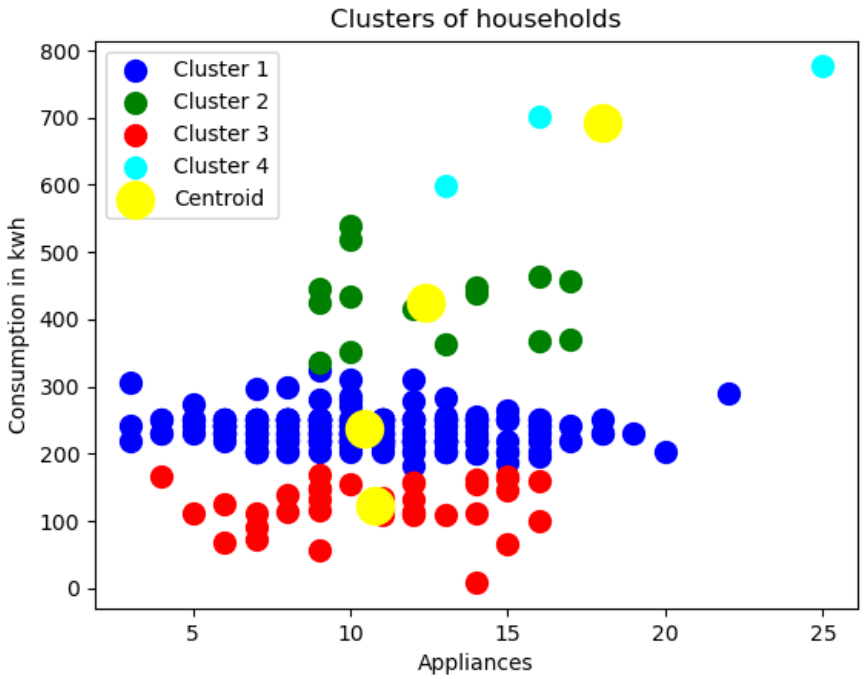


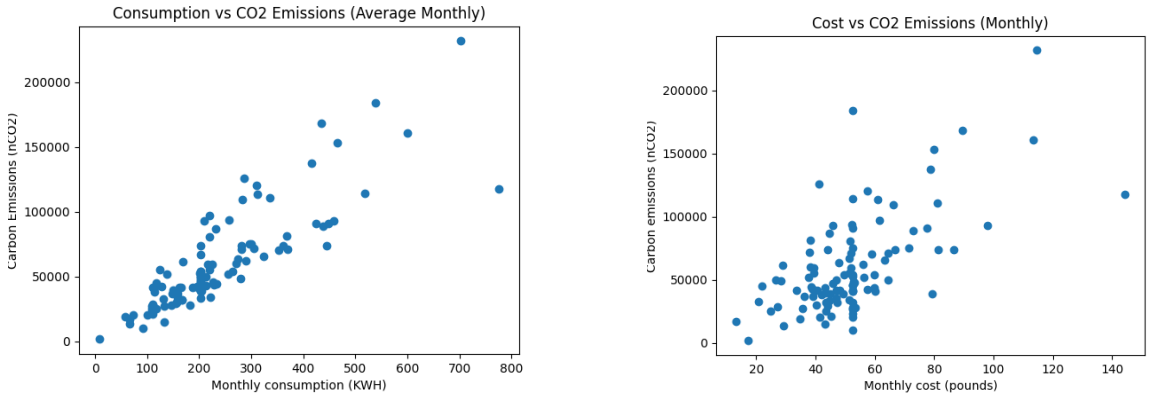
Figure 4.7: 2-D cluster where income band is replaced with Appliances

4.4.4 Carbon Emissions and its relation with consumption and cost

The scatter plot (Figure 4.8 (a)) depicting the relationship between monthly consumption and carbon emissions (measured in nCO2) reveals an interesting trend. As monthly consumption increases, there is a discernible upward trend in carbon emissions, suggesting a positive correlation between these two variables. This linear increase signifies that households with higher energy consumption tend to produce higher levels of carbon emissions, which aligns with our understanding of the environmental impact of energy use.

A notable concentration of data points is observed within a specific range: monthly consumption values ranging from approximately 100 to 300 kilowatt-hours (kWh) correspond to carbon emissions around 50,000 nCO2. This cluster of data points suggests a significant grouping of households with similar energy consumption levels and a consistent emission rate. The fact that a substantial portion of data points aligns closely within this range indicates a consistent carbon emissions pattern across these households.

The dense concentration of data points in this particular consumption-emission range may suggest several implications. It could indicate a common energy usage behavior prevalent among a certain subset of households, which might be influenced by factors such as similar appliance usage patterns, household size, or specific geographic factors. Furthermore, this



(a) Consumption vs Carbon Emissions

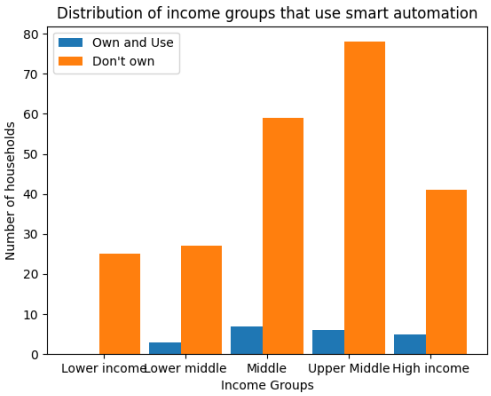
(b) Cost vs Carbon Emissions

Figure 4.8: Carbon Emissions Analysis

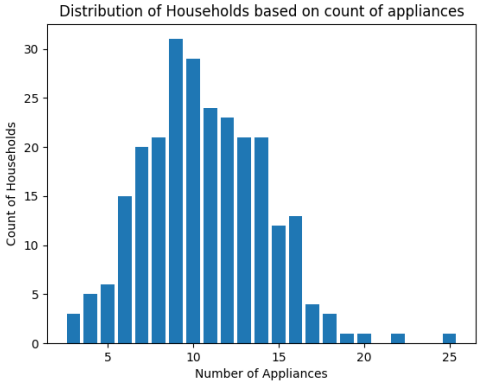
concentration within a specific carbon emissions range could be indicative of consistent energy efficiency practices or technology adoption within this subset.

The cost vs carbon emissions graph (Figure 4.8 (b)) also suggests a linear relationship between these two variables. The majority of data points are situated below 80 GBP for cost and below 100,000 nCO2 for carbon emissions, indicating that there is a correlation between lower costs and relatively lower carbon emissions. As the data pertains to the historical context of 2018, it's essential to recognize that the cost values likely reflect electricity rates prevalent at that time, which are expected to be lower than current rates.

While the historical cost data provides valuable insights into past energy expenses, a compelling aspect lies in the exploration of carbon emissions over time. Given the evolution of energy awareness and potential shifts in energy consumption patterns among households in Scotland, it's intriguing to assess whether carbon emissions have remained stable, increased, or decreased since 2018. The contemporary energy landscape may have witnessed advancements in renewable energy adoption, improved energy efficiency practices, and a growing focus on sustainability. Consequently, there's a possibility that households have become more energy-conscious, potentially leading to reduced carbon emissions despite changing costs. Hence this insight can serve useful for researchers who are collecting current household data to compare with the previous in order to determine what has changed over time. Such an evaluation holds significance not only for energy management strategies but also for understanding the broader environmental implications of evolving energy practices. The potential to observe a divergence between the linear cost-emissions relationship and the actual carbon emissions trend in the present day adds a layer of complexity and underscores the importance of ongoing data analysis in the context of energy conservation and environmental consciousness.



(a) smart devices distirbution



(b) appliance distribution

Figure 4.9

4.4.5 Other Insights

This includes two particular insights that were generated -

1. Distribution of smart home appliances by income band
2. Distribution of count of appliances across households

The examination of the distribution of smart home appliance usage across different income bands in the IDEAL dataset reveals a compelling disparity. The graph in Figure 4.9(a) distinctly highlights a scarcity of households adopting smart home devices, particularly within the lower income brackets. The orange bars indicate the number of household which do not use any smart devices while the blue bars indicate the number of households that own and use smart devices. A striking observation is that none of the families categorized in the lower income group seem to have embraced smart devices during the data collection period. This finding also underscores the influential role of economic factors, symbolized by income levels, in determining the adoption of advanced technological solutions like smart home devices. This is mainly because the data is from the year 2018 and there wasn't much awareness about using smart devices back then and also there weren't many smart devices available in the stores during that time. it is a good sign that smart home devices and smart home automation systems are becoming more prevalent these days and current data would definitely show the start difference in the increasing number of households that use them.

The distribution analysis of the number of appliances per household (Figure 4.9 (b)) reveals an interesting trend reminiscent of a Gaussian-like distribution[17]. The peak of the distribution is observed around 10 appliances, indicating that a substantial number of households utilize this range of devices. However, as the number of appliances surpasses this peak, the frequency of

households with such high appliance counts begins to decline gradually.

An intriguing aspect emerges when considering the presence of an outlier, a household recorded with the highest number of appliances at 25. This outlier stands out within the dataset due to its significantly elevated number of appliances compared to the majority of households. The presence of such an outlier provides a unique opportunity for further exploration and analysis.

Studying this outlier household, which deviates from the typical distribution, can yield valuable insights. By examining the reasons behind the exceptionally high number of appliances in this specific household, we can uncover patterns, usage behaviors, and potentially identify underlying factors driving such a distinctive profile. It's essential to understand whether this household's appliance count is the result of specific requirements, lifestyle choices, or other unique circumstances.

Furthermore, exploring the characteristics of this outlier household might help distinguish between factors that significantly impact energy consumption patterns. By contrasting this outlier with the broader distribution, we can assess the potential implications of such high appliance usage on energy consumption, costs, and environmental impact. This analysis could contribute to a deeper understanding of the relationship between the number of appliances and other variables within the dataset, such as monthly consumption, cost, or carbon emissions.

The presence of this outlier serves as a valuable case study, offering insights that go beyond the general distribution pattern. By delving into the unique features of this household, we can expand our comprehension of the diverse factors influencing energy consumption and inform strategies for energy management and efficiency, even in scenarios that deviate from the norm.

Chapter 5

Predicting monthly household consumption and cost

5.1 The need of a machine learning predictive tool

The incorporation of a machine learning tool was crucial due to the intricacies inherent in household energy consumption. These patterns are influenced by a myriad of variables like user behaviors, appliance usage, etc. Traditional methods often struggle to capture these nonlinear relationships and dynamic dependencies comprehensively. Machine learning, with its ability to handle complex, multidimensional data, emerged as an ideal solution to unravel the intricate connections driving consumption trends.

Predictive capability stands as a pivotal advantage. By training the machine learning model on historical consumption data, it learned underlying patterns, enabling accurate forecasts of future energy usage. This predictive power would empowered users with valuable foresight, helping in proactive energy management.

Furthermore, the machine learning tool would also allow for the identification of influential factors. By analyzing feature importance, the model can reveal the key drivers impacting energy consumption. This knowledge provides a deeper understanding of the intricate dynamics at play, helping users pinpoint areas for energy-saving strategies, promote sustainable practices, and make informed choices about appliances, usage patterns, and resource allocation. The adaptability of these models ensures their relevance across diverse scenarios, effectively bridging the gap between observed consumption data and the actionable insights needed for effective energy management. This approach is aimed at shifting the focus from mere predictions to insightful understanding, mainly about what drives energy usage, leading to more informed energy management decisions, cost optimization, and sustainable consumption practices.

5.2 Choosing the algorithms for model building

Choosing the right algorithms for energy consumption prediction is pivotal, with Random Forest[26] and Multiple Linear Regression[30] emerging as top contenders. The selection decision hinges on their distinctive strengths aligned with the intricacies of the use case. Random Forest's ensemble nature and ability to capture non-linear relationships suit the complexity of energy consumption prediction, offering insights into variables' importance and interactions. This is particularly crucial for deciphering multi-faceted energy usage patterns and understanding the significance of different factors.

Conversely, Linear Regression's simplicity and interpretability rendered it an excellent choice. The algorithm's transparency, along with its emphasis on linear relationships, aligns well with the project's goal of providing users with clear and actionable insights. Given the nature of energy consumption data, where direct correlations between variables might be present, Linear Regression's capability to unveil direct influences aids in making practical and informed decisions. Furthermore, its computational efficiency and speed hold value in real-time applications. The balance between the sophistication of Random Forest and the intuitive nature of Linear Regression empowers the project to offer accurate and accessible energy consumption predictions to users.

5.3 Predicting consumption with 4 features (including income band)

In the process of developing a prediction model for energy consumption using four variables, namely the Number of Residents, Number of Rooms, Number of Appliances, and Income Band, two machine learning algorithms were employed: Random Forest Regression and Linear Regression for reasons stated in the previous section. Both of these methods aimed to uncover relationships between a set of four carefully chosen predictor variables and the energy consumption output variable. The dataset was divided into training and testing sets to evaluate the model's performance accurately.

During training, the algorithms learned the relationships between the input variables and the energy consumption data from historical records. The evaluation results are depicted in section 7.1.

5.4 Predicting consumption with 3 features (without income band)

Since not all users would be willing to share their income bands to which their households belonged too, it was important to build a model that did not use income band as a feature and only made use of the columns - number of residents, number of appliances and number of rooms as features. Also, the cluster analysis also indicated that income band was not that important of a feature in determining the average monthly consumption of the households. Hence, having another model that did not use income band among the dependent variables was significantly important.

Again for this prediction, two different models were run, one that used random forest regression and the other that used multiple linear regression. The model performance for this is evaluated in the section 7.2 as well.

5.5 Predicting the Average Monthly Cost

Predicting the average monthly cost of electricity is a pivotal aspect of empowering users with comprehensive insights into their energy consumption. The process involves a meticulous interplay of data retrieval, calculation, and integration to provide users with accurate estimates of their energy expenses.

To begin, the foundation of this process lies in the data sourced from the Octopus Energy website. Extracting the daily standing charge in pence and the unit charge in pence per kilowatt-hour (kWh) from Octopus Energy forms a critical step. These values, representing essential components of the tariff structure, lay the groundwork for calculating the monthly cost accurately.

The subsequent step involves the computation of the monthly cost based on the predicted energy consumption. Leveraging the energy consumption in kWh forecasted by the model described in sections 5.1 and 5.2, coupled with the unit charge pence per kWh, the variable cost component of the monthly cost is ascertained. This variable cost encapsulates the actual energy usage of the household, translating into a quantifiable expense.

However, the process doesn't stop at variable costs. To paint a holistic picture of the monthly expenditure, the daily standing charge pence plays a significant role. This fixed charge is independent of energy consumption and is added regardless of how much electricity is used. Consequently, the cumulative daily standing charge pence is fused with the variable cost to

derive the comprehensive total cost of electricity for the month.

The result of this meticulous process is the culmination of both variable and fixed charges, yielding the average monthly cost of electricity.

Average Monthly Cost = Daily Standing Charge Pence (Fixed Component) + (Average Monthly consumption (in kWh) * Unit Charge Pence per kWh)

This figure equips users with insights into not only their energy consumption patterns but also the financial implications of their usage behavior. By presenting users with a well-rounded estimation of their monthly expenses, this approach facilitates informed decision-making, enables the optimization of energy consumption habits, and assists in effectively managing monthly costs.

In essence, the process exemplifies the synergy between data-driven insights and practical financial implications. By considering both consumption and tariff structure, users gain a profound understanding of their energy expenses, enabling them to align their consumption habits with their financial goals.

Chapter 6

System Implementation

6.1 Implementation details - Backend

For the several data mining stages, the code was developed in python using various useful libraries like numpy, pandas for dealing with data as dataframes and matplotlib[21] for displaying the graphical visualisations. All the code related to even the predictive machine learning model for consumption was developed in python. For developing the web application. python's simple and lightweight framework, Flask[12], was used. The final features generated with integrated metadata were stored in features.csv for easy access and faster data manipulation and analyses so that the system does not spend too much time integrating the data from different csv files.

The consolidated data features.csv alone serves as a rich data which includes additional columns for each of the households with the tariff info, carbon emissions, number of appliances, rooms, etc.

The python's library - matplotlib has been used extensively throughout the ode to generate the different kinds of plots like box plots, scatter plot, bar graph, histogram, pie chart, etc. The library proved to be convenient as it also allowed to save the figures generated as PNG images which could be rendered in the User interface through the HTML web pages.

6.2 System Implementation - Frontend

The user interface of the system was developed using HTML, CSS and Javascript which were rendered by the Flask's python backend. The visualisations are displayed on five different web pages, each for displaying consumption analysis, tariff analysis, carbon emissions, clusters and other insights. Additionally, a separate web page for predicting the results of the model

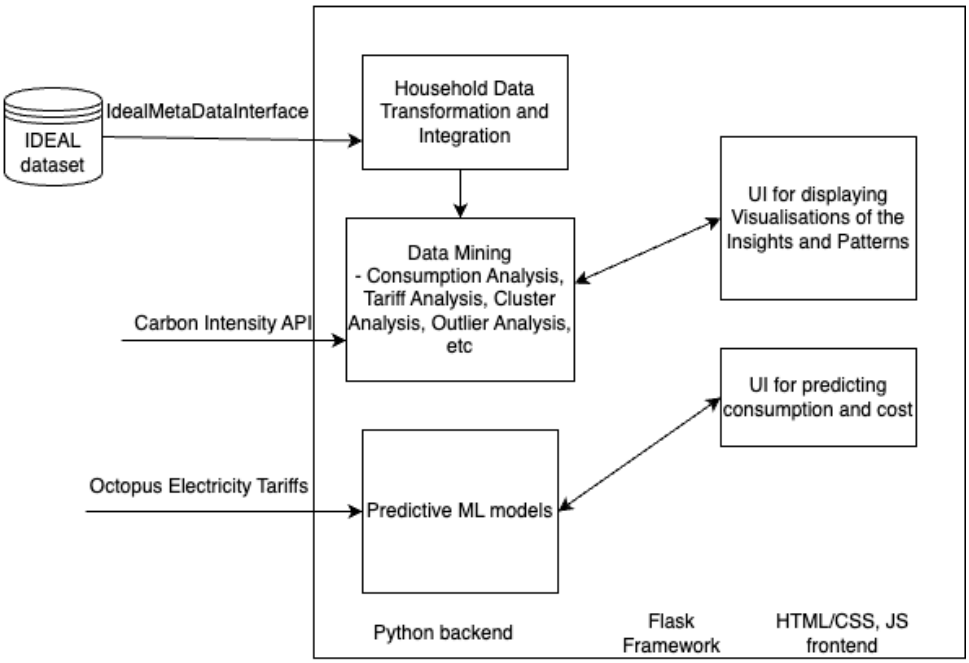


Figure 6.1: System diagram

is present. An html form takes the user input like number of appliances, number of rooms, number of residents and the income band and the button click routes the request to the backend which sends the results of the model to the results page. The navigation bar makes it easy for the user to move to the different web pages to view the visualisations or move to the page where the consumption and costs can be predicted. The UI design screenshots of the system have been added in the Appendix B.

Chapter 7

Evaluation

7.1 Model Performance Analysis

Since both algorithms i.e Random Forest and Linear Regression were applied to the same dataset, their performance could be objectively compared to determine which better captures the dataset's characteristics. The metrics used to evaluate the models' performance are -
1. Mean Absolute Error
MAE measures the average absolute difference between the predicted and actual values. It provides a direct measure of the model's prediction accuracy. Lower MAE values indicate better performance.

2. R2 (R-Sqaured) - Coefficient of Determination
R-squared measures the proportion of the variance in the dependent variable (energy consumption) that is predictable from the independent variables. It ranges from 0 to 1, where 1 indicates perfect predictions and 0 indicates poor predictions. Higher R-squared values suggest a better fit of the model to the data.

The features used in the regression models are -
Na : Number of Appliances that the household owns and uses
Nr : Number of rooms in the house
Np : Number of residents living in the household
IB: Income Band of the Household (1-5)

The Mean Absolute Percentage Error values and the R-square values for each of the models trained (as shown in Table 7.1) indicate that in terms of MAE (percentage), linear regression

Metric	Model	Features	Value
Mean Absolute Error (in percent)	Random Forest	Na, Nr, Np	0.25
	Linear Regression	Na, Nr, Np	0.153
R-Square	Random Forest	Na, Nr, Np	0.569
	Linear Regression	Na, Nr, Np	0.003
Mean Absolute Error (in percent)	Random Forest	Na, Nr, Np, IB	0.209
	Linear Regression	Na, Nr, Np, IB	0.121
R-Square	Random Forest	Na, Nr, Np, IB	0.656
	Linear Regression	Na, Nr, Np, IB	0.033

Table 7.1: Model evaluation results

model performs better. However, in terms of R-square values, the linear regression models have very low values of R-square values indicating that the independent variables in the linear regression model are not effective in explaining the variation in the dependent variable, the MAE values are still good and this is highly because the predicted values are close to the mean value, however the model is not able to rightly identify the dependency of the dependent variables on the independent variable (average monthly consumption) in case of linear regression model. .

The random forest regressor models have comparable Mean absolute errors to their respective linear regression models but have a huge difference in the R-square value (i.e they gave higher R-square values) indicating that random forest is better able to capture the dependency of the features on the independent variable. The R-squared score is particularly valuable as it indicates how well the model explains the variability in the data. In both cases, where three and four features were utilized for prediction, the Random Forest model consistently demonstrated a higher R^2 score compared to other models, indicating its capacity to capture more of the variation in energy consumption. This signifies that the Random Forest model fits the data better, suggesting it is a stronger candidate for predicting consumption patterns. Hence, random forest regression models are chosen in both the cases to predict the monthly average consumption of households.

Figure 7.1 shows the actual vs predicted value graph for the random forest regressor with 4 variables (including income band). It can be easily observed that most of the values lie between 200 and 300 for both actual and predicted ones which is a good sign, but for outliers in the dataset the model finds difficulty in capturing the variations.

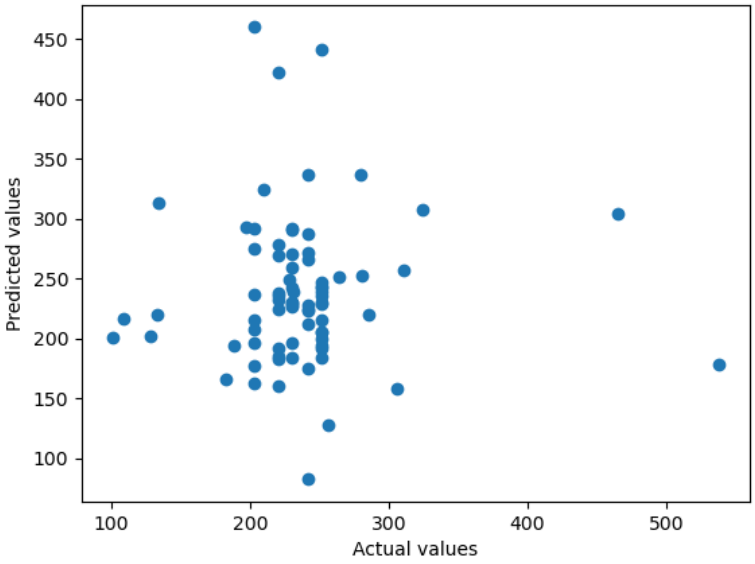


Figure 7.1: Actual vs Predicted Values

7.2 Evaluation of the overall system and UI

Users were asked to evaluate the overall system in terms of the goals and objectives met, about the usefulness of the visualisations presented and were also asked to assess the predictive tool to understand how close it was to their actual consumption and costs and if there were any useful findings which could be derived through this tool to give them more understanding of their consumption and costs and the factors that influence it.

Additionally, two of these, Jonathan Kilgour and Lynda Webb who are researchers in the field of energy informatics at University of Edinburgh were also approached to evaluate the predictive model. Jonathan's house was simple and used a basic non agile tariff plan with no experimental setup. The predicted consumption and cost was quite close the actual values. For Lynda's household, however, which was complex with its use of smart home devices, solar power, electric vehicles and agile tariff plan[5], the consumption predicted was much lower than the actual consumption, with the predicted being 238 KWh and the actual being 632 KWh. Even then, the cost predicted using the basic daily standing charge pence from the Octopus agile website was closer to the actual cost (actual cost was £89 while the predicted was 85.83. This is mainly because of the appliance use time shifting to times of lower tariffs. This was quite useful in understanding the cost savings even though the consumption was high. This also indicated that the historical data could be useful in understanding the energy use patterns in households which used most of the basic electric appliances and had common consumption

patterns like most households. In case of households with sophisticated energy use practices, it can serve as a useful tool in understanding the effects of difference in the electricity usage patterns, especially related to costs.

The user interface's usability, layout and other factors were qualitatively evaluated by Jonathan and about half of the people who were involved in the initial survey of understand user needs. A user satisfaction score of 7 on a scale of 10 was achieved for the overall system.

Chapter 8

Conclusion

This dissertation project delved into the data-driven insights related to the household energy consumption in parts of Scotland leveraging the IDEAL dataset. The journey embarked upon a path of data exploration, visualization, and predictive modeling to unravel the various dynamics that govern energy usage patterns. The visualizations obtained from diverse metadata sources proved to be invaluable insights on their own, offering a window into the interplay of various factors within household energy consumption.

These visualizations, each capturing a facet of household dynamics, stand not only as meaningful findings but also as potential building blocks for further exploration. By integrating these visual insights with external data sources like weather patterns, a comprehensive and integrated understanding of energy consumption patterns can be fostered. This integration promises the potential to discover deeper correlations, providing a more holistic view of energy consumption in the context of diverse environmental factors.

Through the course of this project, it became evident that these insights are not just static snapshots but windows into the evolving energy usage trends. As society shifts towards greater energy consciousness, understanding the underlying factors that contribute to consumption changes over time is crucial.

The predictive model developed within this work, while serving as a foundation for consumption estimation, also sheds light on its limitations. While proficient at providing estimates for basic households with common appliances and straightforward tariff structures, the model struggles with the complexities of modern energy dynamics. Households equipped with sophisticated smart devices, renewable energy sources, and agile tariff plans tend to confound the model's predictions. Nevertheless, the model's performance remains a valuable benchmark, offering insights into the energy consumption of typical households and allowing comparisons that illuminate potential savings or over-expenditures.

In summation, this dissertation's journey through data mining, visualization, and predictive modeling has unveiled insights that are both enlightening and pragmatic. The visualizations stand as testaments to the value of exploring energy data's multifaceted dimensions, while the predictive model serves as a stepping stone towards understanding and optimizing energy consumption. As we peer into the future, the integrated approach of this project forms the foundation for a more comprehensive and dynamic understanding of energy consumption, fostering informed energy choices and contributing to a more sustainable future.

Bibliography

[1] Milad Ashouri, Fariborz Haghighat, Benjamin C.M. Fung, Amine Lazrak, and Hiroshi Yoshino. Development of building energy saving advisory: A data mining approach. *Energy and Buildings*, 172:139–151, 2018.

[2] Hans Auer and Reinhard Haas. On integrating large shares of variable renewables into the electricity system. *Energy*, 115:1592–1601, 2016.

[3] Barry Barton, Sally Blackwell, Gerry Carrington, Rebecca Ford, Rob Lawson, Janet Stephenson, Paul Thorsnes, and John Williams. *Energy cultures: implication for policymakers*. Centre for Sustainability, University of Otago, 2013.

[4] Igor RS da Silva, Ricardo de AL Rabêlo, Joel JPC Rodrigues, Petar Solic, and Arthur Carvalho. A preference-based demand response mechanism for energy management in a microgrid. *Journal of Cleaner Production*, 255:120034, 2020.

[5] Octopus Energy. Agile octopus: A consumer-led shift to a low carbon future, 2018.

[6] STAR ENERGY. Energy star portfolio manager. *Energy Star*, 2019.

[7] Silvia Erba, Francesco Causone, and Roberto Armani. The effect of weather datasets on building energy simulation outputs. *Energy Procedia*, 134:545–554, 2017.

[8] Cheng Fan, Fu Xiao, and Shengwei Wang. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. *Applied Energy*, 127:1–10, 2014.

[9] Pamela Fry. Literature review template. https://www.tru.ca/__shared/assets/Literature_Review_Template3 Thompson Rivers University. Online; accessed 3 September, 2018.

[10] Krzysztof Gajowniczek and Tomasz Zabkowski. Data mining techniques for detecting household characteristics based on smart meter data. *Energies*, 8(7):7407–7427, 2015.

[11] Nigel Goddard, Jonathan Kilgour, Martin Pullinger, D.K Arvind, Heather Lovell, Johanna Moore, David Shipworth, Charles Sutton, Jan Webb, Niklas Berliner, Cillian Brewitt, Myroslava Dzikovska, Edmund Farrow, Elaine Farrow, Janek Mann, Evan Morgan, Lynda Webb, and Mingjun Zhong. IDEAL Household Energy Dataset, April 2021.

[12] Miguel Grinberg. *Flask web development: developing web applications with python*. " O'Reilly Media, Inc.", 2018.

[13] Tomáš Hák, Svatava Janoušková, and Bedřich Moldan. Sustainable development goals: A need for relevant indicators. *Ecological Indicators*, 60:565–573, 2016.

[14] Faridul Islam, Muhammad Shahbaz, Ashraf U. Ahmed, and Md. Mahmudul Alam. Financial development and energy consumption nexus in malaysia: A multivariate time series analysis. *Economic Modelling*, 30:435–441, 2013.

[15] Jing Jiang, Menghan Xu, Sen Pan, and Lipeng Zhu. Intelligent identification of similar customers for electricity demand estimation based on metadata of household background. In *Artificial Intelligence and Robotics: 7th International Symposium, ISAIR 2022, Shanghai, China, October 21-23, 2022, Proceedings, Part I*, pages 271–280. Springer, 2022.

[16] Tae-Young Kim and Sung-Bae Cho. Predicting residential energy consumption using cnn-lstm neural networks. *Energy*, 182:72–81, 2019.

[17] Brett Lantz. *Machine learning with R: expert techniques for predictive modeling*. Packt publishing ltd, 2019.

[18] Alasdair Bruce Lyndon Ruff. Carbon intensity api.

[19] David JC MacKay et al. Introduction to gaussian processes. *NATO ASI series F computer and systems sciences*, 168:133–166, 1998.

[20] William D McGinnis, Chapman Siu, S Andre, and Hanyu Huang. Category encoders: a scikit-learn-contrib package of transformers for encoding categorical data. *Journal of Open Source Software*, 3(21):501, 2018.

[21] Wes McKinney. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.", 2012.

[22] Hiroyuki Mori. State-of-the-art overview on data mining in power systems. In *2006 IEEE PES Power Systems Conference and Exposition*, pages 33–34. IEEE, 2006.

[23] David Murray, Lina Stankovic, and Vladimir Stankovic. An electrical load measurements dataset of united kingdom households from a two-year longitudinal study. *Scientific data*, 4(1):1–12, 2017.

[24] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[25] Kedar Potdar, Taher S Pardawala, and Chinmay D Pai. A comparative study of categorical variable encoding techniques for neural network classifiers. *International journal of computer applications*, 175(4):7–9, 2017.

[26] Mark R Segal. Machine learning benchmarks and random forest regression. 2004.

[27] Hussain Shareef, Maytham Ahmed, Azah Mohamed, and Eslam Al Hassan. Review on home energy management system considering demand responses, smart technologies, and intelligent controllers. *IEEE Access*, PP:1–1, 04 2018.

[28] Artur Trindade. ElectricityLoadDiagrams20112014. UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C58C86.

[29] Alexander Tureczek, Per Sieverts Nielsen, and Henrik Madsen. Electricity consumption clustering using smart meter data. *Energies*, 11(4):859, 2018.

[30] Gülden Kaya Uyanık and Neşe Güler. A study on multiple linear regression analysis. *Procedia-Social and Behavioral Sciences*, 106:234–240, 2013.

[31] Chad Zanocco, Tao Sun, Gregory Stelmach, June Flora, Ram Rajagopal, and Hilary Boudet. Assessing californians' awareness of their daily electricity use patterns. *Nature Energy*, pages 1–9, 2022.

[32] Xiaoling Zhang, Lizi Luo, and Martin Skitmore. Household carbon emission research: an analytical review of measurement, influencing factors and mitigation prospects. *Journal of Cleaner Production*, 103:873–883, 2015.

Appendix A

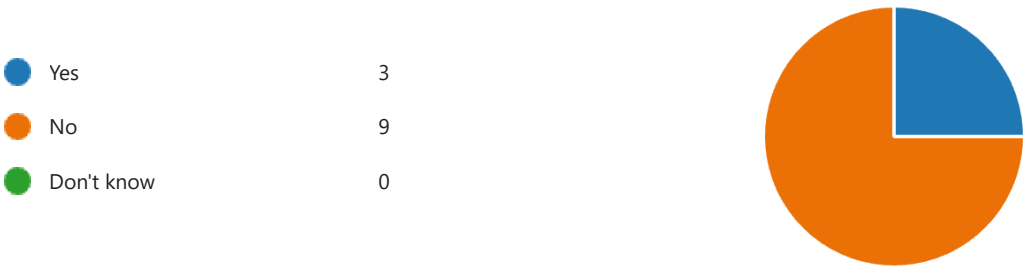
Survey Questionnaire and Survey Responses

Survey - User Requirements for Data-Driven Energy Management System

12	03:03	Active
Responses	Average time to complete	Status

1. Rank these in the decreasing order of importance to you: 1. Reducing energy costs, 2. Increasing use of renewable energy resources 3. Reducing carbon emissions

2. Do you use Electric Vehicles?



3. Which of these weather forecast related metrics are useful to you in order to schedule appliances or manage your electricity use better?

4. Do you use Home energy storage system?

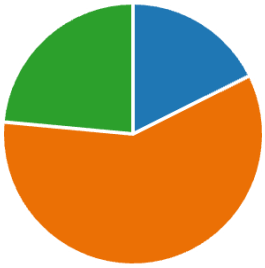
Yes	1
No	9
Don't know	2

5. Which variables or metrics related to energy usage trends are most relevant to you? Select ALL that apply.

Total carbon emissions	7
Seasonal energy consumption	2
Carbon intensity (per unit of ele...	5
Total renewable energy generati...	5
Energy consumption per applia...	5

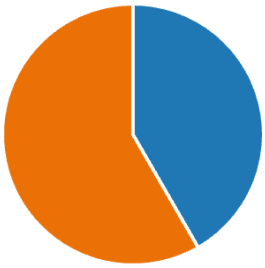
6. Which of these energy efficiency metrics or benchmarks would you like to track or visualise? Select all that apply.

Energy usage per square foot	3
Energy efficiency rating	10
Carbon Intensity	4



7. Do you use solar panels?

Yes	5
No	7
Don't Know	0

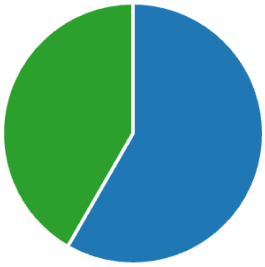


8. Which of these demographic factors would you like the ability to compare and benchmark energy usage trends across? Select ALL that apply.

Household size	4
Total household income	10
Location	10
Time of the year (season)	5

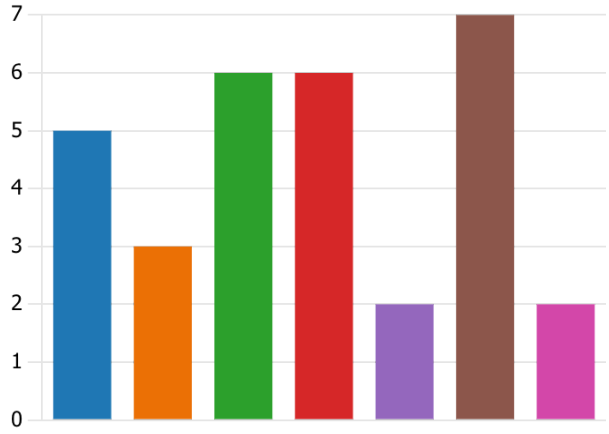
9. How do you utilise the generated renewable energy? Are you primarily self-consuming it or exporting it back to the grid?

Self-consuming	7
Exporting back to the grid	0
Don't Know	5



10. Are there any specific visualization formats or tools that you prefer when analyzing energy usage trends data? Select all that apply.

Bar chart	5
Scatter Plot	3
Heat Maps	6
Line Chart	6
Histogram	2
Pie Chart	7
Box Plot	2



11. Are there any other visualisations that would be useful to you? Please mention them here.

12
Responses

Latest Responses

"No"

"Nope"

"NA"

5 respondents (**42**%) answered **No** for this question.

12. If you are conducting any research in Energy Informatics or related fields, would it be helpful for you to have access to aggregated data or individual household-level data for conducting your research?

Not conducting research	5
Aggregated data	3
Individual household-level data	4
Both	0

13. Any additional comments?

12
Responses

Latest Responses

"No"

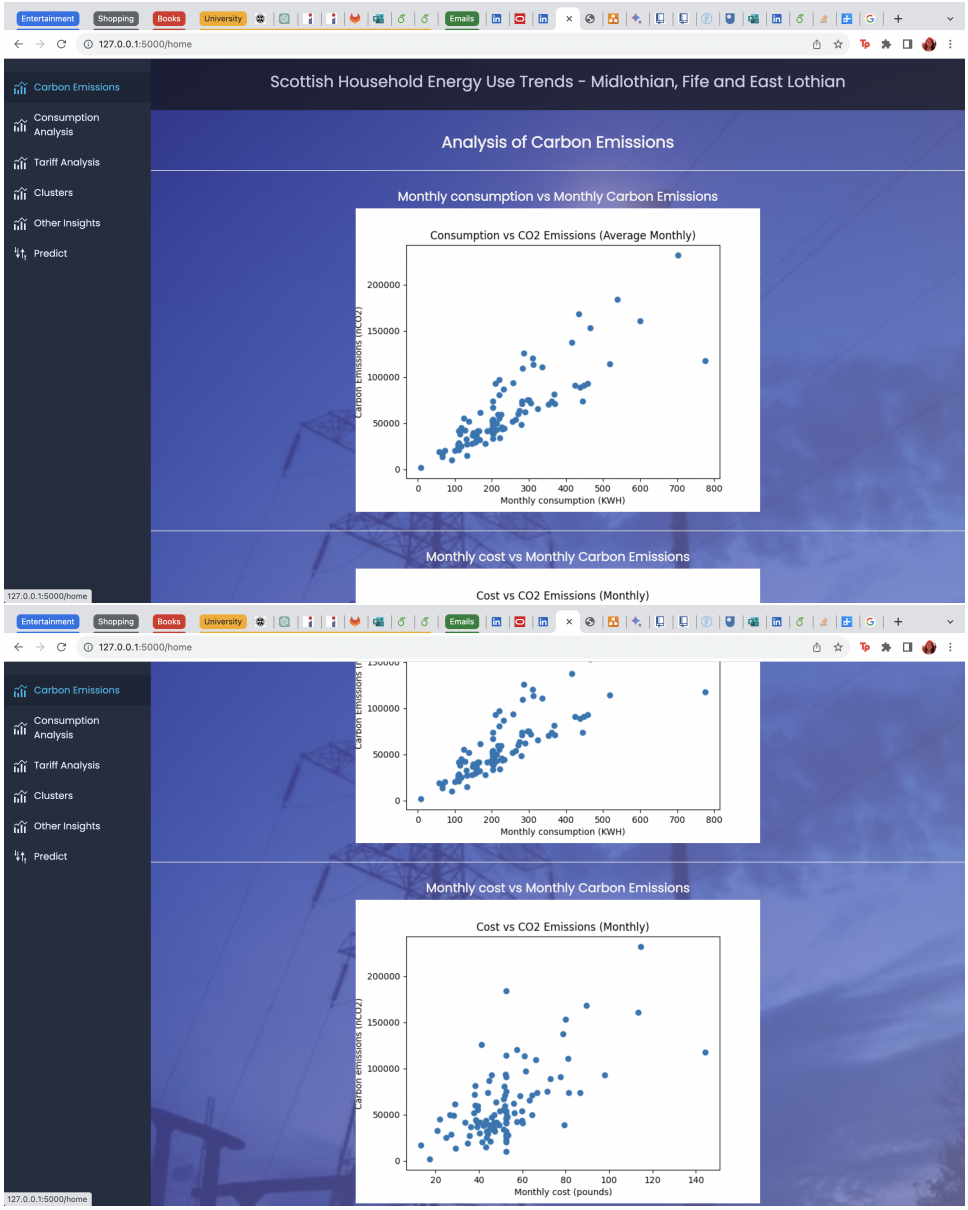
"None"

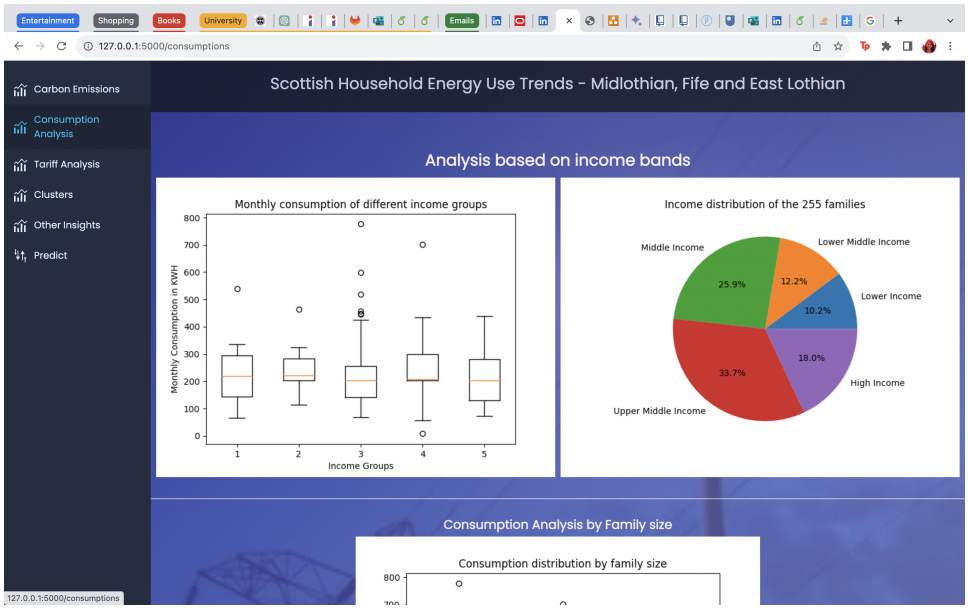
"NA"

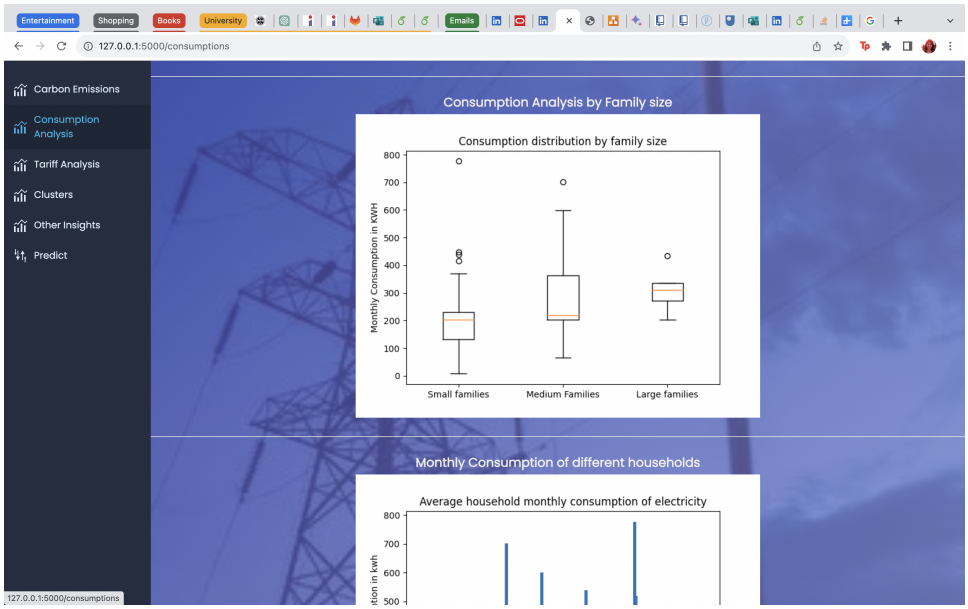
5 respondents (**42**%) answered **No** for this question.

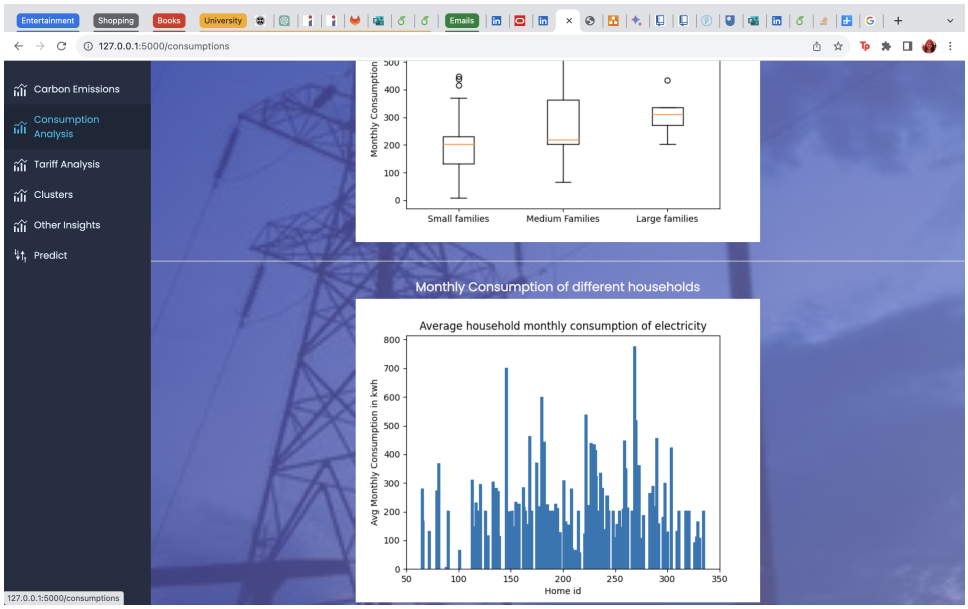
Appendix B

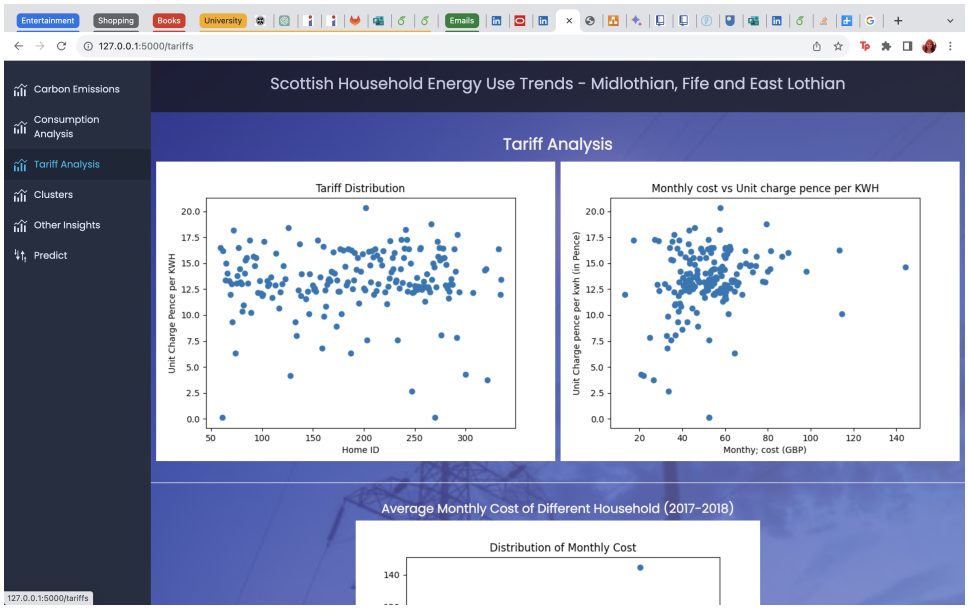
System UI screenshots

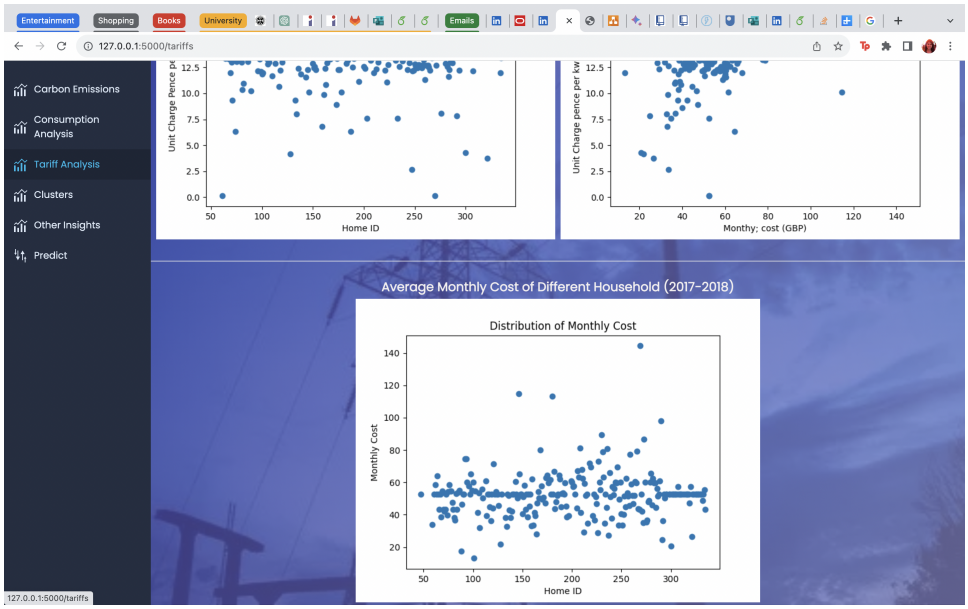


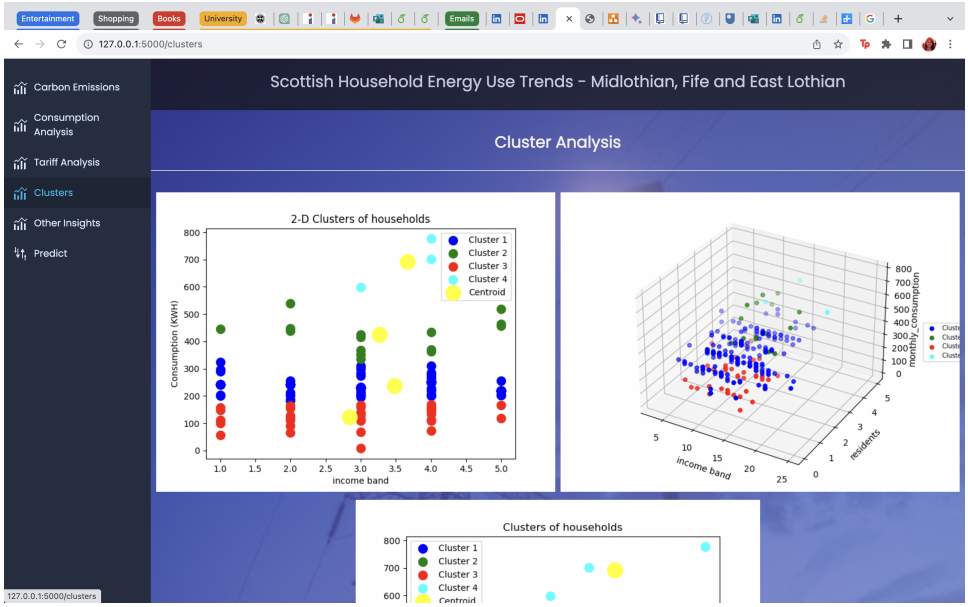

Entertainment
Shopping
Books
University
Emails
127.0.0.1:5000/consumptions
Carbon Emissions
Consumption Analysis
Tariff Analysis
Clusters
Other Insights
Predict
Scottish Household Energy Use Trends - Midlothian, Fife and East Lothian
Analysis based on income bands
Monthly consumption of different income groups
Monthly Consumption in KWH
Income Groups
Income distribution of the 255 families
Middle Income
Lower Middle Income
Lower Income
High Income
Upper Middle Income
25.9%
12.2%
10.2%
18.0%
33.7%
Consumption Analysis by Family size
Consumption distribution by family size
127.0.0.1:5000/consumptions

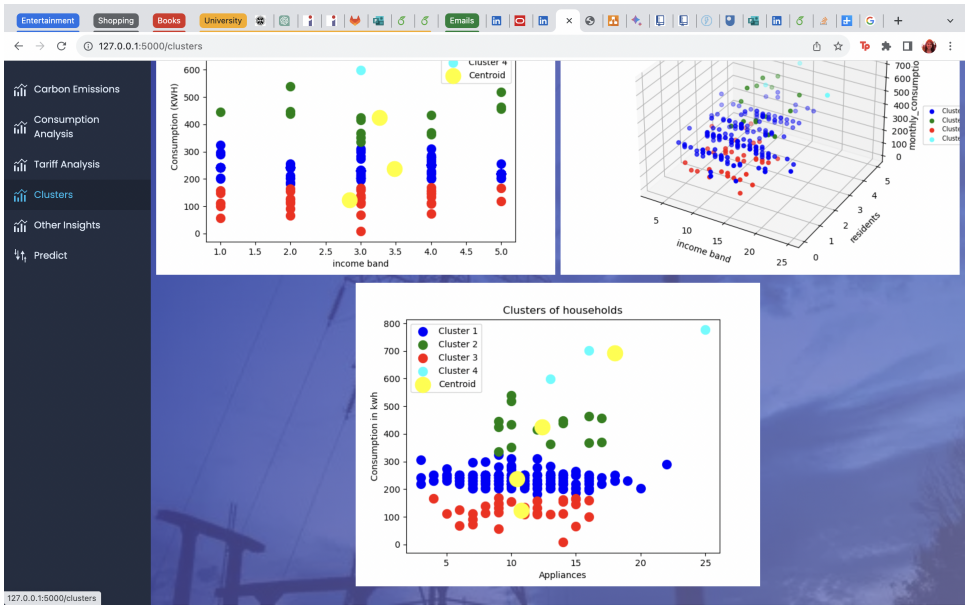

Entertainment
Shopping
Books
University
Emails
127.0.0.1:5000/consumptions
Carbon Emissions
Consumption Analysis
Tariff Analysis
Clusters
Other Insights
Predict
Consumption Analysis by Family size
Consumption distribution by family size
Monthly Consumption in KWH
Small families
Medium Families
Large families
Monthly Consumption of different households
Average household monthly consumption of electricity
127.0.0.1:5000/consumptions

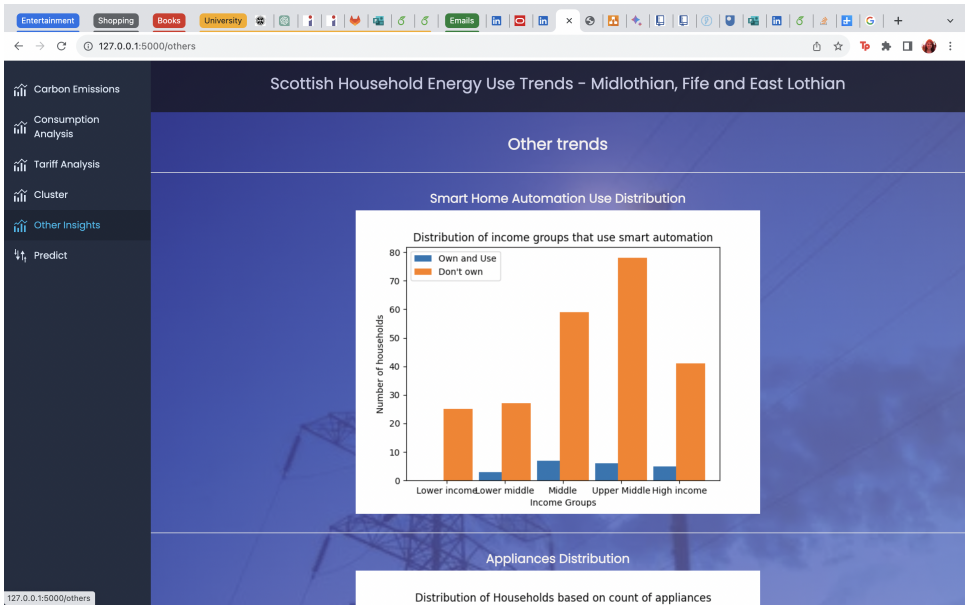

Entertainment
Shopping
Books
University
Emails
127.0.0.1:5000/consumptions
Carbon Emissions
Consumption Analysis
Tariff Analysis
Clusters
Other Insights
Predict
Monthly Consumption
Small families
Medium Families
Large families
Monthly Consumption of different households
Average household monthly consumption of electricity
Avg Monthly Consumption in kwh
Home id
127.0.0.1:5000/consumptions

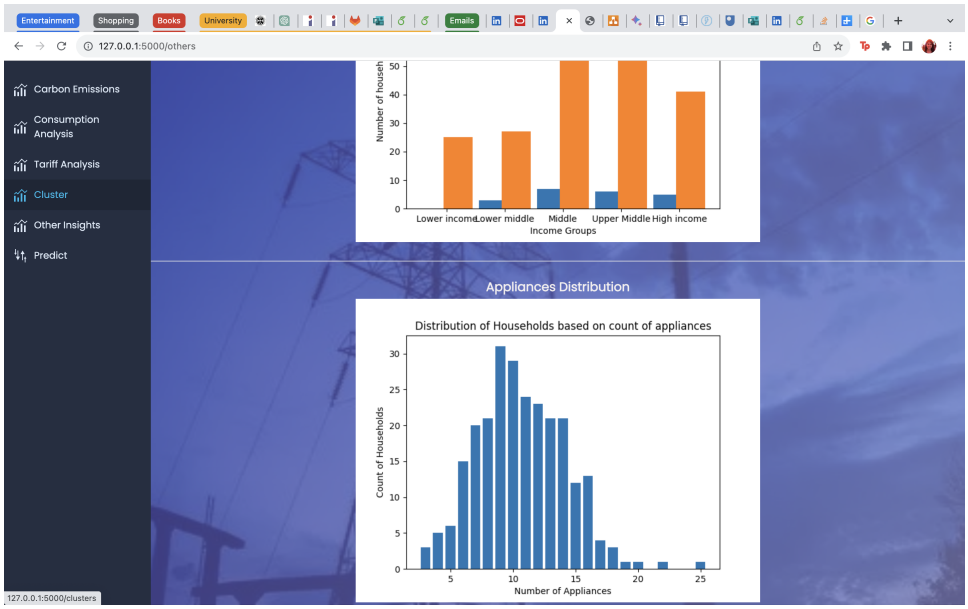

Entertainment
Shopping
Books
University
Emails
127.0.0.1:5000/tariffs
Carbon Emissions
Consumption Analysis
Tariff Analysis
Clusters
Other Insights
Predict
Scottish Household Energy Use Trends - Midlothian, Fife and East Lothian
Tariff Analysis
Tariff Distribution
Unit Charge Pence per KWH
Home ID
Monthly cost vs Unit charge pence per KWH
Unit Charge pence per kwh (in Pence)
Monthy; cost (GBP)
Average Monthly Cost of Different Household (2017-2018)
Distribution of Monthly Cost
127.0.0.1:5000/tariffs

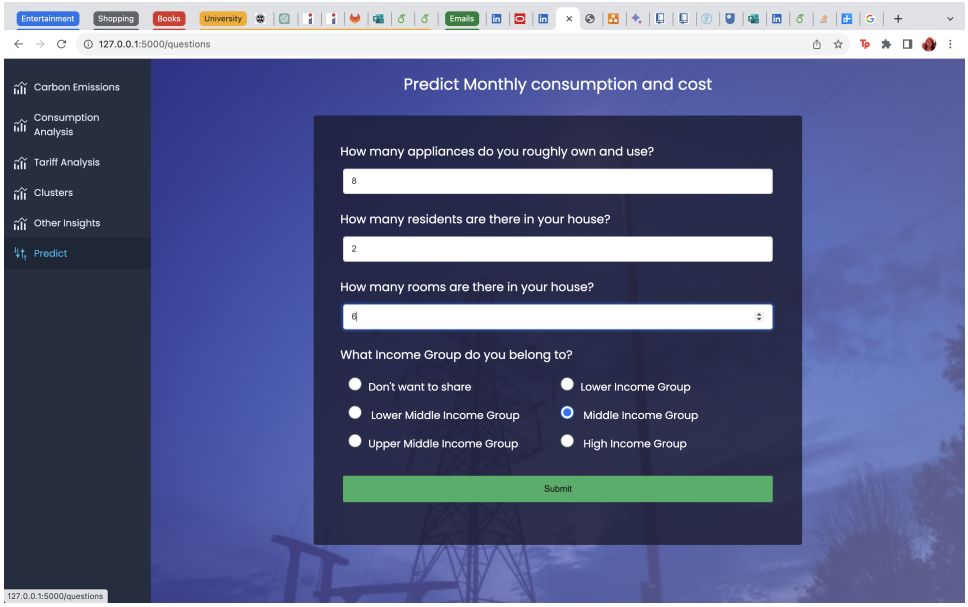

Carbon Emissions
Consumption Analysis
Tariff Analysis
Clusters
Other Insights
Predict
Unit Charge Pence per kWh
Home ID
Unit Charge pence per kWh
Monthy; cost (GBP)
Average Monthly Cost of Different Household (2017-2018)
Distribution of Monthly Cost
Monthly Cost
Home ID

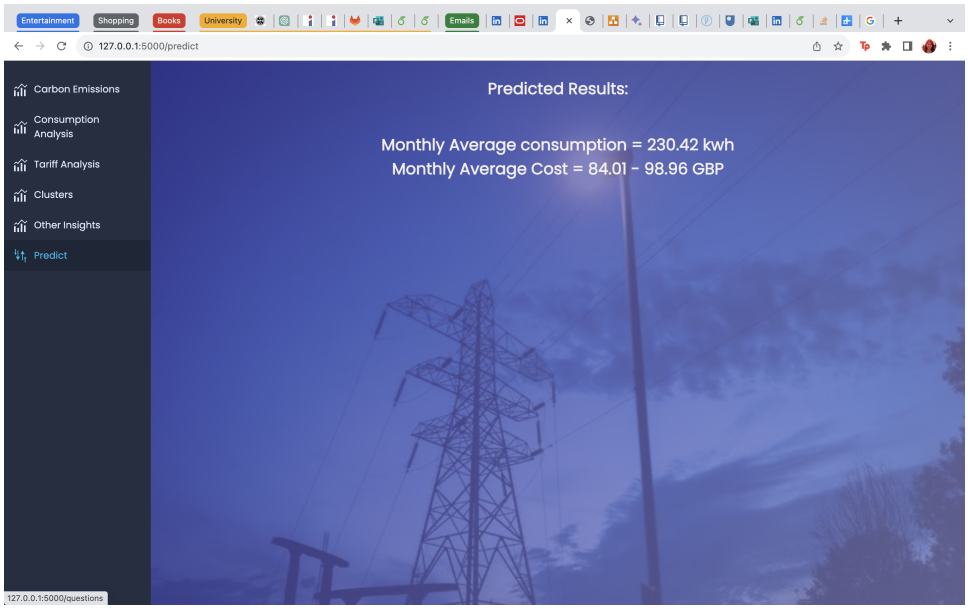

Carbon Emissions
Consumption Analysis
Tariff Analysis
Clusters
Other Insights
Predict
Scottish Household Energy Use Trends - Midlothian, Fife and East Lothian
Cluster Analysis
2-D Clusters of households
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Centroid
Consumption (KWH)
income band
income band
residents
monthly consumption
Clusters of households


Entertainment
Shopping
Books
University
Emails
127.0.0.1:5000/clusters
Carbon Emissions
Consumption Analysis
Tariff Analysis
Clusters
Other Insights
Predict
Consumption (KWH)
income band
Centroid
monthly Consumption
residents
Clusters of households
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Centroid
Consumption in kwh
Appliances
127.0.0.1:5000/clusters


Entertainment
Shopping
Books
University
Emails
127.0.0.1:5000/others
Carbon Emissions
Consumption Analysis
Tariff Analysis
Cluster
Other Insights
Predict
Scottish Household Energy Use Trends - Midlothian, Fife and East Lothian
Other trends
Smart Home Automation Use Distribution
Distribution of income groups that use smart automation
Own and Use
Don't own
Number of households
Lower income
Lower middle
Middle
Upper Middle
High income
Income Groups
Appliances Distribution
Distribution of Households based on count of appliances
127.0.0.1:5000/others


127.0.0.1:5000/others
Carbon Emissions
Consumption Analysis
Tariff Analysis
Cluster
Other Insights
Predict
Number of household
Lower income
Lower middle
Middle
Upper Middle
High income
Income Groups
Appliances Distribution
Distribution of Households based on count of appliances
Count of Households
Number of Appliances
127.0.0.1:5000/clusters


127.0.0.1:5000/questions
Carbon Emissions
Consumption Analysis
Tariff Analysis
Clusters
Other Insights
Predict
Predict Monthly consumption and cost
How many appliances do you roughly own and use?
8
How many residents are there in your house?
2
How many rooms are there in your house?
6
What Income Group do you belong to?
Don't want to share
Lower Income Group
Lower Middle Income Group
Middle Income Group
Upper Middle Income Group
High Income Group
Submit
127.0.0.1:5000/questions


Entertainment
Shopping
Books
University
Emails
127.0.0.1:5000/predict
Carbon Emissions
Consumption Analysis
Tariff Analysis
Clusters
Other Insights
Predict
Predicted Results:
Monthly Average consumption = 230.42 kwh
Monthly Average Cost = 84.01 - 98.96 GBP
127.0.0.1:5000/questions