Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Astrid Solheim

Including New Customers in The Prediction of Electricity Consumption

Master's thesis in Applied Physics and Mathematics
Supervisor: John Sølve Tyssedal
Co-supervisor: Tibber - Victor Sundberg and Gian-Luca Francone
December 2023

NTNU
Norwegian University of
Science and Technology

tibber

Astrid Solheim

Including New Customers in The Prediction of Electricity Consumption

Master's thesis in Applied Physics and Mathematics
Supervisor: John Sølve Tyssedal
Co-supervisor: Tibber - Victor Sundberg and Gian-Luca Francone
December 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Preface

This master thesis is the finishing assignment of my degree in Physics and Mathematics within Industrial Mathematics. The master thesis has been performed during the fall 2023 under the supervision of professor John Sølve Tyssedal and delivered to the department of Mathematical Sciences at the Norwegian University of Science and Technology (NTNU). The thesis was written in collaboration with Tibber AS with their Data Scientist team.

The primary objective of this study has been to analyze Tibber's customer base and explore whether statistical methods could offer improved ways to forecast customer consumption patterns.

I would like to thank my supervisor John Sølve Tyssedal for his guidance and support during both the project and master thesis. His advice and response in physical and digital meetings, as well as mails, have been of great help. I feel very fortunate to have had such a dedicated supervisor. I also want to express my appreciation to Tibber for the opportunity to write this thesis and the help from Victor Sundberg and Gian-Luca Francone by biweekly meetings. A special thanks to Victor for the additional assistance with coding in our digital interactions. Finally, I would like to thank my family and friends for supporting and encouraging me.

Astrid Solheim
Trondheim, December 2023

Abstract

For companies that purchase electricity in advance, achieving accurate predictions of consumption is crucial. Using machine learning and historical data enables the creation of predictive response models. The main topic in this thesis is to accurately classify new customers into appropriate clusters/groups based on their initial variables. The idea is that classification can develop effective prediction models for new homes lacking historical hourly consumption data. Tibber, an energy company operating in the Nordic region, has generously provided time series data related to customers' electricity consumption. The insights gained from the master analysis are anticipated to provide valuable information for the energy company.

The primary goal of this thesis involves to evaluate a range of classification methods to determine if any models exhibit exceptional performance in delivering an acceptable accuracy. Clusters were used as response variable to evaluate different classification methods. These clusters were created based on the mean hourly consumption within a subset of the data set. Several classification models have been trained using the clusters as response variables, including K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Support Vector Machine (SVM), Random Forest (RF), and Kernel Density Classifier (KDC). Among these, LDA demonstrated the best performance on the test set, leading to its selection for further analysis.

Later on, with the new customers together with the old customers, consumption values were predicted using a Generalized Linear Model (GLM) and a Linear Model (LM) for each group. The overall predicted consumption was then compared against the actual overall consumption, measured by Mean Absolute Error (MAE). Given the diverse variables and missing values in the data, various approaches to handle the data were explored.

The final MAE value achieved with the LDA model, where 248 of 829 were considered new customers, was 165.7 kWh. When considering the entire data set as one group, the MAE value was 218.3 kWh. Alongside with the linear model, an additional variable was later incorporated, involving the consumption variable from the previous day. This inclusion resulted in a MAE of 159.7 kWh for the clusters. Consequently, the combination of clustering and classification methods resulted in an improvement in accuracy compared to all the data as one single cluster.

Samandrag

Det er avgjerande for bedrifter som kjøper straum på førehand å oppnå nøyaktige prognosar for forbruket. Ved å nytte maskinlæring og historiske data kan ein utvikle prediktive responssmodellar. Hovudmålet med masteroppgåva er å klassifisere nye kundar inn i passande klynger/grupper basert på startvariablane deira. Tanken er at klassifisering kan utvikle effektive prognosemodellar for nye kundar som manglar historisk times forbruksdata. Tibber, eit energiselskap som opererer i Norden, har velvillig levert oss tidsseriedata knytt til straumforbruket til kundane sine. Kunnskapen frå masteranalysen vil forhåpentlegvis gje verdifull innsikt for energiselskapet.

Dei viktigaste måla med denne oppgåva involverer å evaluere eit spekter av klassifikasjonsmetodar for å avgjere om nokre modellar viser god ytelse med akseptabel nøyaktigheit. Klynger vart nytta som responsvariabel for å evaluere ulike klassifiseringsmetodar. Desse klyngene blei til frå det gjennomsnittlege timesforbruket av ei undergruppe av datasettet. Fleire klassifikasjonsmodellar vart trente ved å nytte klyngene som responsvariablar, inkludert K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Support Vector Machine (SVM), Random Forest (RF) og Kernel Density Classifier (KDC). Blant desse viste LDA den beste ytelsen på testsettet, noko som førte til at den blei vald for vidare analyse.

Etterpå vart forbruksverdiane, for nye og gamle kundar, predikert ved å nytte ein Generalized Linear Model (GLM) og ein Linear Model (LM) for kvar gruppe. Det samla predikerte forbruket vart deretter samanlikna med det faktiske samla forbruket, målt i Mean Absolute Error (MAE). Sidan datasettet inneheldt fleire variablar og manglande verdiar, vart det utforska ulike tilnærmingsmåtar for å handtere dataen.

Den endelege MAE-verdien oppnådd med LDA-modellen for 248 nye kundar var 165.7 kWh. Då heile datasettet vart vurdert som ei gruppe, var MAE-verdien 218.3 kWh. Saman med den linære modellen vart det integrert ein ekstra variabel som omfatta forbruket frå dagen før. Denne inkluderinga resulterte i ein MAE på 159.7 kWh for klyngene. Som eit resultat viste kombinasjonen av klynge- og klassifikasjonsmetodar å gi ei auka nøyaktigheit i forhold til å sjå på alle dataane som ei klynge.

Table of Contents

List of Figures

List of Tables

1 Introduction

Electricity plays a crucial role in the Norwegian energy market and is a fundamental element of the ongoing transition to green energy in Norway. The accessibility and affordability of electricity are of vital importance, particularly during the winter months when electricity prices tend to rise. This can pose challenges for many households, making it essential to seek continuous improvements in energy efficiency and cost management.

The Nordic region is recognized as a unified electricity market divided into several price areas, representing specific regions for power market transactions. Taking Norway as an example, Oslo is located in price area NO1. Power prices are determined by the Nord Pool power exchange, and since all Nordic countries trade from Nord Pool, prices change based on supply and demand from neighboring countries. The power and market conditions in each area are influenced by the power flow between the spot market areas, allowing for the transfer of power from one region to another based on demand. This segmentation into price areas aims to maintain balance in the power marked and reduce the risk of local or regional power shortages [4, 20].

Tibber is a digital electricity company that try to provide as cheap and well-timed electricty for their customers. They do this by forecasting the consumption of a household based on earlier consumption pattern and outside factors that affect the customers usage. They have a customer base of 400 000 people located in Norway, Sweden, Germany and also the Netherlands. Rather than imposing a fixed fee on customers, Tibber provides spot prices, refraining from any profit on customers' consumption [22]. Additionally, they leverage smart home devices to enhance cost savings for customers. While it is certainly feasible to forecast each individual home's consumption, this approach can be time-consuming. Hence, simplifying the forecasting process and uncovering trends that may not be immediately evident when examining individual homes, we aim to investigate the potential benefits of clustering homes into groups for making forecasts.

The data sets utilized in this thesis have been supplied by Tibber, derived from their Swedish customer base in price area SE3. Two data sets have been provided, one containing time series data and the other comprising properties of various households. The first data set primarily consists of continuous variables, including 11 explanatory variables and a length of 9 117 535 rows. The second is predominantly composed of categorical variables, containing 141 108 rows and 9 explanatory variables. Although both data sets have been employed in the thesis, the primary emphasis has been on the home properties data as they were used in the classification. A more in-depth exploration of these data sets will be undertaken in Section 3.

Certain sections in this thesis have been adapted, with minor adjustments, from my project thesis [16]. This is particularly evident in Chapter 1, including the problem and motivation discription. In Chapter 2, Section 2.4 and Section 2.5 have been added from [16] with alterations. Additionally, certain plots from [16] have been included in both Chapter 2 and Chapter 4, appropriately credited as referenced. The problem solving is based on the results of the project thesis and the same data set has also been used, thereas the same explanation of the data in Chapter 3.

1.1 Motivation

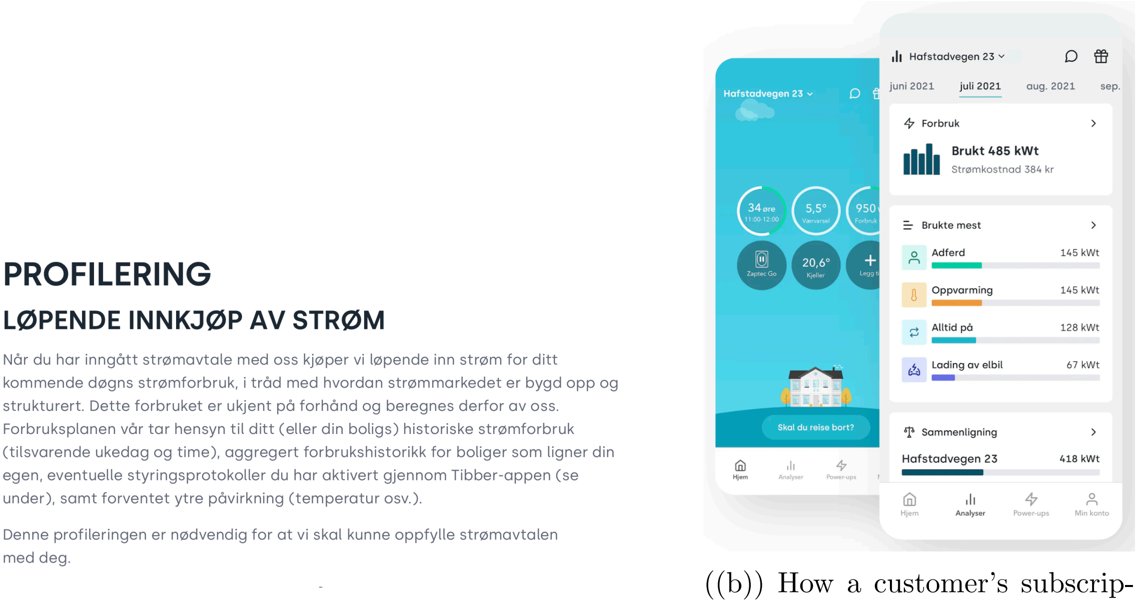
Tibber buy a large amount of electricity in advance based on the predicted consumption of their customers. This is naturally a very large expense and there is a lot of money that can go to waste if the forecast predicts poorly. In the event of an underprediction, Tibber risks acquiring an insufficient amount of electricity, potentially leaving customers without a satisfactory supply. Conversely, overprediction results in unnecessary financial outlays for Tibber. Furthermore, there is a preference to maintain a balance between supply and demand within the price areas, as previously mentioned, with the intention of minimizing the potential for power shortages. Ensuring a reliable electricity supply for customers remains the top priority for an electricity company like Tibber. Figure 1 shows the marketing strategy for their electricity offering to customers and offers a visual representation of the app's configuration.

Tibber derives its forecasts from the historical consumption patterns of its customers. However, a challenge arises when dealing with new customers for whom the hourly consumption history is unknown. To provide an accurate estimate of their consumption, the prediction model requires some time to adjust. Addressing this issue is crucial for the company to obtain a reliable forecast of future electricity. Despite the absence of historical hour consumption data for new customers, they provide essential home properties information, including the size, type of their household and their annual consumption. Using this initial data, our aim is to categorize new customers into appropriate groups, providing a reliable estimate of their future consumption.

1.2 Outline

The goal of the master thesis is to address new customers properly, so that the forecast is as good as possible from start of their Tibber subscription. Several classification methods were examined, taking into account the presence of both continuous and categorical variables in the problem. Certain models, such as LDA and QDA, are limited to handling continuous variables exclusively. Furthermore, when incorporating categorical variables into a classification method, it is necessary to convert them, and in this context, the use of dummy variables has been explored.

The thesis is structured as followed. It begins with an introduction that presents the problem and provides the motivation behind the research. This is followed by Chapter 2, the theory section, that provides an in-depth understanding of the various methods explored, including discussions on data transformation, as well as explanations of classification and clustering techniques. Chapter 3 delves into the details of the data set, providing insights into its structure, visual representation, and addressing associated challenges. The method section, Chapter 4, outlines the experimental implementation and the approach taken to address the data-related challenge. Subsequently, Chapter 5 is where the results and analysis are presented. Chapter 6 is the final chapter that concludes the thesis and offers recommendations for future work.



PROFILERING

LØPENDE INNKJØP AV STRØM

Når du har inngått strømavtale med oss kjøper vi løpende inn strøm for ditt kommende døgns strømforbruk, i tråd med hvordan strømmarkedet er bygd opp og strukturert. Dette forbruket er ukjent på forhånd og beregnes derfor av oss. Forbruksplanen vår tar hensyn til ditt (eller din boligs) historiske strømforbruk (tilsvarende ukedag og time), aggregert forbrukshistorikk for boliger som ligner din egen, eventuelle styringsprotokoller du har aktivert gjennom Tibber-appen (se under), samt forventet ytre påvirkning (temperatur osv.).

Denne profileringen er nødvendig for at vi skal kunne oppfylle strømavtalen med deg.

((a)) How they explain the usage of the electricity profile of customers to buy electricity.

Hafstadvegen 23

juni 2021 juli 2021 aug. 2021 sep.

Forbruk

Brukt 485 kWt

Strømkostnad 384 kr

Brukte mest

Adferd	145 kWt
Oppvarming	145 kWt
Alltid på	128 kWt
Lading av elbil	67 kWt

Sammenligning

Hafstadvegen 23 418 kWt

((b)) How a customer's subscription and usage are displayed in the Tibber app.

Figure 1: Two images of the marketing on the website of Tibber, taken 11.10.23.

2 Theoretical Background

This section delves into the different classification methods tested in this thesis. The theory also includes various ways the data could be preprocessed and presented before usage.

2.1 Continuous and categorical variables

There are two groups of variables that can be characterized, continuous or categorical also known as quantitative and qualitative. Continuous variables can have different values as in age and size which can have an impact depending on the value. Categorical variables are split into N different classes. These values are often independent and will not affect each other. Variables like these can be an animal type, a name or only binary answers (yes or no). Usually, problems with quantitative response are solved with regression models, while classification models are mostly used when the response is qualitative. There are, however, many different cases and methods where both continuous and categorical variables have to be taken into consideration. Many of the known statistical methods can handle both as long as the variables are properly preprocessed and ready for usage [11].

2.2 Classification

Classification is a technique used to assign new objects to a predefined number of groups. In this process, the chosen classification method's goal is to categorize objects into two or more labeled classes, with the primary aim being the optimal assignment of new objects to these labeled classes [12]. Discriminants are often used to seperate groups or collections as much as possible. Classification rules are typically derived from training samples, which consist of randomly selected objects with known associations to specific populations. As a result, the potential sample outcomes are constrained to specific defined regions. For instance, in a scenario with two regions, if a new observation falls into region 1 (R_1), it is assigned to population π_1, and conversely, if it falls into R_2, it is assigned to population π_2.

For two groups, classification can stay rather simple as we only need to know which class is above a given threshold, for example 50%. When dealing with more than two groups, the situation becomes more complex. A challenge is that the behavior of a linear statistics, for instance, depends significantly on where the population is located [12].

Consider a classification problem involving g classes. Let $f_i(\mathbf{x})$ be the probability density function associated with an observation $\mathbf{x}$ from population π_i and p_i denotes the prior probability that an observation originates from π_i, $i = 1, 2, ..., g$. Applying Bayes' Theorem allows for the calculation of the posterior probability when a specific data point x_0 is observed.

$$P(\pi_i|x_0) = \frac{f(\pi_i \cap \mathbf{x}_0)}{f(\mathbf{x}_0)} = \frac{f(\mathbf{x}_0 \cap \pi_i)}{\sum_{i=1}^{q} f(\mathbf{x}_0|\pi_i)p_i}$$

$$= \frac{f(\mathbf{x}_0|\pi_i)p_i}{\sum_{i=1}^{q} f(\mathbf{x}_0|\pi_i)p_i} = \frac{f_i(\mathbf{x}_0)p_i}{\sum_{i=1}^{q} f_i(\mathbf{x}_0)p_i} \tag{1}$$

Since the denominator remains consistent across all cases, it is natural to classify x_0 to belong to the population π_k which fulfills

$$p_k f_k(\mathbf{x}_0) > p_i f_i(\mathbf{x}_0) \ \forall \ i \neq k. \tag{2}$$

Now, let $c(k|i)$ be the cost of allocating an observation to π_k when it belongs to $\pi_i \neq \pi_k$. In classification, the Minimum Expected Cost of Misclassification Method (ECM) is usually considered. This method calculates the cost of being classified wrong for each class.

Misclassification of a x from π_1 to one of the other groups $\pi_2, \pi_3, ..., \pi_g$ will make $\text{ECM}(1) = \sum_{k=2}^{g} P(\text{classify to } \pi_k|\pi_1)c(k|1)$. Here, we have

$$P(\text{classify to } \pi_k|\pi_1) = P(k|1) = \int_{R_k} f_1(\mathbf{x})d\mathbf{x}.$$

Similarly, the conditional expected costs of misclassification, $\text{ECM}(2), ..., \text{ECM}(g)$, can be calculated. Then, the total $\text{ECM} = \sum_{i=1}^{q} p_i \text{ECM}(i)$ can be computed as the sum of each conditional ECM, multiplied by the prior probability.

$$\text{ECM} = p_1\text{ECM}(1) + p_2\text{ECM}(2) + ... + p_g\text{ECM}(g)$$
$$= \sum_{i=1}^{q} p_i \left(\text{ECM}(i)\right) = \sum_{i=1}^{q} p_i \left(\sum_{\substack{i=1 \\ i\neq k}}^{g} P(k|i)c(k|i) \right). \tag{3}$$

Selecting the ideal classification approach involves choosing regions that are non-overlapping and all-inclusive, in order to minimize Equation 3.

It can be shown that the ECM allocate x_0 to the group π_k where

$$\sum_{\substack{i=1 \\ i\neq k}}^{g} p_i f_i(\mathbf{x})c(k|i)$$

has the lowest value [12]. Here, p_i is the prior probability of the chosen group i, f_i is the density connected to the population and $c(k|i)$ is the cost of assigning the the group π_k to the group π_i. If the cost is equal for two or more groups, $\mathbf{x}$ can be placed in any of those populations with the same, minimal cost.

2.2.1 The Bayes Classifier

Bayes is a simple classifier that allocate the test observations to a class based on the predictor values [11]. A new observation will be classified to the class for which the probability

$$\Pr(Y = j|X = x_0),\ j = 1, 2, ..., g \tag{4}$$

is the largest. Here, j is the class and x_0 is the predictor value. In a simple classification problem, where there are only 2 classes, the Bayes Classifier would assign a test observation a class based on

$$\Pr(Y = 1|X = x_0) > 0.5.$$

In this case, the test observation would, according to the classifier, belong to class 1.

Figure 2 is an example from p. 38 in [11], where we can see how the classifier works on simulated data. As the figure shows, there are two classes, orange and blue, made from the training observations. There is a lined distinction between the bright orange and blue shades, this shows which class a new test observation is more likely to belong to. If a test point X is placed in the blue shade, this means that $\Pr(Y = \text{blue}|X) > 0.5$. Therefore, the test point will be assigned to class blue. The dashed line is called the Bayes decision boundary and is where the probability is 50% for both classes.

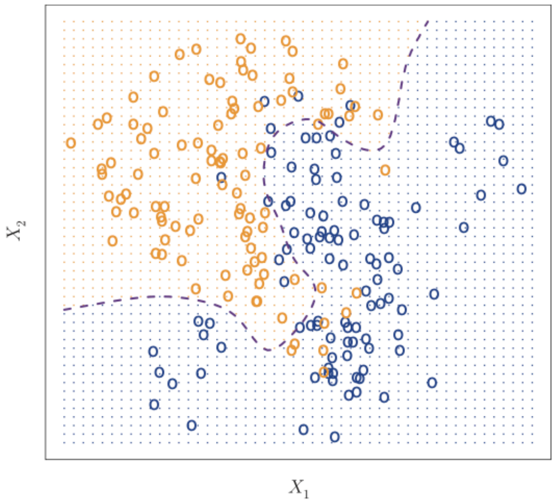


Figure 2: How the Bayes Classifier operates on a two-dimensional space with the predictors X_1 and X_2.

Source: Page 38 in [11]

In addition, the Bayes classifier finds the minimum test error rate, referred to the Bayes error rate. The Bayer classifier selects the class with the highest value for equation 4. The expectation involves averaging the probability across all possible values of X.

$$1 - E\left(\max_{j} \Pr(Y = j|X)\right).$$

Several classification methods are built upon the Bayes Classifier, this includes both in KDC and LDA, with a detailed description to follow later on.

2.2.2 K-nearest neighbour

The K-nearest neighbour (KNN) involves estimating the conditional distribution of Y given X, determining the highest estimated probability for assigning to a specific class. The KNN algorithm measures the distance between the observations using Euclidean distance. For a test observation, the algorithm considers its closest neighbors, where the integer K denotes the number of training observations nearest to x_0 to be examined during the KNN process. The set of these K nearest neighbours is denoted as $\mathcal{N}_0$. The method then calculates the conditional probability for class j as the fraction of points in $\mathcal{N}_0$ where response values equal the class j. The target class for the new observation is determined by the majority of the class within this neighbourhood [11].

$$\mathrm{Pr(Y{=}j|X{=}x_0)} = \frac{1}{K}\sum_{i\in\mathcal{N}_0} \mathrm{I(y}_i = \mathrm{j)} \tag{5}$$

A disadvantage with KNN is that the amount of neighbours, K, compared with the new object has to be identified at the beginning. Testing can certainly be performed in this context, like applying the elbow method. This technique involves executing KNN clustering with various K intergers to determine the optimal number of clusters. This involves setting up a loop using the training set, where the KNN model is employed with different initial K values. Throughout this process, the error rate is monitored by comparing the model's predictions on the test set against the actual values. The final results are visualized in a plot, where the x-axis represents the integer K, and the y-axis depicts the corresponding error rates. Typically, a distinctive "elbow" shape emerges, and at the point where the slope decreases minimally, you can identify the desired K value. This approach ensures an appropriate number of clusters while preventing overfitting of the data set [7]. However, when testing for a lot of different data, the K integer might change for the best fits, and it has a large effect on how the classifier is obtained.

Figure 3 shows how the K-Nearest Neighbour operate when it gets a new observation. In this example, the method takes K=3 as input, indicating that it examines the three closest neighbors of a new observation and subsequently selects the class based on the highest estimated probability.

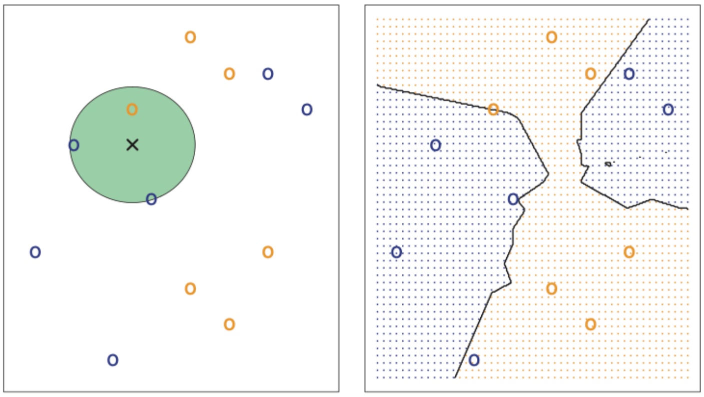


Figure 3: How the KNN method operates.

Source: Page 40 in [11]

2.2.3 LDA and QDA

Linear- and Quadratic discriminant analysis, LDA and QDA respectively, are very common methods to use when we are only considering continuous variables.

LDA is a popular classification method as it handles well-seperated classes well, is stable and often used with more than two response classes. In the problem of the master thesis, there are multiple response classes as well as multiple predictor variables. For LDA to work with a several predictors, it assumes that $X = (X_1, X_2, ..., X_p)$ is drawn from a multivariate Gaussian distribution. This distribution is a generalization of the normal density, designed for multiple dimensions [12]. The probability density function is given as

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}(\mathbf{x}-\boldsymbol{\mu})\right).$$

In this case, $\mathbf{\Sigma}$ is the covariance matrix of $\mathbf{X}$, $\boldsymbol{\mu}$ is the mean vector and $\mathbf{x}$ is the observation vector.

The linear discrimination score can be found by maximizing $p_i f_i(\mathbf{x}), i = 1, 2, ..., g$, as was shown in equation (2). Since maximizing $p_i f_i(\mathbf{x})$ is equivalent to maximizing $\ln(p_i f_i(\mathbf{x}))$, $\mathbf{x}$ should be classified to π_k if

$$\begin{aligned} \ln p_k f_k(\mathbf{x}) &= \ln p_k - \left(\frac{p}{2}\right)\ln 2\pi - \frac{1}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\left((\mathbf{x}-\boldsymbol{\mu}_k)^T\mathbf{\Sigma}(\mathbf{x}-\boldsymbol{\mu}_k)\right) \\ &= \max_i \ln p_i f_i(\mathbf{x}). \end{aligned}$$

The second term in the equation can be ignored, as this is constant for all the populations i. When expressing the term, we obtain $-\frac{1}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}\mathbf{x}$, which can be disregarded since it remains constant across $\delta_1^L(\mathbf{x}), \delta_2^L(\mathbf{x}), ..., \delta_g^L(\mathbf{x})$. Disregarding this term leaves us with a constant and a linear combination of components of $\mathbf{x}$. Therefore, the linear discrimination score for the i-th population can be defined as

$$\delta_i^L(\mathbf{x}) = \mathbf{x}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\boldsymbol{\mu}_i^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p_i, \;\; i = 1, 2, , ..., g.$$

The LDA classification works by comparing $\delta_i^L(\mathbf{x})$, $i = 1, 2, ..., g$, and assign $\mathbf{x}$ to the class for which $\delta_i^L(\mathbf{x})$ is largest. An essential component for this approach is the distinction in mean vectors for observations in different classes, while they share a common covariance matrix. In practice, both $\boldsymbol{\mu}_i$ and $\mathbf{\Sigma}$, $i = 1, 2, ..., g$ are unknown. However, utilizing a training set with correctly classified observations enables the construction of estimations for these parameters [12]. With n_j as the observations from π_i and $n = \sum_{j=1}^{g} n_j$, the estimated mean and covariance for a normal population can be given as

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{n_j}\sum_{j=1}^{n_j}\mathbf{x}_{ji} \quad \text{and} \quad \hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{j=1}^{n}(\mathbf{x}_j - \overline{\mathbf{x}})(\mathbf{x}_j - \overline{\mathbf{x}})^T. \tag{6}$$

Figure 4 shows how a LDA can look for three different classes. Here again, the dashed line represents the Bayes boundary descision based on the true values of $\boldsymbol{\mu}_i$ and $\mathbf{\Sigma}$. The colored ellipses regions the 95% probability of belonging to that specific class. In this example the observations are drawn from a mulitvariate Gaussian distribution with a specific mean vector of a chosen class and a common covariance matrix. As earlier, the Bayer Classifier assign an observation a class depending on where it is located. The image on the right-hand side shows the LDA descion boundaries based on the estimated $\boldsymbol{\mu}_i$ and $\mathbf{\Sigma}$ that are visualized with the black solid lines.

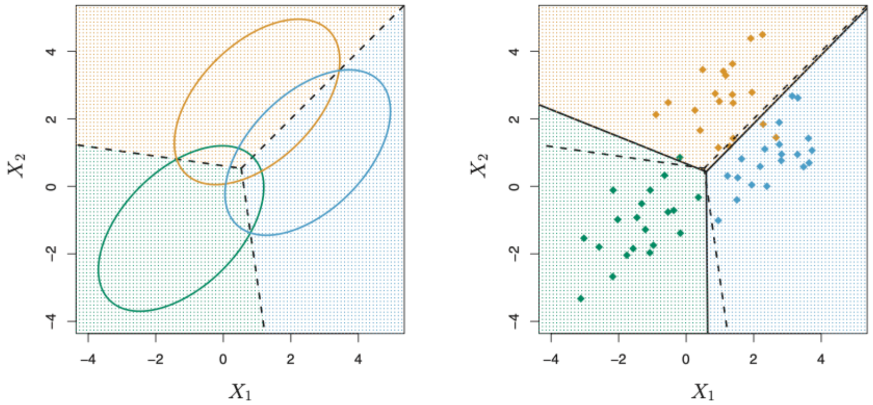


Figure 4: How the LDA method operates. The left picture shows the region of the classes, while the picture to the right shows the observations.

Source: Page 143 in [11]

Quadratic Discrimimant Analysis (QDA) is similar to LDA in many ways best differs in the way that QDA, unlike LDA, assumes each class has their individual covariance matrix. Therefore, for the quadratic discrimination score, the term $-\frac{1}{2}\log|\mathbf{\Sigma}_i|$ cannot be taken out from $p_i f_i(\mathbf{x})$ and the dicriminant score becomes

$$\delta_i^Q(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| + \log p_i, \ i = 1, 2, ..., g.$$

We assign the observation to the class for which $\delta_i^Q(\mathbf{x})$ is largest.

Unlike the earlier LDA expression, it shows from this equation that the covariance matrix is dependent on which class k that is considered. Also, observations will enter in a quadratic manner. Hence, in QDA, a covariance matrix needs to be estimated for each class. To accomplish this, there must be a sufficient number of training observations within each class, allowing for a reliable estimation. The estimated covariance matrix can be found by

$$\hat{\boldsymbol{\Sigma}}_i = \frac{1}{n_j}\sum_{j=1}^{n_j}(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)^T.$$

The reason why there could be an advantage of choosing QDA instead of LDA is because of the bias-variance trade-off. As QDA calculate a covariance matrix for each class ($\boldsymbol{\Sigma}_i$), there are $gp(p+1)/2$ parameters to estimate for the covariance matrices, while as for LDA there are only $p(p+1)/2$. For multiple classes these calculations in QDA becomes larger. However, that also makes QDA much more flexible than LDA. There are advantages and disadvantages to consider when deciding which method to choose. As LDA assume g classes share a common covariance matrix, LDA can suffer from high bias. In essence, LDA tends to perform more favorably than QDA when dealing with a limited number of training observations, prioritizing the reduction of variance. If the training data set is relatively large, diminishing the impact of classifier variance, or assumption of distinct covariance matrices, QDA is recommended.

Figure 5 display a possible outcome from utilitizing QDA and LDA on a simulated data set. It really shows the diversity within the methods and the flexibility of QDA.

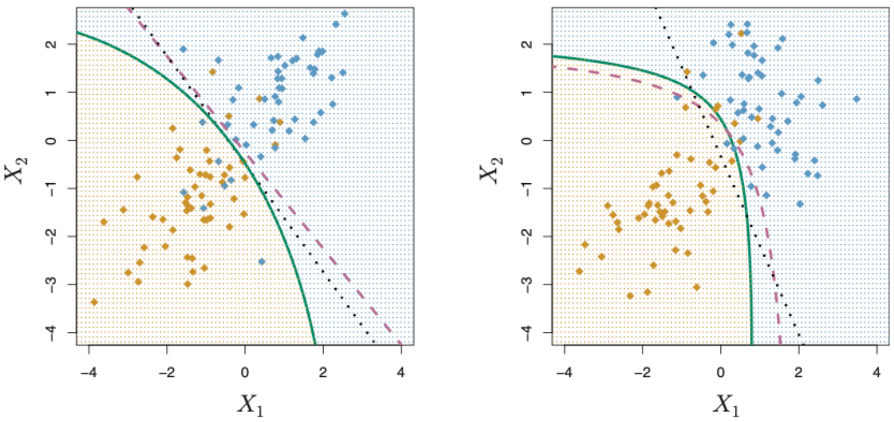


Figure 5: How the QDA method operates. The green line is QDA, the dotted black is the LDA and the dashed read is Bayes Classifier where the distribution is known.

Source: Page 150 in [11]

2.2.4 KDC

Kernel Density Clustering (KDC) is based on two things, Kernel Density Estimation (KDE) and then clustering with a classifier. The method focused on in this thesis is the Naive Bayes Classifier.

First, let us focus how the Kernel Density Estimation operates. KDE estimates the PDF of each data point [23]. Numerous investigations aim to estimate $f(x)$ using a sample of observations $x_1, x_2, ..., x_n$. The relationship between the probability distribution and the PDF can be defined as

$$P(a \leq X \leq b) = \int_a^b f(x)dx,$$

where X is a continuous-valued random variable and $f(x)$ is the PDF.

The parametric approach assumes that the PDF of a data set belongs to a specific parametric family of distributions. However, the main drawback of this approach is its lack of flexibility, as it may not accurately capture the true underlying distribution. When flexibility and adaptability are essential, a non-parametric approach is favored, as it allows for more adaptable and data-driven modeling of the distribution.

To estimate the PDF from data when the underlying distribution is unknown, it is necessary to have both a weight function and a kernel. These components are essential for conducting KDE. The weight function specifies how each data point influences the estimation, while the kernel dictates the smoothing applied to the data, shaping the approximation of the PDF. This function tells us the degree of smoothness of the PDF.

First, we will examine KDE and the process of establishing the weight function. From the definition of the PDF, denoted as f(x), for a random variable X, it follows that

$$P(x-h < X < x+h) = \int_{x-h}^{x+h} f(t)dt \approx 2hf(x) \;\Rightarrow\; f(x) \approx \frac{1}{2h}P(x-h < X < x+h).$$

By examining the relative frequency, $f(x)$ can be approximated

$$\hat{f}(x) = \frac{1}{2h}\frac{\text{data samples in } (x-h, x+h)}{n},$$

which can be rewritten as

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} w(x - x_i, h).$$

In this case, the observed values are defined as $x_1, x_2, ..., x_n$. Also, we can now observe from the expression there has appeared a weight function as

$$w(t,h) = \begin{cases} \frac{1}{2h} \text{ for } |t| < h, \\ 0 \text{ otherwise.} \end{cases}$$

There are various methods to articulate the weight function, but they all ultimately trace back to the form of

$$w(t,h) = \frac{1}{h} K \left(\frac{t}{h} \right).$$

Here, K, referred to as the kernel, is a function that depends on a single variable. The kernel plays a large role in the shape of the weight function. While the bandwidth parameter, also called smoothing constant, is presented as h. The bandwidth determines the amount of smoothing to apply to the estimated PDF. The combination of the bandwidth and the kernel estimation produces the properties of $\hat{f}(x)$.

Some simple kernels are the Triangular and Rectangular kernels, respectively $1 - |t|$ for $|t| < 1, 0$ otherwise and $\frac{1}{2}$ for $|t| < 1, 0$ otherwise. The most common kernel is the Gaussian kernel, which is given as $K(t) = \frac{1}{\sqrt{2\pi}} e^{-(1/2)t^2}$. Generally, any function that possesses the following properties can serve as a kernel:

- $\int K(z)dz = 1$
- $\int zK(z)dz = 0$
- $\int z^2 K(z)dz := k_2 < \infty$

Figure 6 show plots from an example in [23] of kernel estimates with a Gaussian kernel and various bandwidth sizes.

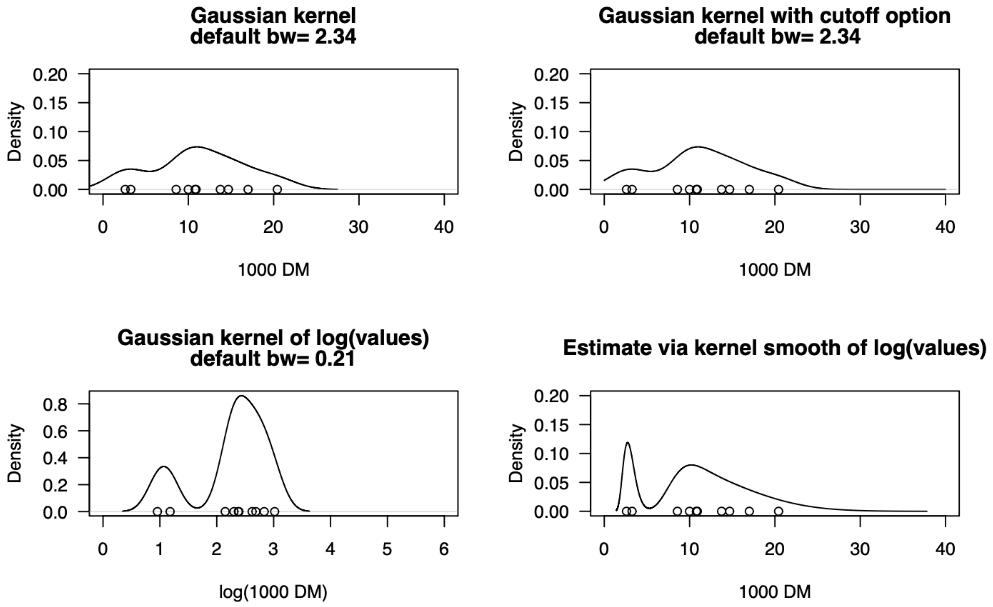


Figure 6: Example of kernel estiamtes with a Gaussian kernel and various inputs for the bandwidth.

Source: page 10 in [23]

The aim is to identify the optimal bandwidth and kernels, determined through an assessment of their efficiency based on the Measured Integrated Squared Error (MISE),

$$\begin{gathered}\text{MISE}(\hat{f}) = E\int_{-\infty}^{\infty}(\hat{f}(x)-f(x))^2dx \\ = \int_{-\infty}^{\infty}\left[E(\hat{f}(x)-f(x))^2\right]dx = \int_{-\infty}^{\infty}\text{MSE}(\hat{f}(x))dx \\ = \int_{-\infty}^{\infty}\left[\text{Bias}^2\hat{f}(x)+\text{Var}(\hat{f}(x))\right]dx = \int_{-\infty}^{\infty}\text{Bias}^2\hat{f}(x)dx + \int_{-\infty}^{\infty}\text{Var}(\hat{f}(x))dx.\end{gathered}$$

This is a measure of the global accuracy of $\hat{f}(x)$, indicating the quality of the density estimator. A more detailed calculation of bias and variance can be found in [23].

2.2.5 Support Vector Machine

The Support Vector Machine (SVM) is a supervised machine learning algorithm and is a well known method within classification. This method is referred to as the *maximal margin classifier*, because of its simple and intituive classification style [11]. The primary goal of the SVM algorithm is to identify the optimal hyperplane within a p-dimensional space, which can effectively segregate data points across various classes within the feature space [10]. A p-dimensional hyperplane with a point X and beta parameters is defined as

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p = 0.$$

Now, imagine there are points based on n training and p dimensions

$$x_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{1p} \end{pmatrix}, ..., x_n = \begin{pmatrix} x_{n1} \\ \vdots \\ x_{np} \end{pmatrix}.$$

Then the separating hyperplane classifies after

$$f(\mathbf{x}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} > 0 \text{ if } y_i = 1$$

and

$$f(\mathbf{x}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} < 0 \text{ if } y_i = -1,$$

for all $i = 1, ..., n$.

The support vectors represent the closest points to this hyperplane, and consequently, they exert significant influence on determining the margin. The margin is defined as the distance between the support vectors and the hyperplane. The primary goal of the support vector machine algorithm is to maximize this margin because a wider margin signifies improved classification performance. All these elements are depicted in Figure 7, where the support vectors represent observations closest to the hyperplane, and the margin illustrates the distance extending outward from the hyperplane.

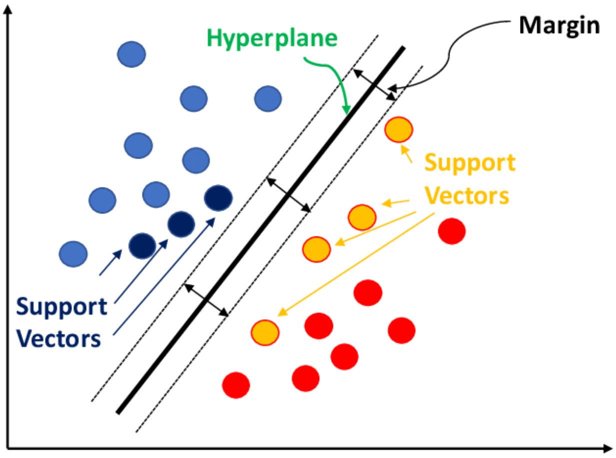


Figure 7: Visual representation of the support vectors, hyperplane, and margin.

Source: [9]

The distance between the hyperplane and the support vectors can be determined by

$$d_i = \frac{\mathbf{x}^T \beta_i + \beta_0}{||\mathbf{x}||},$$

where the denominator indicates the Euclidean norm of $\mathbf{x}$ [10].

An example of how a separating hyperplane can act is shown in Figure 8. On the left side, numerous potential hyperplanes capable of distinguishing between the two classes are depicted. Meanwhile, the right-hand side illustrates the specific hyperplane that the classifier ultimately selects.

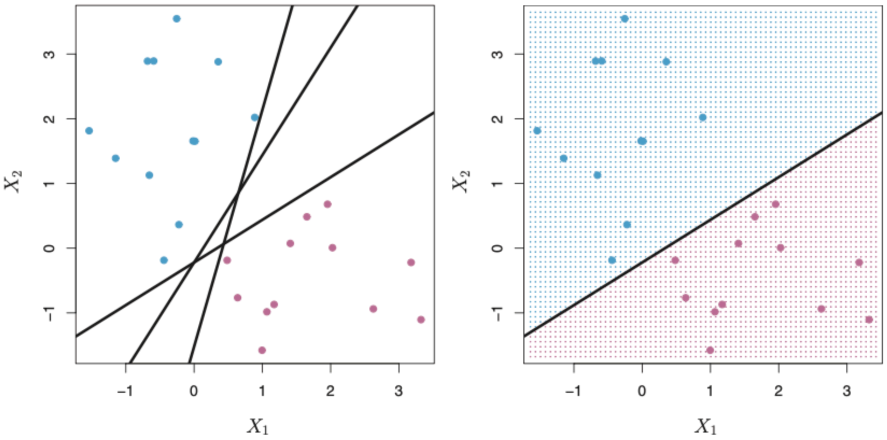


Figure 8: Separating hyperplanes in SVM.

Source: page 340 in [11]

We opt for the hyperplane that maximizes the distance from itself to the nearest data point on each side. If such a hyperplane can be found, it is termed the maximum-margin hyperplane or hard margin. However, in cases where it is impossible to completely separate the classes due to outliers, the SVM seeks to find the maximum margin while imposing a penalty each time a data point crosses the margin. In such situations, these margins are referred to as soft margins [10]. Therefore, there can be two different optimization problem. First, for har margin linear SVM classifier

$$\min_{\mathbf{x},\beta_0} \frac{1}{2}\mathbf{x}^T\mathbf{x} = \min_{\mathbf{x},\beta_0} \frac{1}{2}||\mathbf{x}||^2$$

$$\text{subject to } f_i(\mathbf{x}^T\boldsymbol{\beta} + \beta_0) \geq 1 \text{ for } i = 1, 2, ..., n.$$

Second, the optimization problem of a soft margin will look like

$$\min_{\mathbf{x},\beta_0} \frac{1}{2}\mathbf{x}^T\mathbf{x} + C\sum_{i=1}^{n} \zeta_i$$

$$\text{subject to } f_i(\mathbf{x}^T\boldsymbol{\beta} + \beta_0) \geq 1 - \zeta_i \text{ and } \zeta_i \geq 0 \text{ for } i = 1, 2, ..., n.$$

In this context, the term $C \geq 0$ denotes the cost parameter, which regulates the penalty obtained for misclassifying an observation. Additionally, ζ_i signifies the permissible slack allowed for the given problem.

2.2.6 Random Forests

A classification tree predicts for each observation to be in the same class as most other training observations in its region. The trees grows by using recursive binary splitting. This splitting method is a top-down, greedy approach. Meaning, it begins from the top of the tree and works its way down, splitting by the best possible choice in that particular step. Figure 9 displays how a decision tree can be structured.

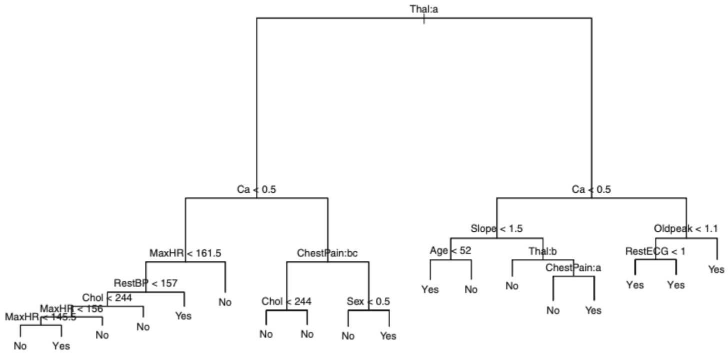


Figure 9: Decision tree build on a recursive binary splitting approach.

Source: page 313 in [11]

Random Forests construct their decision trees using bootstrapped training samples. During the process of building these decision trees, a random subset of n predictors is selected as candidate predictors for each split, chosen from the complete set of p predictors [11]. Typically, this random sample is chosen as $n \approx \sqrt{p}$. The advantage of using random forests over other decison tree methods primarily attributed to the introduction of randomness. In many cases, when a strong predictor dominates, most decision tree methods tend to select it as the top split, resulting in very similar trees. Random forests, by limiting each split to involve only a subset of predictors, creates diversity in the collection of trees, diminishing the likelihood of depending too heavily on a single dominant predictor.

When a Random Forest is used for classification, it takes a subset of data points and constructs an individual decision tree for each sample. Each tree generates its own prediction, and the chosen class for the overall classification is the one that appears most frequently across all the trees [15].

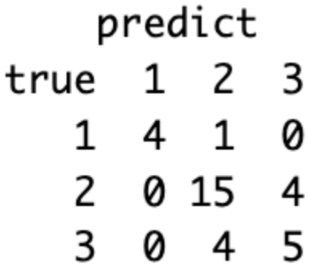
2.2.7 Accuracy

For the classification accuracy, the response is no longer a numerical value. One can simply find the accuracy rate by comparing the similarity between the training and test observations.

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i),$$

where y_i is the actual class and $\hat{y}_i$ is the predicted class. $I(y_i = \hat{y}_i)$ is an indicator variable and says if the prediction has been classified correctly. This means that when the actual and the predicted class equals each other, the indicator variable equals 1. The sum of correct placements over the amount of observations considered, gives the accuracy rate of the classification model. The accuracy rate is the opposite of the error rate, where the amount of misplaced classes are in focus instead.

A confusion matrix provides an effective output to present accuracy information. A matrix representing the class sizes will be shown, with correctly placed classes displayed along the diagonal. Figure 10 shows a very simple confusion matrix. Here, the columns are the predicted class and the rows are the true class of the observation.



```
       predict
true  1  2  3
   1  4  1  0
   2  0 15  4
   3  0  4  5
```

Figure 10: An example of a confusion matrix.

2.3 Transformation

Several machine learning algorithms assumes that the given data set has a Gaussian distribution, though this is often not the case for real data. Therefore, the data often needs to be preproccesed before used. Often a data set gives us more information if it is transformed somehow before being used. This will often make the machine learning algorithm perform better. Even algorithms that do not expect variables to have Gaussian distribution, often perform better when data are close to Gaussian.

2.3.1 Standarization

Standarization is a very usual data preprocessing method. This method tranforms the variables to have a mean of 0 and a standard deviation of 1. This is done by subtracting the sample mean from each observation and dividing the variable by its standard deviation [17]. This is beneficial because it gives the data set the same start construction.

2.3.2 Box-Cox

The Box-Cox transformation changes the output variables so that they are close to having a Gaussian distribution. This means that the method stabilize the variation of the distribution. Box-Cox transformation is a method that helps choosing a good lambda to transform the data set.

$$\begin{cases} \frac{x^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0, \\ \log(x) & \text{if } \lambda = 0 \end{cases}$$

The lambda ranges between 0 and 1, and indicates what type of transformation should be done with the data set. For each column in the data set, it might be a different optimal lambda. This helps us stabilize the variables with a zero mean and a standard deviation of 1. Often, this tranformation leads to a more favorable outcome.

The Box-Cox transformation can find a specific, numerical, optimal lambda, and it can visualize a range of values where the lambda can be positioned, especially from the most used transformations. An example of this visualization is shown in Figure 11. The most common Box-Cox transformations, written about in [2], are

$$\begin{cases} \lambda = -1 & \Rightarrow \text{ reciprocal transform,} \\ \lambda = -0.5 & \Rightarrow \text{ reciprocal square root transform,} \\ \lambda = 0 & \Rightarrow \text{ log transform,} \\ \lambda = 0.5 & \Rightarrow \text{ squareroot transform,} \\ \lambda = 1 & \Rightarrow \text{ no transformation.} \end{cases}$$

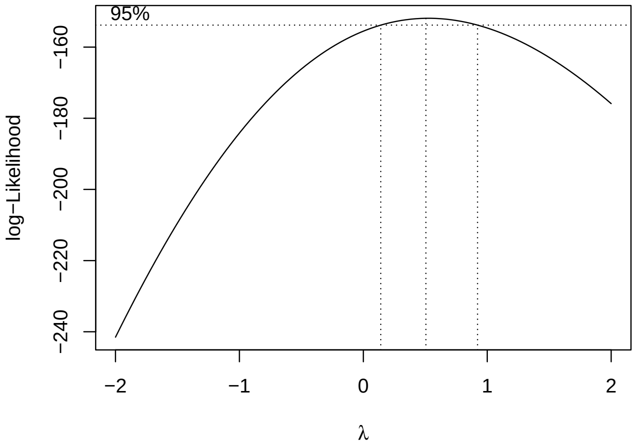


Figure 11: Shows how a Box-Cox transformation plot can look.

2.4 Clustering

Clustering is an approach that organizes data exhibiting comparable patterns into cohesive groups. This allows us to reduce the number of models required, as opposed to creating an individual model for each data point. Clustering analysis is an example of unsupervised machine learning, eliminating the necessity to label input and output values for the training data [19]. Two very common methods are K-means and Hierarchical Clustering. Other methods worth mentioning include Self-organising Map (SOM), Fuzzy K-means (FKM) and Support Vector Clustering (SVC), where further information about all of these methods can be found in [16].

A clear distinction between classification and clustering to note is the knowledge of the amount of groups. As written earlier, classification knows the number of classes and assigns new observations to these already made classes [12]. Clustering, on the other hand, take no assumptions of amount of groups or their structure. This method forms groups based on the similarities of the data.

2.4.1 K-means

K-means stands as one of the most frequently utilized clustering techniques, designed to minimize the pairwise distance between data points within each cluster [19].

The observations $x_1, ..., x_n \in R^d$ are divided into K clusters $C_1, ..., C_k$, which are created based on the objective

$$\min_{\substack{C_1,...,C_K \\ \mu_1,...,\mu_K}} \sum_{j=1}^{K} \sum_{i \in C_j} ||x_i - \mu_j||_2^2, \tag{7}$$

where $\mu_l = \frac{1}{|C_l|} \sum_{i \in C_l}^{n}$. The cost function, $|C_l|$, is an average of the variance for each cluster, taking into account their respective sizes.

The K-means clustering method entails selecting a specified number (K) of clusters, with the initial center points of each cluster determined randomly. The number of clusters can be predefined or established using cluster analysis. Each data point is then assigned to the cluster with the nearest center, as determined by the Euclidean distance formula:

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2},$$

which means that observations are placed in the cluster with the closest mean. Subsequently, the variance for the clusters and their observations is computed and saved. This process is iterated multiple times, and the final result is the cluster configuration with the smallest overall variance within the clusters. This is accomplished by minimizing the total within-cluster variation through the calculation of the squared sum of the Euclidean distance between each observation and its associated cluster center.

Lloyd's algorithm, often referred to as the K-means algorithm, is a potential approach for addressing the K-means objective. This algorithm aims to minimize the following expression:

$$\min_j ||x_i - \mu_j||_2^2.$$

In this expression, x_i represents the data points, and μ_j represents the cluster centers. The algorithm iteratively updates the cluster assignments and the cluster centers to minimize the overall distance between data points and their assigned cluster centers. More of this algorithm and possible withdraws with the K-means approach can be found in [16].

Figure 12 illustrates the appearance of a clustering plot after applying a K-means approach to a set of observations. The figure displays the mean hourly consumption for five distinct clusters derived from a three-month time interval. It underscores differences between these clusters, showcasing variations in daily consumption patterns among different households.

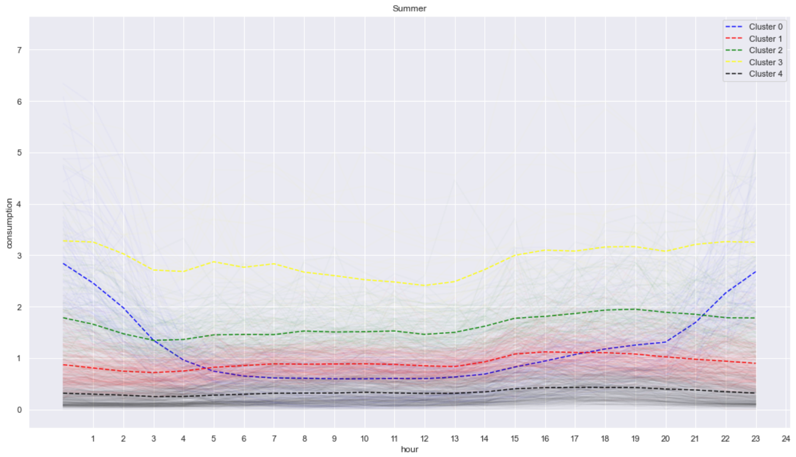


Figure 12: A visualization of a K-means clustering approach from the project thesis.

Source: [16]

2.4.2 Hierarchical Clustering

The hierarchical clustering technique is a widely used and straightforward method for clustering. It operates through a sequence of either successive mergers or successive divisions [12]. Initially, an equal number of clusters is created, matching the number of objects, and these clusters are formed based on the similarities among the objects. As the similarities decrease, subclusters are gradually merged into a single cluster. The final clustering result is visualized in a dendrogram, a tree-structured graph, with different colors representing larger subclusters. Figure 13 shows an example of how a dendrogram can appear after applying hierarchical clustering to a set of observations.

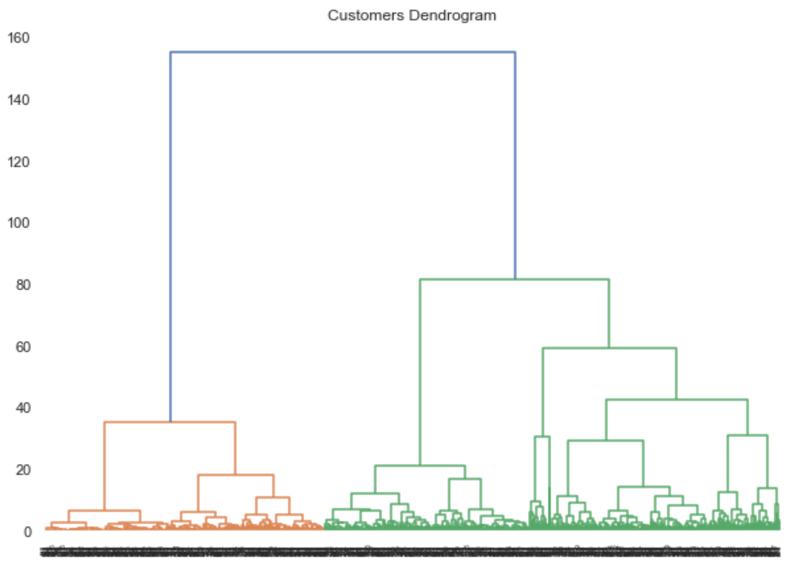


Figure 13: A dendogram created from a hierarchical clustering approach from the project thesis.

Source: [16]

Hierarchical clustering relies on linkage methods to determine cluster distances and subsequently group the objects accordingly. These linkage methods are criteria based on distance measurements to identify the nearest neighbors for each object. The various methods offer different approaches to calculating distances between observations, both within and outside the clusters. Some commonly used methods for hierarchical clustering are single linkage, complete linkage, average linkage and Ward linkage. For more knowledge of these methods, see [16].

2.5 Regression Analysis

Regression analysis is a statistical approach used for building relationships between one or more dependent variables and a set of predictor independent variable values [12]. The relationships can be used for explainations or predictions.

2.5.1 Linear and Non-Linear Models

A linear regression model is of the form

$$y = \beta_0 + \beta_1 z_1 + ... + \beta_r z_r + \epsilon. \tag{8}$$

Here, y is the response, while the set of $\boldsymbol{\beta}$ are the unknown paramters and $z_1, z_2, ..., z_r$ are r predictor variables meant to be associated to y. The random error variable is expressed as $\epsilon \sim N(0, \sigma^2)$ and is assumed to be independent. The word linear comes from the mean being a linear function of $\boldsymbol{\beta}$ [12].

A model is regarded non-linear when the outcome is not a linear function in the unknown parameters. A non-linear model may be expressed as followed

$$y = f(z, \beta) + \epsilon.$$

2.5.2 Generalized Linear Models

Generalized linear models (GLM) are designed to handle univariate response variables that can be modeled using distributions from the exponential family [8]. This method is widely employed in statistics and serves as a robust tool for constructing predictive models. Distributions within the exponential family encompass gamma, binomial, negative binomial, exponential, Poisson, Gaussian, inverse normal, and geometric distributions. In contrast to linear models, GLMs introduce non-linearity, offering flexibility that can be advantageous for fitting diverse types of data. While linear models assume that the response variables are aligned with the explanatory variables effects under the conditions of normality of errors and variance homogeneity, these assumptions are not mandatory for GLMs. Let Y_i for $i = 1, ..., n$ denote the dependent response variables and let $\mathbf{x}_i = (1, x_{i1}, ..., x_{ik})^T$ represent the explanatory variables. GLMs comprise three components: the random, the systematic, and the link component. These elements are integrated into the GLMs' linear predictor, response variable distribution, and link function.

Members of the exponential family exhibit probability density functions expressible as

$$f(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}, \tag{9}$$

where $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ represent functions. In this context, θ denotes the canonical parameter, and ϕ is the dispersion parameter. There are three components that build the GLM and they are as followed [3, 8]:

1. *The Random Component*: This component comes from equation 9 and is the probability distribution of the response variable Y_i for $i = 1, ..., n$.

2. *The Systematic Component*: comes from the quantity

   $$\eta_i(\boldsymbol{\beta}) = \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik},$$

   where $\eta_i(\boldsymbol{\beta})$ is referred to the linear predictor and $\boldsymbol{\beta} = (\beta_0, ..., \beta_k)^T$ are unknown parameters.

3. *The Parametric Component*: The link function comes from the relationship between the mean μ_i of y and the linear predictor η_i found in step 2. The link function is written as

   $$g(\mu_i) = \eta_i(\boldsymbol{\beta}) = \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

2.5.3 Distributions

The normal distribution, also called the Gaussian distribution, is one of the most common and used distribution. When employing the normal distribution in a GLM, no transformations are needed, resulting in a standard linear model [8]. In this context, the use of the normal distribution implies that the response variable is assumed to have a normal distribution and have a linear relationship with the predictor variables.

Consider equation 8 with $\epsilon \sim N(0, \sigma^2)$, then the density of the response variable y can be written in exponential family form as

$$f(y_i; \mu_i, \sigma) = \exp\left\{ \frac{y_i\mu_i - \frac{\mu_i^2}{2}}{\sigma^2} - \frac{1}{2}\left[\ln(2\pi\sigma^2) + \frac{y_i^2}{\sigma^2} \right] \right\}.$$

Then, when considering equation 9, it follows that $\theta_i = \mu_i$ and $\phi = \sigma^2$ [3, 8]. Thereby, the Gaussian distribution has the identity function as link function, meaning

$$g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} = \mu_i.$$

The Gamma distribution is a continuous probability distribution and belongs to the exponential family. For a given $Y_i \sim \text{Gamma}(\mu_i, \nu)$, we have that $E(Y_i) = \mu_i$ and $\text{Var}(Y_i) = \frac{\mu_i^2}{\nu}$. The gamma distribution is characterized by its probability density function, which is defined for Y_i as

$$f(y_i; \mu_i, \nu) = \frac{1}{\Gamma(\alpha, \beta)} \left(\frac{y_i}{\beta} \right)^{\alpha} e^{-\frac{y_i}{\beta}} \frac{1}{y_i}, \tag{10}$$

where $y_i \geq 0, \alpha = \nu > 0, \beta = \frac{\mu_i}{\nu} > 0$.

The normal distribution and the gamma distribution have different domains. The gamma distribution exclusively selects random variables within the range $[0, \infty)$, whereas the normal distribution considers variables from the entire range $(-\infty, \infty)$.

The gamma distribution in GLM typically employs either the canonical link or the log link. The log link is a common choice because, unlike the canonical link, it ensures that estimated responses are not negative. The canonical link function for the gamma distribution is an inverse power function, allowing the model of the mean to be expressed as

$$\frac{1}{\mu} = \mathbf{x}^T \boldsymbol{\beta}. \tag{11}$$

2.5.4 Maximum Likelihood Estimation

The regression parameters from model, $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_r)^T$, are made through maximum likelihood estimation (MLE). The log likelihood in a GLM can be defined as

$$l(\boldsymbol{\beta}, \sigma^2) := \log(L(\boldsymbol{\beta}, \sigma^2)) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta}, \sigma^2),$$

where $\mathbf{y} = (y_1, ..., y_n)^T$ is the observed data of $\mathbf{Y}$ and $l_i(\boldsymbol{\beta}, \sigma^2)$ denotes the log likelihood for observation y_i. To obtain the maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, the unscaled score equations need to be solved. Numerical solutions are required for these equations since they are nonlinear in $\boldsymbol{\beta}$, and this entails the use of iterative algorithms [8].

2.5.5 MAE and MAPE

Mean Absolute Error (MAE) is the absolute error between the predicted response and the actual value and is given as

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|.$$

The MAE value will serve as an indicator of the accuarcy of the prediction model.

The expression for the Mean Absolute Percentage Error (MAPE) is given by

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|.$$

The MAPE is a valuable metric because it presents the error as a percentage, providing a more intuitive and easily interpretable measure compared to dealing with large numerical values [13].

Given that Tibber aims to minimize the MAE, both the MAE and MAPE are employed as the preferred error estimates in the subsequent analysis.

3 The Data Set

In this section, we initially outline the process of data collection and provide an introduction of the explanatory variables used in the analysis. Following this, we visualize the correlation between the variables and address the need for data preprocessing.

3.1 Data Description

Tibber has provided the used data set in the thesis. Tibber serves numerous customers across the Nordic countries, each falling into different price areas according to Nord pool [16]. For simplicity's sake, only the SE3 price area has been utilized. SE3 represents a price area in Sweden, encompassing regions such as Stockholm, see Figure 14. The decision of only using the swedish customer data was made to focus on developing effective models with a smaller data set initially, with the possibility of expanding the data in future projects.

The data set consists primarily of two dataframes. The first dataframe includes home-ids, serving as unique identifiers for each home, along with corresponding property information that provides specific details about the contents of each home. The second dataframe is a time series data set, with each home-id having approximately one year of hourly data, that provides a wealth of information about the surroundings and the specific events occurring at each given hour. This chapter provides insight into how the data set is constructed through the use of display and visualization techniques.

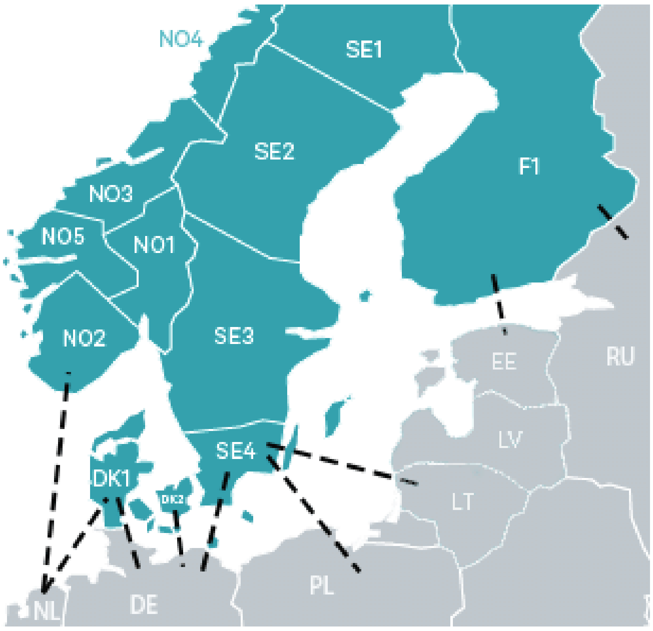


Figure 14: An image displaying the different price areas in Scandinavia.

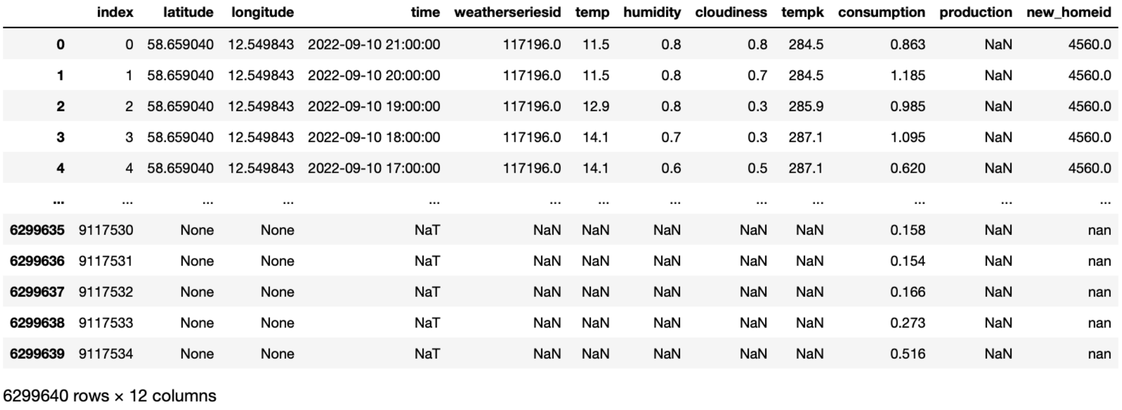
Source: [1]

3.2 Explanatory Variables

The two data sets given are one time series data set and one properties data set, displayed in Figure 15 and 16. The time series data set was used in the project thesis to cluster the home-ids after their consumption, while the master thesis placed greater emphasis on the properties data set, supplemented by the inclusion of the time series data.

Figure 15 represents an hourly time series data set comprising 12 variables. The original data set consisted of 9 117 535 rows. In the present context, a smaller dataframe is shown, resulting from the removal of rows containing NaN values for consumption. This step was taken due to the clustering process relying on hourly consumption values. The home-id variable serves as a unique identifier for each customer's home. For each hour, the data set includes data on home consumption, given in kWh, and production, considering that some households may have solar panels, enabling production. The weatherseries-id variable is linked to weather forecasts, providing information about humidity, cloudiness, and temperature in both Celsius and Kelvin for that specific time. It is worth noting that both humidity and cloudiness are constrained to a range of 0 to 1, allowing for selection from only five discrete values: 0, 0.2, 0.4, 0.6, 0.8, and 1.

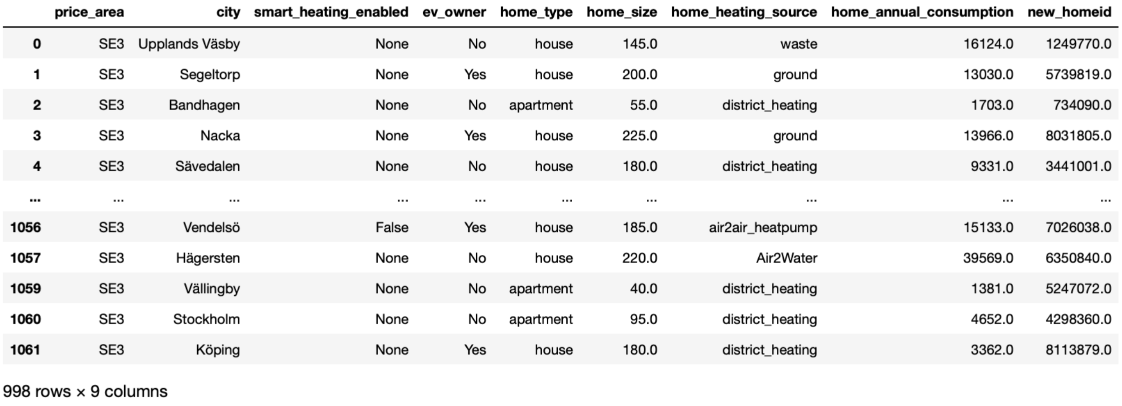
The home properties data set is visualized in Figure 16. The initial data set consisted of 141 108 rows. During the preprocessing stage, several rows were removed due to the presence of duplicates. Furthermore, in consideration of the analysis, rows with NaN values in the home_annual_consumption and home_size columns were excluded. However, the primary factor contributing to the substantial decrease in the home properties data set occurred due to the evaluation of home-ids connected to the time series data. Due to this step, numerous homes that were unsuitable for the further analysis were removed. This dataframe is also associated with the home-id variable and provides additional insights into household properties. The home properties data set includes details like the presence of an electric vehicle (EV), home size, residence type, heating source, and annual consumption. Moreover, the smart_heating_enabled variable indicates whether the household utilizes the Tibber app and its smart products, which adjust the home's heating based on occupancy [21]. The properties data set is a mixture of continuous and categorical variables. The continuous variables are variables as home-size and home-annual-consumption, where we have different numbers based on their size. Examples of categorical variables are type of home as "apartment" and "house". The distinction between these variables has played a crucial role in the thesis experimentation.



	index	latitude	longitude	time	weatherseriesid	temp	humidity	cloudiness	tempk	consumption	production	new_homeid
0	0	58.659040	12.549843	2022-09-10 21:00:00	117196.0	11.5	0.8	0.8	284.5	0.863	NaN	4560.0
1	1	58.659040	12.549843	2022-09-10 20:00:00	117196.0	11.5	0.8	0.7	284.5	1.185	NaN	4560.0
2	2	58.659040	12.549843	2022-09-10 19:00:00	117196.0	12.9	0.8	0.3	285.9	0.985	NaN	4560.0
3	3	58.659040	12.549843	2022-09-10 18:00:00	117196.0	14.1	0.7	0.3	287.1	1.095	NaN	4560.0
4	4	58.659040	12.549843	2022-09-10 17:00:00	117196.0	14.1	0.6	0.5	287.1	0.620	NaN	4560.0
...	...	...	...	...	...	...	...	...	...	...	...	...
6299635	9117530	None	None	NaT	NaN	NaN	NaN	NaN	NaN	0.158	NaN	nan
6299636	9117531	None	None	NaT	NaN	NaN	NaN	NaN	NaN	0.154	NaN	nan
6299637	9117532	None	None	NaT	NaN	NaN	NaN	NaN	NaN	0.166	NaN	nan
6299638	9117533	None	None	NaT	NaN	NaN	NaN	NaN	NaN	0.273	NaN	nan
6299639	9117534	None	None	NaT	NaN	NaN	NaN	NaN	NaN	0.516	NaN	nan

6299640 rows × 12 columns

Figure 15: Given time series data in pandas format.



	price_area	city	smart_heating_enabled	ev_owner	home_type	home_size	home_heating_source	home_annual_consumption	new_homeid
0	SE3	Upplands Väsby	None	No	house	145.0	waste	16124.0	1249770.0
1	SE3	Segeltorp	None	Yes	house	200.0	ground	13030.0	5739819.0
2	SE3	Bandhagen	None	No	apartment	55.0	district_heating	1703.0	734090.0
3	SE3	Nacka	None	Yes	house	225.0	ground	13966.0	8031805.0
4	SE3	Sävedalen	None	No	house	180.0	district_heating	9331.0	3441001.0
...	...	...	...	...	...	...	...	...	...
1056	SE3	Vendelsö	False	Yes	house	185.0	air2air_heatpump	15133.0	7026038.0
1057	SE3	Hägersten	None	No	house	220.0	Air2Water	39569.0	6350840.0
1059	SE3	Vällingby	None	No	apartment	40.0	district_heating	1381.0	5247072.0
1060	SE3	Stockholm	None	No	apartment	95.0	district_heating	4652.0	4298360.0
1061	SE3	Köping	None	Yes	house	180.0	district_heating	3362.0	8113879.0

998 rows × 9 columns

Figure 16: Given property data in pandas format.

In consideration for the privacy security for the customers, a ”new_homeid” as unique identiy for the customers was added. This identifier also serves a future purpose for Tibber, ensuring that it does not interfere with or alter the information they already possess. See Table 1 and 2 for a description of the time series data and the home properties data, respectively.

Table 1: Explanation of variables from the time series data

Variable Name	***Explanation***
new home-id	The unique home-id to a household
latitude	North-south coordinate
longitude	East-west coordinate
time	Date and hourly time
weatherseries-id	Unique id for a geographic area connected to a weather forecast
temp	The hourly temperature in Celsius
humidity	The amount of water vapor in the air
cloudiness	How cloudy the sky is
tempk	The hourly temperature in Kelvin
consumption	The amount of electricity that is used in an hour
production	The amount of electricity that has been self-produced in an hour

Table 2: Explanation of variables from the home properties data

Variable Name	***Explanation***
price_area	Geographic area with specific electricity price
city	The city where the household is located
smart_heating_enabled	If the home use smart products to control the heating in the house
ev_owner	If the household contains an electric vehicle
home_type	What type of household it is, e.g. house, apartment, etc.
home_size	The size of the household
home_heating_source	Which type of source the household uses for heating, e.g. district_heating, electricity, etc
home_annual_consumption	The yearly electricity consumption of the household
cluster	This column is added afterward, calculated from the electricity consumption data obtained from the time series data

3.3 Response

The primary aim of this master thesis is to effectively classify new households into groups established by clustering the hourly consumption of the old households. The response variable for this classification task is represented by these clusters, which categorize customers based on their hourly consumption into five distinct groups. The established clusters were incorporated into the dataframe as a distinct column, each associated with its corresponding home-id. When classifying new customers, the cluster column serves as the target variable. Given that we are utilizing a classification model, our response variable is categorical. The properties associated to the new customers do not have a specific meaning by themselves. Instead, their importance lies in determining the group to which they belong. Consequently, we are working with categorical variables.

3.4 Visualization

Figure 17 shows a correlation plot featuring three variables derived from the home properties data. The plot illustrates a correlation at 0.67 between home_size and home_annual_consumption, two variables that have seen substantial usage in this study. The variable ev_owner has relatively lower correlations, with values of 0.27 for home_size and 0.20 for home_annual_consumption.

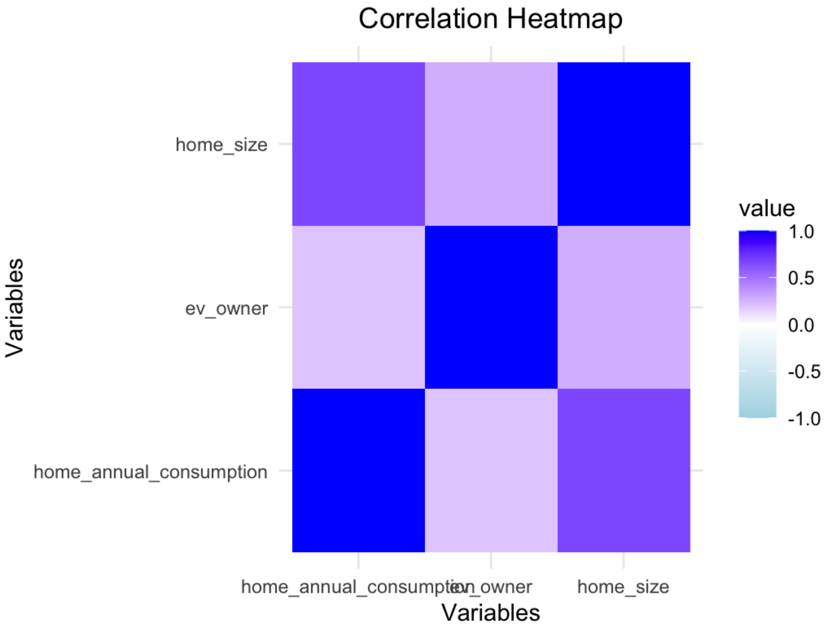


Figure 17: Correlation heatmap of 3 numeric variables from the data properties

Figure 18 also illustrates a correlation heatmap for four categorical variables. In this representation, the binary variable ev_owner is transformed from yes/no to 1/0, while the remaining variables are converted into dummy variables. This highlights the challenge of dealing with categorical variables, particularly when the explanatory variables have numerous categories. The resulting matrix becomes large, making it difficult to conduct a meaningful analysis of the correlations due to the multitude of categories within each explanatory variable.

Some of the outliers that stand out, there is a notable correlation of 0.70 between home_heating_sourcedistinct_heating and home_typeapartment. Furthermore, an association is observed between EV owners and customers with houses, as indicated by a correlation of 0.27. Additionally, it is worth noting that explanatory variables display expected negative correlations, as there is a -0.65 correlation between house and apartment, reflecting the differences in home types. Several of the variables demonstrate a tendency to become nearly uncorrelated. Notably, the smart_heating_enabled variables stays in a range between +/- 0.2, except for the negative correlation between true and false for the variable itself. A similar pattern is observed among many of the home_heating_source variables.

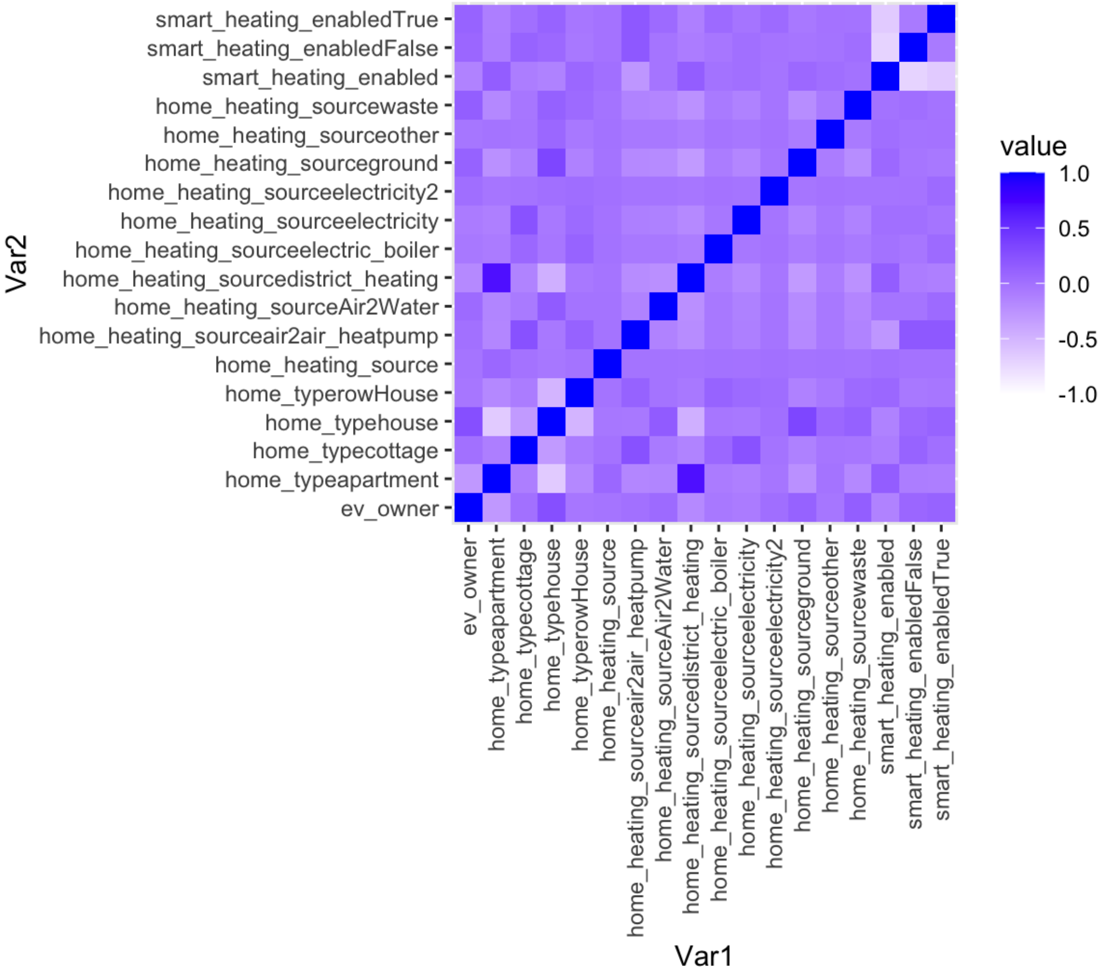


Figure 18: Correlation heatmap of 4 categorical variables from the data properties after being transformed to dummy variables.

Figure 19 shows again a correlation heatmap including ev_owner, home_type, home_size, smart_heating_enabled, and home_annual_consumption. One observation worth noting is that the home type "house" exhibits a slightly higher correlation with both the size of the home at 0.61 and the annual consumption of electricity at 0.56. This correlation is logical, considering that houses are typically larger than other types of homes, leading to a higher electricity consumption. In contrast, the home type ”apartment” demonstrates a negative correlation at -0.56 with annual consumption, indicating that this particular home category likely consumes less electricity. It is also notable that the three variables with the strongest correlation are consistent with those highlighted in the first correlation plot (Figure 17).

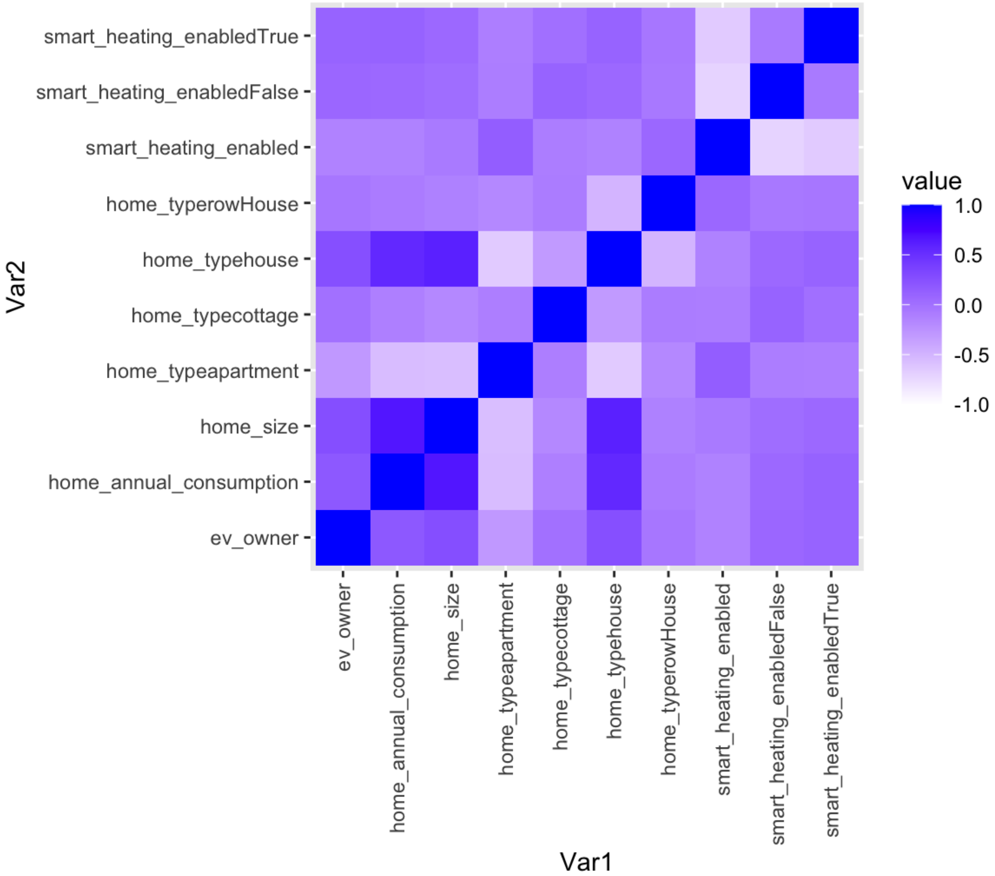


Figure 19: Correlation heatmap of 2 categorical, 1 binary and 2 continous variables from the data properties after being transformed to dummy variables.

3.5 Data Pre-Processing

Before utilizing the provided data sets and those created from the project task in classification methods, a necessary step involves data pre-processing. Data pre-processing involves manipulating raw data to make them accessible and ready to analyse. Typically, data sets contain numerous missing values (NaN) which must be addressed. This is a common problem in data analysis with real data. The most important aspect is to understand the disadvantages and find ways to navigate around the issues.

3.5.1 Missing values

The collected data contain missing values, as is evident in both Figure 15 and Figure 16, where these are a substantial number of NaN values in certain columns. Below, we provide visual representations of the missing data in the data set. These missing values pose a challenge for the classification task, as we aim to assign homes to clusters based on their properties. Without the home_annual_consumption information, classifying a household becomes quite difficult. In some cases, like with the smart_heating_enabled variable, it might be possible to find workarounds, introducing a third category instead of just a binary choice. As an example, rather than having NaN values, an alternative is to substitute them with ”Unknown” or with the most occuring group [14].

Figure 20 illustrates the count of missing observations for each explanatory variable in the time series data set. This orginal data set is extensive, containing a total of 9 117 535 entries. Our focus here is specifically on the outliers within the consumption and production variables. During the earlier clustering, data points had to be excluded due to missing values in consumption. Therefore, 30% of the initial data was disregarded, which can be seen in the length of Figure 15. Although the production variable has a missing value percentage exceeding 90%, its significance is relatively lower. This is because: 1) we did not utilize the production information, and 2) only 7.9 % of the given households engage in electricity production. It is relatively uncommon for customers to generate electricity, as it is limited to homes equipped with devices such as solar panels.

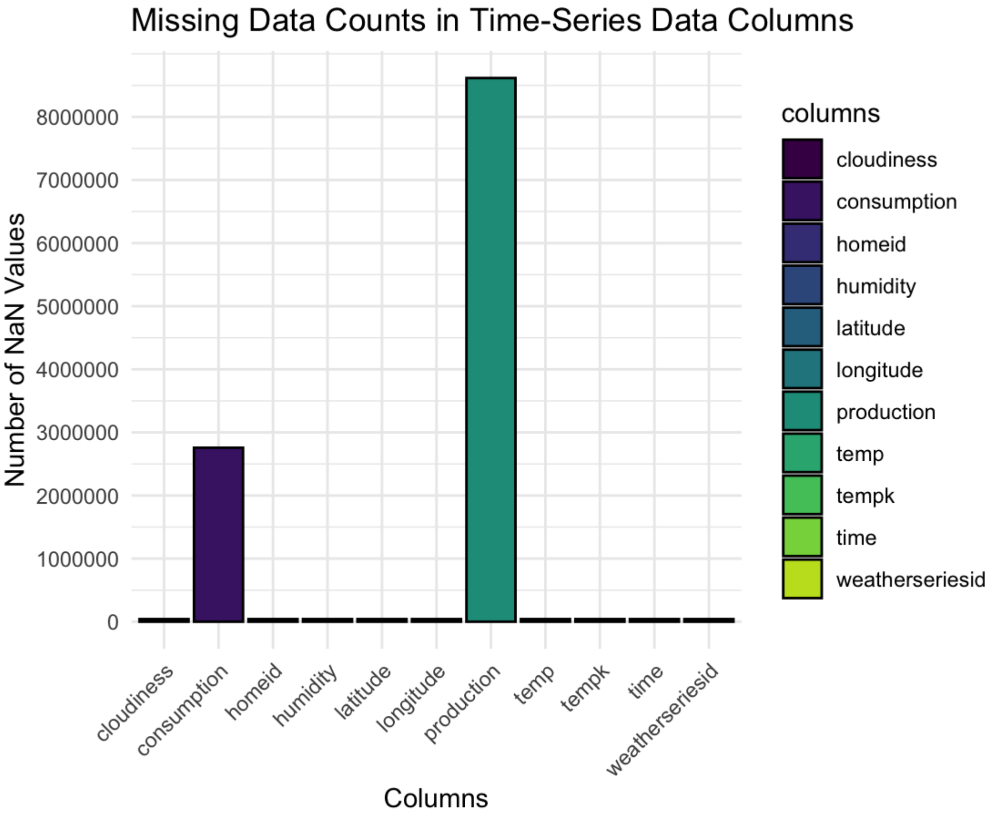


Figure 20: Plot of the number of missing values in the time series data.

Figure 21 illustrates the number of missing observations for each explanatory variable in the home properties data set. Given that the data set comprises 141 108 entries, the majority of explanatory variables exhibit a low percentage of missing data. However, an exception is observed in the case of smart_heating_enabled, where more than 90% of entries are missing, rendering the variable relatively unreliable. In comparison, the next variable with the highest percentage of missing values is home_heating_source, missing approximately 30%. As home_size will be a key variable in the subsequent phases of the thesis, it is not desirable for it to have a high missing value rate.

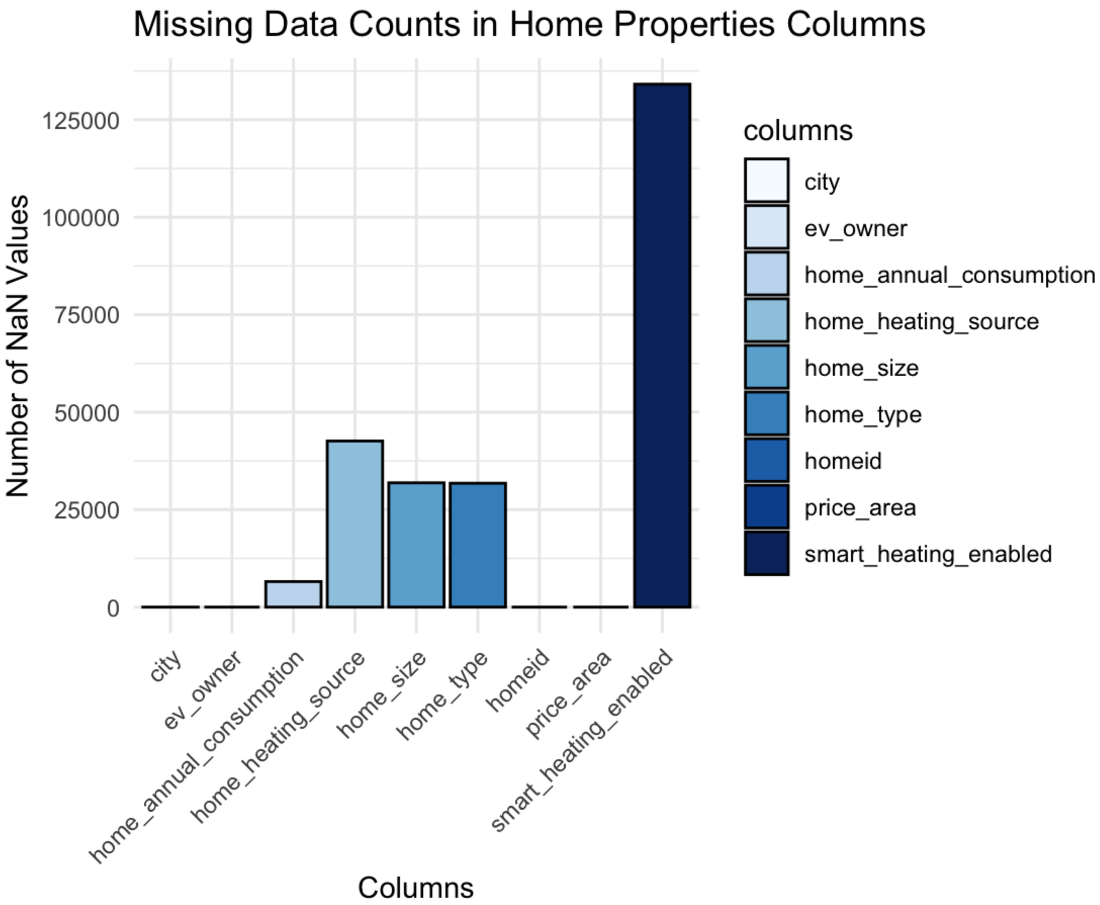


Figure 21: Plot of the number of missing values in the home properties data.

4 The Method of Experimentation

This chapter focuses on the practical implementation of the statistical methods introduced in the theoretical framework. All aspects of classification and diagnostic procedures are performed utilizing R, an open source flexible software environment well known for its capabilities in statistical analysis and data visualization. It is worth noting that previous project tasks were executed in Python, predominantly utilizing Pandas for handling data in dataframe structures. Consequently, the final stages involving the implementation of prediction models and error calculation, mirroring the approach taken in the project thesis, were also conducted using Python.

4.1 Problem Approach

The approach applied to the problem can be split into three distinct cases. Here, one year of historical consumption data for a subset of customers has been used. The data set is divided at a specific date, forming clusters based on data occuring before that date, and predictions are made for the subsequent time interval.

1. **One Cluster Approach**:

   Initially, we will treat all the data as a single cluster and forecast the consumption for this entire cluster. Subsequently, we will evaluate the predicted consumption against the actual consumption of customers at that specific timestamp, using the MAE.

2. **Historical Clustered Forecast Approach**:

   For the second approach, we will form clusters based on historical consumption patterns until the given date. The hierarichical analysis use the hourly mean consumption during a day to assign the data to distinct clusters. Then, individual forecasts for each cluster is generated using GLM and linear models. The predicted consumption for each cluster is calculated by multiplying the number of homes in the cluster by its predicted consumption. The cluster predictions are summed to obtain the overall predicted consumption for the test time interval. This consumption is compared against the actual consumption, using the MAE. Then, we compare the MAE from this approach with the MAE from the first approach. This was implemented in the project thesis, resulting in a lower MAE for the second approach. Consequently, we proceeded with further analysis based on this methodology.

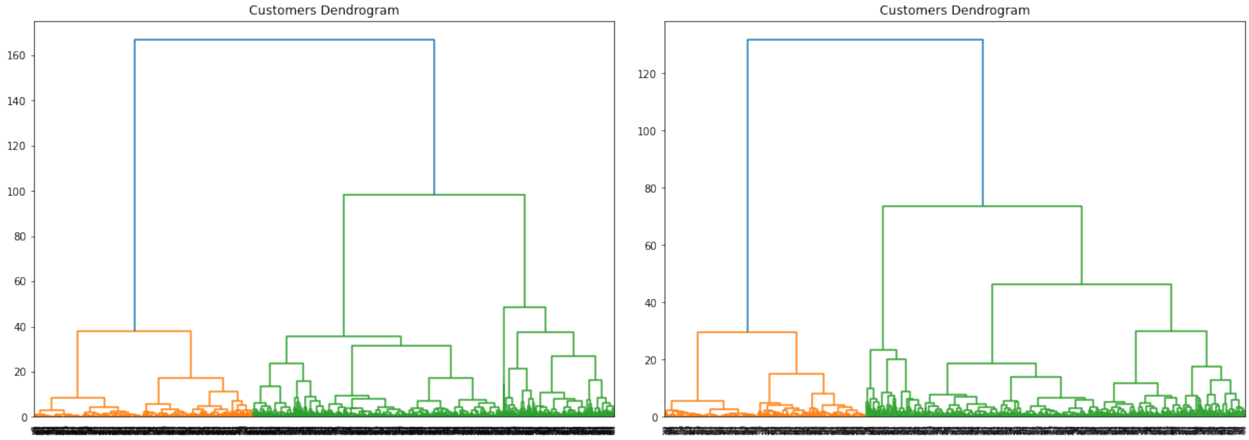
3. **Integrating ”New Customers” Approach**:

   We need to designate a subset of homes as ”new customers” from the given data set and re-cluster using historical hourly consumption data, as was done in approach 2. As with real new data, new homes lack historical information of the hourly consumption data and pattern during the day. Therefore, new homes need to be classified into their supposed clusters based on their home properties using a classification method. We add the newly classified homes to the respective subclusters. Conduct a GLM or LM prediction and a MAE calculation, following the procedure in approach 2. The MAE from this approach is compared with the MAEs from the previous two approaches. For the result to be valuable for Tibber, the MAE from this approach should be lower than the MAE from the first approach. This is the main purpose of our investigation.

4.2 Clustering

As we are operating with the data set used in the project thesis [16], the necessity to create new instances of customers, could only be derived from the existing data set. In the previous analysis, all data had been clustered. However, because we intend to set aside a portion of the data to represent new customers, a re-clustering is necessary for the remaining data.

The clusters where made in Python with the imported package "scipy.cluster.hierarchy" [6]. The clusteres are formed by linking points using the Ward method and measuring the Euclidean distance between the points. Figure 22 presents two dendrograms, one obtained through hierarchical analysis on the entire data set and another where the "new customers" of 248 households have been excluded. When looking at a dendrogram, one can choose a preferred number of clusters by inspecting the dendrogram or establish clusters by setting a threshold. For example, setting a threshold of 60 would in these two dendrograms result in forming 3 clusters.



((a)) Dendrogram of the entire data set (829 households).

((b)) Dendrogram without new customers (248 households).

Figure 22: Displaying two dendrograms illustrating the clustering of the different data sets.

As observed in the dendrograms, they exhibit a certain degree of similarity in the overall structure of the clusters, but are not entirely identical. This may influence the outcome of the predicted consumption. It is anticipated that clusters containing all the data would likely provide a more robust fit for future data points, given their larger data set. However, this is not guaranteed, as outliers, for instance, could potentially disrupt the formation of a well-defined general cluster.

4.3 Dummy Variables

To be able to use the categorical string values, we need to give them a number. Categorical variables, being distinct entities, pose challenges when assigned numerical values as these numbers may lack inherent meaning. For instance, in the explanatory variable home_heating_source, assigning "ground" number 1 and "electric_boiler" number 5 does not necessarily mean that "electric_boiler" has a greater impact on consumption simply because it has a larger numerical value. Therefore, we make our categorical values into dummy variables. Dummy variables transform the output into a binary format, producing values of 0 or 1. When dealing with more than two classes within a categorical column, the result is a matrix consisting of dummy variables, each converted to 0 or 1 based on its truth value [5].

The dummy variables were generated in R using the dummyVar() function from the "caret" package. Subsequently, the resulting dummy variables for categorical features were combined with the continuous variables to form an integrated dataframe.

4.4 Exploration of the Home Properties Variables

As mentioned above, the ongoing exploration of the data set is based on both the clusters established in the project thesis and the adjusted clusters. The clusters has been formed through hierarchical clustering, where an explanation of this method can be found in Section 2.4.2.

The analysis assessed the potential influence of various factors, including property type and home size, on electricity consumption based on the home properties data. It was hypothesized that these factors could significantly affect electricity usage, and this relationship was further examined.

In the classification approach, the variables from the home properties data that were considered for analysis are as follows:

- city
- smart_heating_enabled
- ev_owner
- home_type
- home_size
- home_heating_source
- home_annual_consumption

A more detailed description of the variables is given in Table 2. Numerous tests were conducted using various combinations of the variables together and evaluating their performance in the classification models. Each variable was individually tested, and they were also evaluated in progressively larger combinations with other variables.

Certainly, one can anticipate variations in the different variables. For example, individuals who own electric vehicles might naturally require additional electricity, suggesting a potential correlation. However, customers may also employ smart heating systems that optimize charging when electricity prices are lowest until the next use. Additionally, the size of a home is a relevant factor, with larger homes typically necessitating more electricity and, consequently, exhibiting higher consumption levels. The type of dwelling is another consideration; apartment buildings can benefit from shared heating between neighbour units, which differs from detached houses. Moreover, if we have access to the home_annual_consumption data in home properties, it offers insights into the overall annual consumption. Lastly, the city's geographic location, even within SE3, can have varying weather effects, impacting electricity consumption patterns.

4.4.1 Backward Elimination of Categorical Variables

Dealing with the interplay of continuous and categorical variables has posed a persistent challenge throughout the thesis. Struggling to strike the right balance in determining which variables to include, we explored an innovative approach. Observing that categorical variables had minimal impact on accuracy rates due to the dominance of continuous variables, we focused solely on the categorical ones. Converting them into dummy variables, we constructed a linear model with the home_annual_consumption as the response variable. Then, we performed backward elimination using the output of the linear model as input to assess their significance. This method aims to create a "new" variable, incorporating the most influential categorical factors. Therefore, we intended to integrate this variable into the classification alongside the continuous variables. It is worth noting that we excluded city from this process since we confined our analysis to the same price area, and considering all cities would result in a large and computationally intensive matrix of dummy variables.

This process was executed using R, employing lm() and step() with the input parameter 'direction = "backward"' [18]. Both of these functions belong to the base R package. In the linear model, the response variable was incorporated into the backward model, and the significance of the variables was assessed. The significant variables were retained, and the fitted values were introduced as a new response in the subsequent modeling step.

4.4.2 Creating a Continuous Variable from a Linear Model

The backward elimination process reveals the most significant variables. With this information, we can selectively include only the most impactful variables in the matrix consisting of dummy variables. A response variable can then be created from the remaining variables through linear model fitting. The home_annual_consumption is employed as the response variable in the linear model, with the chosen categorical variables transformed into a matrix of dummy variables serving as predictors. This yields a new variable that can be treated as continous and incorporated into the classification models.

4.5 Training and Test Sets

To utilize our data set effectively, we must partition it into training and testing sets. This division is crucial because evaluating the performance of our models based predominantly on existing clustered customers does not address the main problem of determining where to place new customers. We must examine how the model performs with new customers, where we lack prior knowledge of their consumption patterns. Our objective is to ensure that a new customer is accurately assigned to the correct cluster from the beginning, allowing for optimal consumption predictions.

The process involves randomly dividing the existing home-ids into a 70% training set, consisting of already clustered households, and a 30% test set representing "new customers." The training labels serve as targets for the classification method, while the test set receives predicted labels. Given that clusters had been assigned to all home-ids up to a specific date in a previous stage, adjustments were necessary for the clusters conducted in the project thesis. Since our data set is limited, we treat the 30% test set as if it comprises entirely new customers with no historical hourly consumption patterns. Therefore, we randomly select a set of home-ids from the data, exclude them, and then create new clusters from the remaining data. Consequently, when classifying these new homes, they are treated as entirely new entities in terms of clustering.

As all households were previously clustered, we retain their original cluster labels for accuracy verification. The predicted targets from the classification are then compared to the existing test labels derived from earlier clustering, allowing us to assess the accuracy rate of the classification model's predictions.

4.6 Classification

The classification methods, as discussed in Section 2.2, were implemented in R. The built-in functions employed include knn(), lda(), qda(), randomForest(), NaiveBayes(), and svm(). These functions are part of the packages "class," "MASS," "randomForest," and "e1071".

When employing these techniques, the models were trained on the training set, which consisted of the included the home properties data, and the response variable. The response variable was labeled as training labels, derived from the clusters generated through hierarchical clustering. Subsequently, the trained model was applied to the test set, and a predicted class for each observation was obtained. In this context, the predicted class corresponds to the cluster to which the new home is assigned.

4.7 Prediction

While we evaluate the test accuracies of various classification methods, it is necessary to ascertain whether grouping contributes to improved predicted consumption, especially if it aligns more closely with the true value of customers' consumption. In the project thesis, we developed a prediction model for clustered data and another for treating all data as a single cluster. The comparison involved assessing the MAE of these models against the actual consumption outcomes. Now, a similar procedure is required for new households. When onboarding new customers, the absence of historical information about the hourly consumption data hinders the use of clustering for placement, making this approach of the new data insufficient.

Despite the positive results in the project thesis, it is essential to recognize that the outcome may not guarantee success in this new setting. When considering more extensive test sets, as in our case, classification models will not achieve flawless categorization into presumed clusters. The classification model could potentially assign a new home to a less fitted cluster. Because different prediction models exist for the different clusters, this may lead to a lower MAE when considering all clustered data compared to the combination of newly classified homes and the clustered data. Additionally, since new clusters are formed of the data without incorporating the new homes, there is no assurance that these new clusters will yield more accurate predicted consumption than treating all the data as one single cluster.

When forecasting the future consumption of clustered customers, we employed a straightforward GLM analysis based on "temp," "humidity,", "cloudiness" and "production" from the time series data as explanatory variables (see Table 1 for details). We constructed our prediction model using a GLM with the smf.glm() function from the Python package statsmodels.formula.api. Initially, we employed a gamma distribution for the model. However, as will be discussed later in the analysis, we eventually transitioned to a normal distribution, effectively transforming our GLM model into a linear model (LM). In addition, we introduced a new column named "hour" in this data set, derived from the timestamp of each respective row. This straightforward modification was intended to enhance the performance of our linear model, as it is designed for predicting hourly consumption.

For new customers, integration into existing prediction models generated from the clustering process is necessary. This involves utilizing the GLM models for the old clustered data. However, when summarizing the consumption of each cluster, it is crucial to account for any changes in the number of customers in a specific cluster after classification. For instance, in cluster 1, we use the prediction model made for cluster 1 based on the households without the new customers. We then multiply the model's predictions by the respective counts of old homes and new homes within cluster 1, yielding an overall forecasted consumption for these homes in this specific time interval.

Subsequently, we calculate the MAE based on the difference between the clusters predicted consumption and the known actual consumption. This MAE will be compared against the entire data as one single cluster and the old clustered data of all the data to evaluate the degree of difference. Figure 23 illustrates the hourly consumption pattern for all homes over a year from the time series data set. The specific date marking the split between the training and test sets is evident in the plot where the colors transition from black to red. The forecasted consumption for all data treated as a single cluster using the GLM with a gamma distribution is visualized in the green plot.

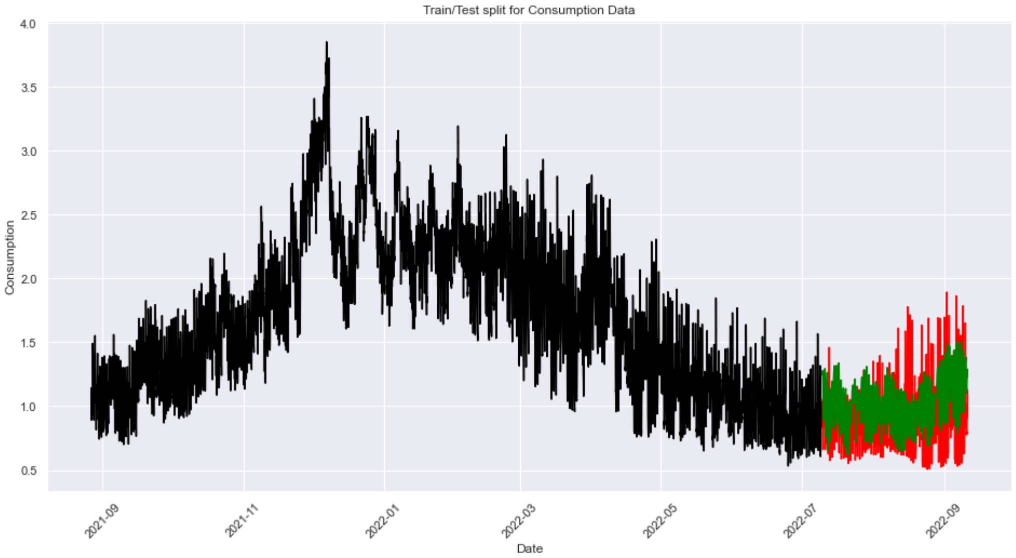


Figure 23: Shows the predicted consumption by GLM with gamma distribution for all the data as one cluster.

Source: [16]

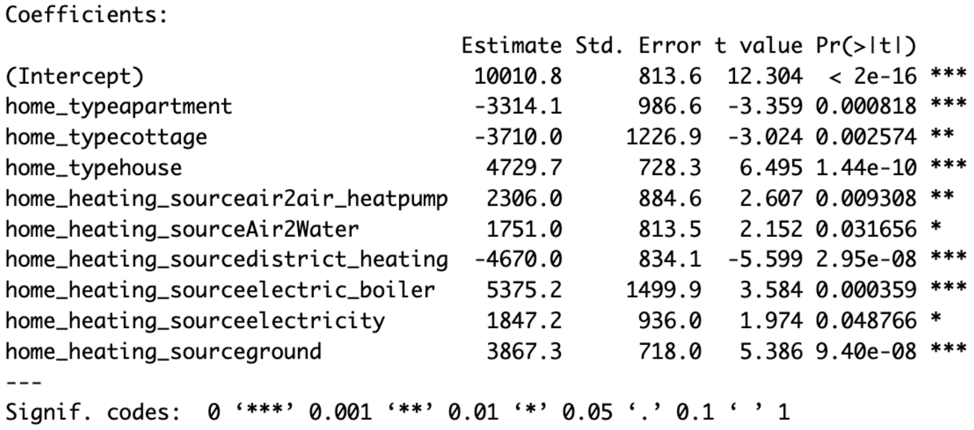
5 Results and Analysis

This chapter delves into the presentation and evaluation of the achieved results. Initially, various experimented approaches will be introduced, followed by the final chosen approach. Subsequently, the different prediction models and their corresponding MAE results will be detailed.

5.1 Several Tried Approaches

5.1.1 Backward Elimination

As mentioned in Section 4.4.1, we employed a backward elimination method to assess the significance of categorical variables, particularly as they were overshadowed by the continuous variables. Combining these distinct types of variables posed challenges for classification methods to effectively process the input. Initially, the categorical variables underwent transformation into dummy variables. Thereafter, a linear model was constructed with home_annual_consumption as the response variable. The backward elimination method was then applied to assess the significance of various categorical variables, leading to the removal of those deemed uninteresting. Figure 24 illustrates the outcome, highlighting the categorical variables that were most significant. Notably, "home_typerowHouse" did not meet the criteria, and the entire explanatory variable smart_heating_enabled also failed to qualify as significant. Through backward elimination, we obtained an R-squared value of 0.4624, utilizing AIC as it is the default criterion in the step() function.



```
Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)                          10010.8      813.6  12.304  < 2e-16 ***
home_typeapartment                   -3314.1      986.6  -3.359 0.000818 ***
home_typecottage                     -3710.0     1226.9  -3.024 0.002574 ** 
home_typehouse                        4729.7      728.3   6.495 1.44e-10 ***
home_heating_sourceair2air_heatpump   2306.0      884.6   2.607 0.009308 ** 
home_heating_sourceAir2Water          1751.0      813.5   2.152 0.031656 *  
home_heating_sourcedistrict_heating  -4670.0      834.1  -5.599 2.95e-08 ***
home_heating_sourceelectric_boiler    5375.2     1499.9   3.584 0.000359 ***
home_heating_sourceelectricity        1847.2      936.0   1.974 0.048766 *  
home_heating_sourceground             3867.3      718.0   5.386 9.40e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 24: Display of the summary of the Backward Elimination method applied to the categorical variables.

After preserving the categorical variables, we constructed a linear model using these variables to generate a continuous variable named "new_response." This variable was intended to be incorporated into the classification models as an additional continuous feature. This experiment aimed to assess whether augmenting the data set with more of the home properties data through an additional variable would enhance the performance of the classification methods. However, the results did not demonstrate improvement.

5.1.2 Transformations

Various transformations, elaborated in Section 2.3, were tried out. One might assume that Box-Cox would be a suitable transformation as it selects the best-fitted lambda to make output variables more Gaussian-like. However, contrary to expectations, this approach resulted in a higher MAE. Conversely, standardization significantly enhanced the overall accuracy of predicted consumption. Consequently, we proceeded to standardize the selected variables before incorporating them into the classification methods. Standardization was also tested along with the backward elimination method, yielding a small improvement. However, despite the enhancement in this method, applying backward elimination still led to an overall higher MAE. This became evident when incorporating the backward variable as an input for the classification model, predicting the clusters, and subsequently comparing the MAE with the values from Table 3.

5.1.3 Re-clustering

The thesis involved a comprehensive series of tests. Given the necessity to re-cluster homes after excluding supposedly new homes, the optimal number of clusters could vary. As detailed in Section 4.2, the clusters underwent slight changes following the removal of new homes. Although the disparity seen from Figure 22 is not very large, the distinction between clusters for the entire data set and the data set excluding new homes increased notably based on the quantity of homes designated as "new customers." Consequently, numerous tests were conducted with varying amounts of new homes to determine the optimal configuration for predicting consumption. Ultimately, we decided to include 248 households as new additions, comprising a 30% subset of the total households considered in the thesis (829 homes). Additionally, diverse cluster quantities were examined including 3, 4, 5, and 6 clusters, all of which yielded acceptable outcomes. Nevertheless, in line with the project thesis, five cluster groups were retained as they provided an overall improved consumption result and were deemed reasonable upon analyzing the dendrograms.

Furthermore, considering the continuous variable home_size as one of the home properties input during the classification methods, we aimed at addressing some missing values in this variable. It is important to subtract these values from the start. If we only address this issue in the training and test sets for the classification method, the count of "clustered and classified" homes might be lower than for all the homes in a single cluster. This is because that cluster relies solely on consumption and not on the home properties. Consequently, it was necessary to eliminate NaN values for home_size for all home-ids within the cluster, resulting in a reduction of the number of customers analyzed from 843 homes to 829 homes.

5.2 Accuracy Rates

After experimenting with six distinct classification models, our choice ultimately settled on using LDA. While the other methods are generally considered reliable, their performance varied significantly throughout the entire testing phase of the thesis. LDA is typically designed for continuous variables, but during extensive testing, we opted to exclusively employ continuous variables and tested them together with the continuous variable derived from the categorical variables. This decision was driven by the challenges encountered when combining continuous and categorical variables. Ultimately, the categorical variables were excluded from the models due to challenges arising from the complexity of managing the matrices including the dummy variables. The size of this matrix and the minimal impact of the categorical variables in the classification methods, made their inclusion less effective in the context of classification models.

Figure 25 illustrates the accuracy rates of the different classification models tested. In this particular scenario, we employed five clusters and identified 248 homes as new homes. This configuration was chosen as our final test and it revealed that LDA achieved the highest accuracy rate at 48.4%. Despite potential variations in classification methods with the highest accuracy in different cases, LDA exhibited greater stability across multiple tests, including cases with varying quantities of new homes. In the final case, the classification involves the standardized variables home_size and home_annual_consumption, as these variables yielded the best MAE in the end. Over an extended period, the explanatory variable ev_owner was included. Nevertheless, the MAE decreased by 8% when we excluded it and utilized only the standardized variables home_size and home_annual_consumption.

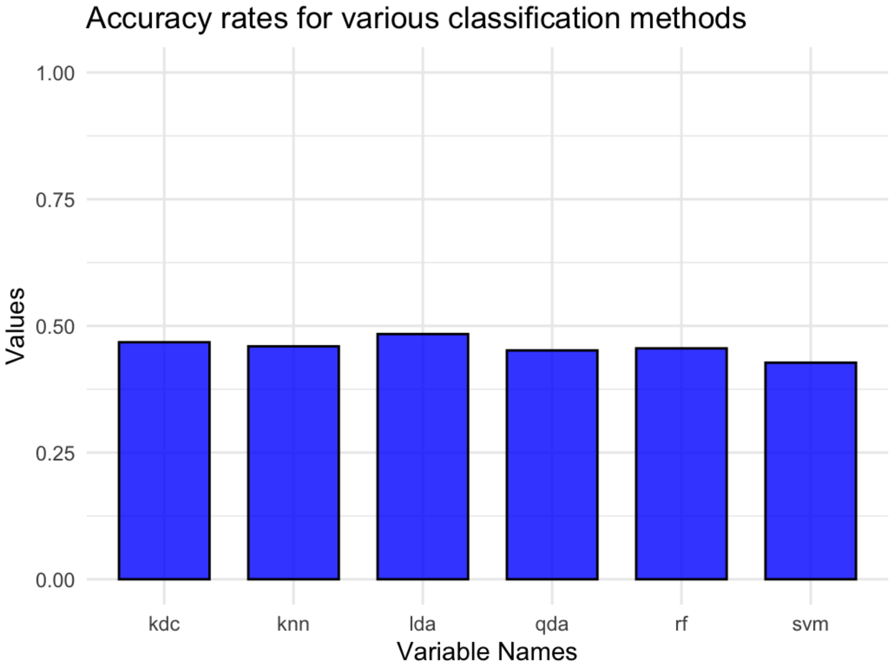


Figure 25: Display of the outcome of accuracy rates of different classification methods when assigning 248 new households to 5 distinct clusters.

It is important to highlight once more that the accuracy rates presented here stem from the comparison between the old cluster labels and the newly classified cluster labels. However, since we have extracted a subset of customers to form new clusters, the accuracy rates and old labels may not offer complete insights, as they reflect the group labels of the all the data clustered into five groups (approach 2). Therefore, the primary focus should be on the MAE output, as it offers a more meaningful success rate of the model in this context.

Despite the modest accuracy, just slightly below 50%, this information remains valuable. Considering the task involves assigning new objects to five different clusters, a 50% accuracy rate is noteworthy, as it surpasses the expected performance from random placement.

Figure 26 illustrates the segregation of test observations into five classes using the LDA method. Upon examining the variables home_size and home_annual_consumption, distinct patterns emerge. For instance, cluster 1 exhibits characteristic features in LDA, such as a low annual consumption and a small household size. One can also observe that the method appears to face challenges when assigning objects that lie in the intermediate region between cluster 3 and 4. The plot indicates that these two clusters share a similar range of home sizes, and both groups can have annual consumption values around 20,000 kWh. The figure also illustrates that only one observation has been assigned to cluster 5. Further examination seen in Figure 22(b), reveals that clusters 2 and 5 are notably the smallest. Consequently, in this LDA separation with the test set, fewer points have been assigned to cluster 2 and 5. These two groups also appear to have home sizes above 200 m^2 and annual consumption above 20,000 kWh.

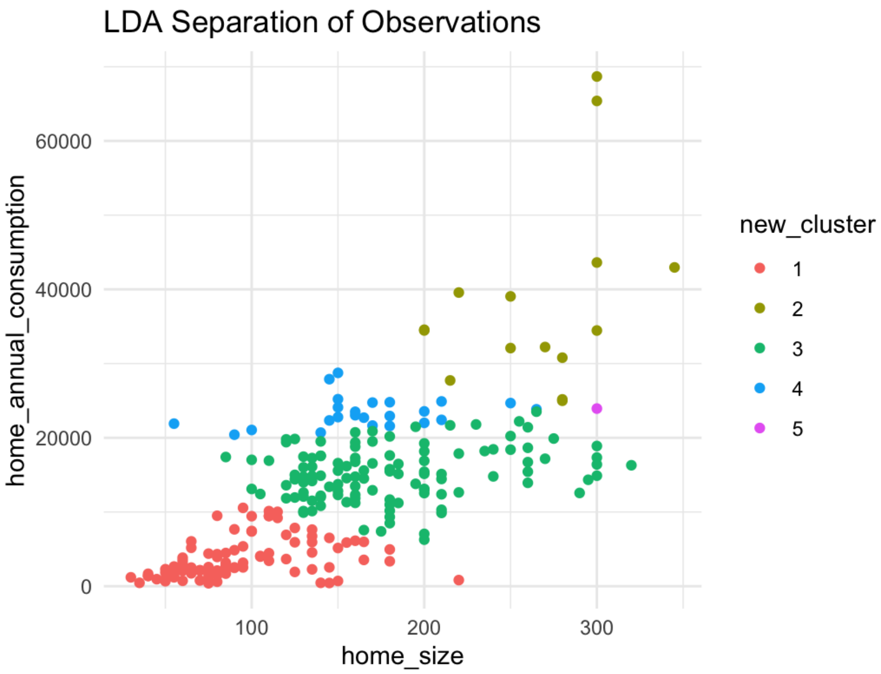
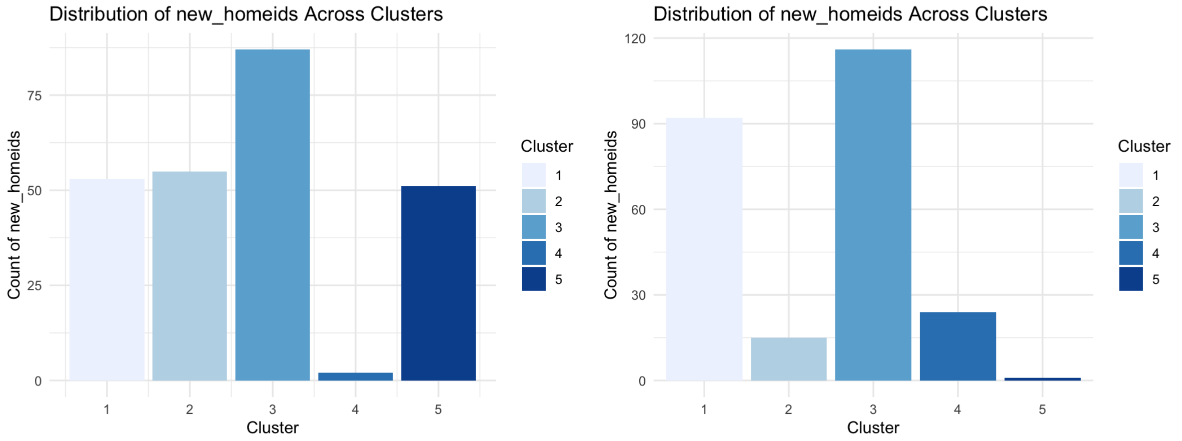


Figure 26: Display of the LDA separation of observations plotting for x = home_size and y = home_annual_consumption.

5.3 Homes Classified to Clusters

The data in Figure 27 illustrate a comparison within the clusters between the total count of households derived from cluster-based analysis and the count of households from newly classified homes combined with the re-clustered old homes. A noticeable difference is evident in the distribution of homes, marked by an increase in the clusters 1, 3 and 4 for the newly classified homes, diverging from the clustered results encompassing the entire data set. Simultaneously, there is a corresponding decline in cluster 2 and 5. The figure might lead one to consider that four clusters would be a more suitable choice, given the initial clustering indicates a relatively low number of homes in the forth cluster. However, the MAE, to be presented below, revealed that five clusters were actually preferred. This outcome might vary depending on the selection of new homes, but in most instances, five clusters yielded superior results. Hence, we proceeded with that number.

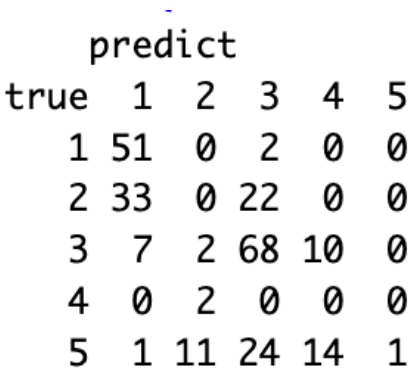


((a)) The count of cluster-based households in the clusters.

((b)) The count of newly classified households in the clusters.

Figure 27: The difference of count of households in the clusters.

It is important to note that there is no confirmation that the housing distribution depicted in Figure 27(a) is the true solution. Instead, it highlights how employing different methodologies can yield varied outcomes. Figure 28 displays the confusion matrix obtained from the LDA separation applied to the test set. A noticeable distinction is evident between the newly classified homes and their initial placement of the test homes through hierarchical clustering. For instance, none of the test observations have been assigned to cluster 2, where they were previously positioned. However, a crucial step involves evaluating the MAE to determine the quality of these results.



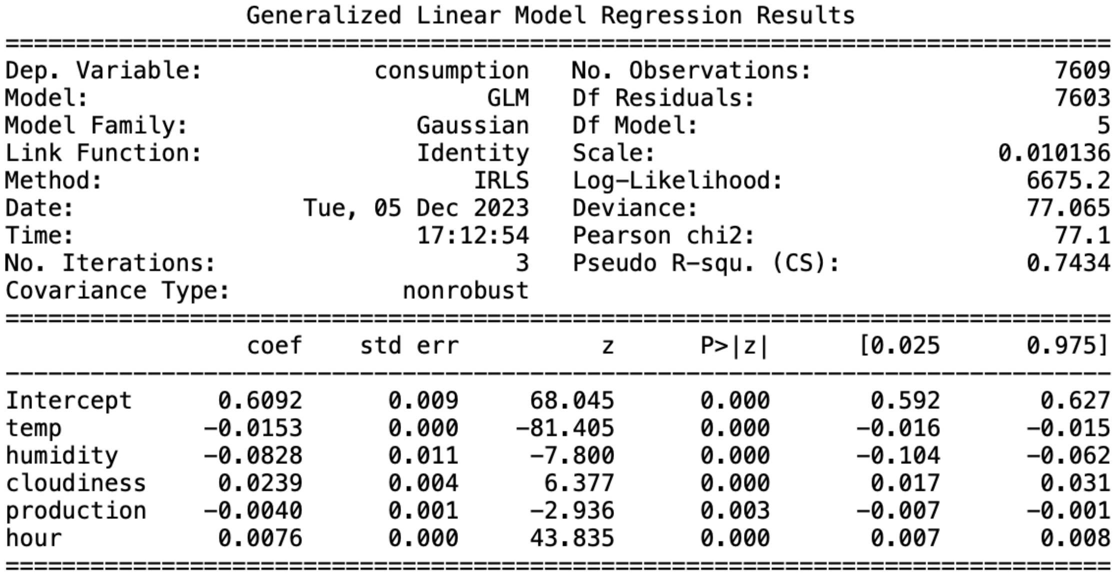
```
     predict
true  1  2  3  4  5
   1 51  0  2  0  0
   2 33  0 22  0  0
   3  7  2 68 10  0
   4  0  2  0  0  0
   5  1 11 24 14  1
```

Figure 28: The confusion matrix resulting from the application of the LDA method on the test set.

5.4 Modelling

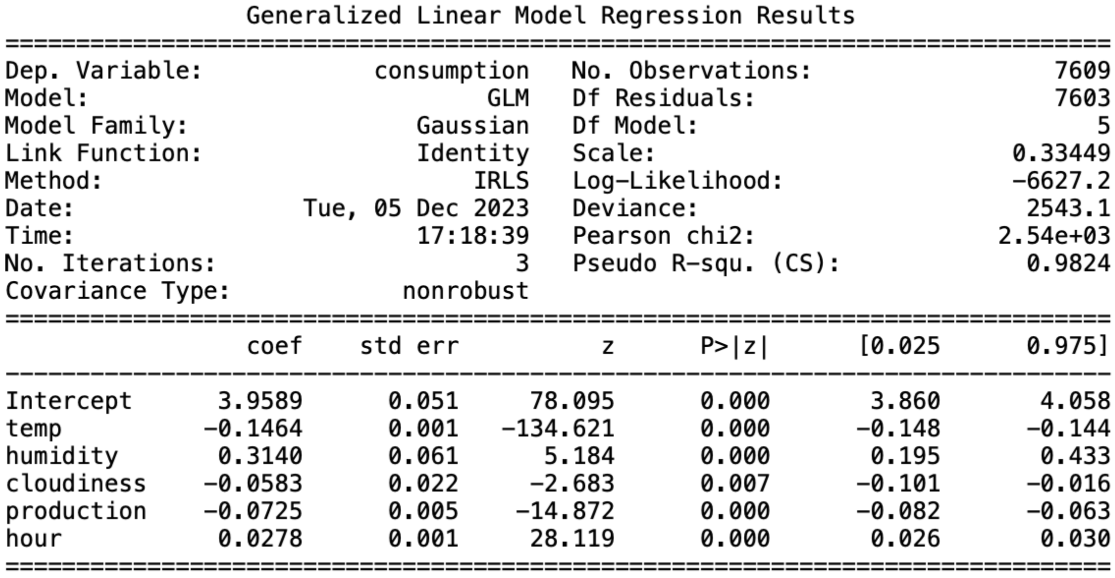
In the initial stages of the project thesis, GLM models were constructed using the Gamma distribution. However, in the subsequent master's thesis, it was observed that employing the normal distribution yielded significantly improved MAE. Consequently, the Gaussian distribution was adopted for the modeling of new clusters.

Figure 29, 30, 31, 32 and 33 displays the overview of the results from the GLM with identity function as link function corresponding to distinct clusters. This pertains to the linear model predictions generated for the data set excluding new customers, as the new customers are required to utilize the prediction models established for existing customers based on their cluster labels.



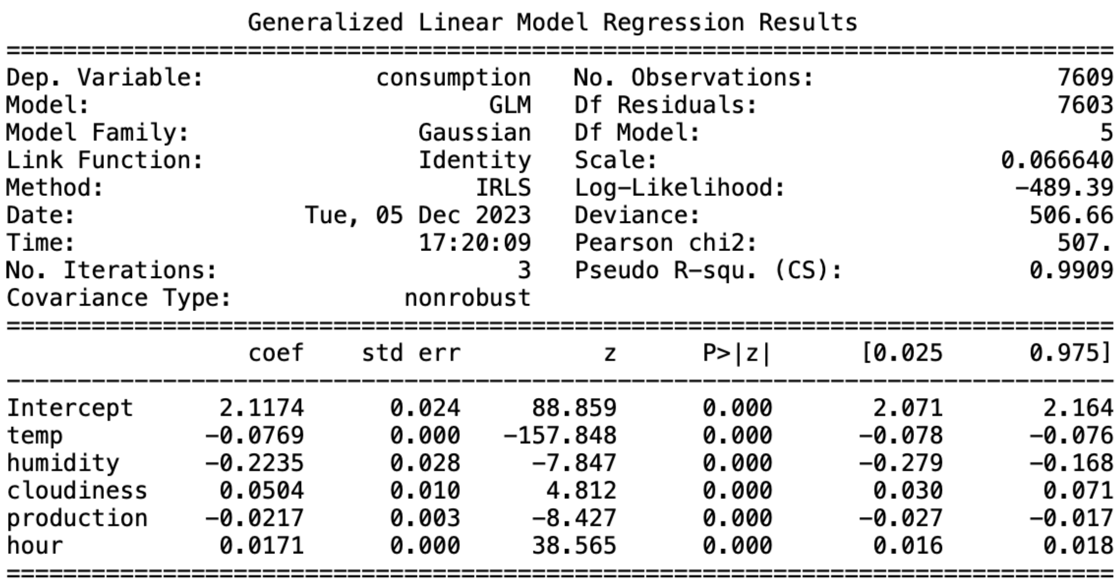
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:            consumption   No. Observations:                 7609
Model:                            GLM   Df Residuals:                     7603
Model Family:                Gaussian   Df Model:                            5
Link Function:               Identity   Scale:                        0.010136
Method:                          IRLS   Log-Likelihood:                 6675.2
Date:                Tue, 05 Dec 2023   Deviance:                       77.065
Time:                        17:12:54   Pearson chi2:                     77.1
No. Iterations:                     3   Pseudo R-squ. (CS):             0.7434
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.6092      0.009     68.045      0.000       0.592       0.627
temp          -0.0153      0.000    -81.405      0.000      -0.016      -0.015
humidity      -0.0828      0.011     -7.800      0.000      -0.104      -0.062
cloudiness     0.0239      0.004      6.377      0.000       0.017       0.031
production    -0.0040      0.001     -2.936      0.003      -0.007      -0.001
hour           0.0076      0.000     43.835      0.000       0.007       0.008
==============================================================================
```

Figure 29: Summary of the linear model of the first cluster.



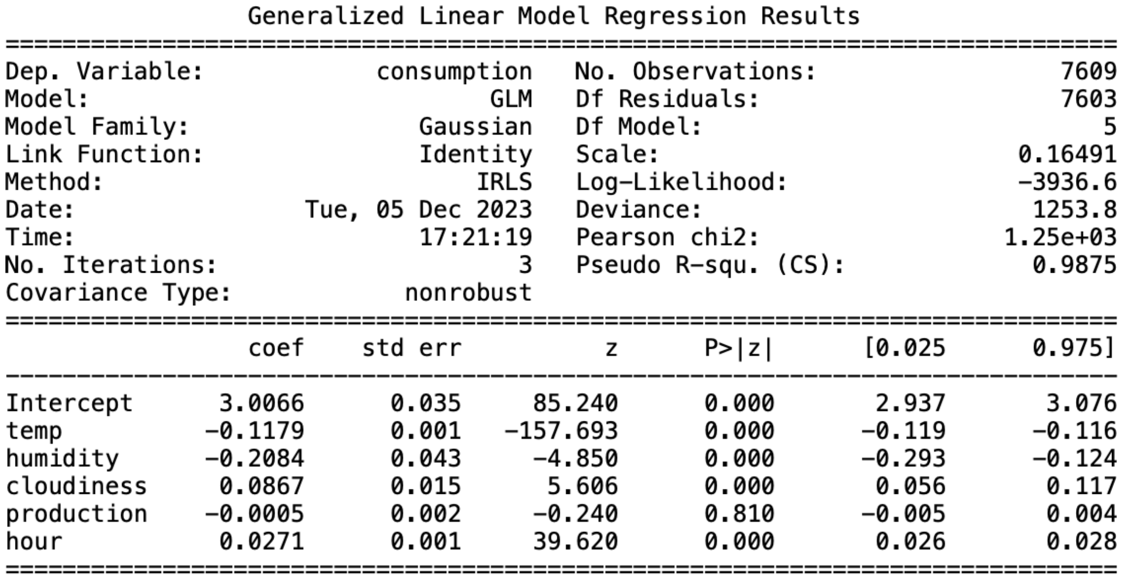
```
                 Generalized Linear Model Regression Results                  
==============================================================================
Dep. Variable:            consumption   No. Observations:                 7609
Model:                            GLM   Df Residuals:                     7603
Model Family:                Gaussian   Df Model:                            5
Link Function:               Identity   Scale:                         0.33449
Method:                          IRLS   Log-Likelihood:                -6627.2
Date:                Tue, 05 Dec 2023   Deviance:                       2543.1
Time:                        17:18:39   Pearson chi2:                 2.54e+03
No. Iterations:                     3   Pseudo R-squ. (CS):             0.9824
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      3.9589      0.051     78.095      0.000       3.860       4.058
temp          -0.1464      0.001   -134.621      0.000      -0.148      -0.144
humidity       0.3140      0.061      5.184      0.000       0.195       0.433
cloudiness    -0.0583      0.022     -2.683      0.007      -0.101      -0.016
production    -0.0725      0.005    -14.872      0.000      -0.082      -0.063
hour           0.0278      0.001     28.119      0.000       0.026       0.030
==============================================================================
```

Figure 30: Summary of the linear model of the second cluster.



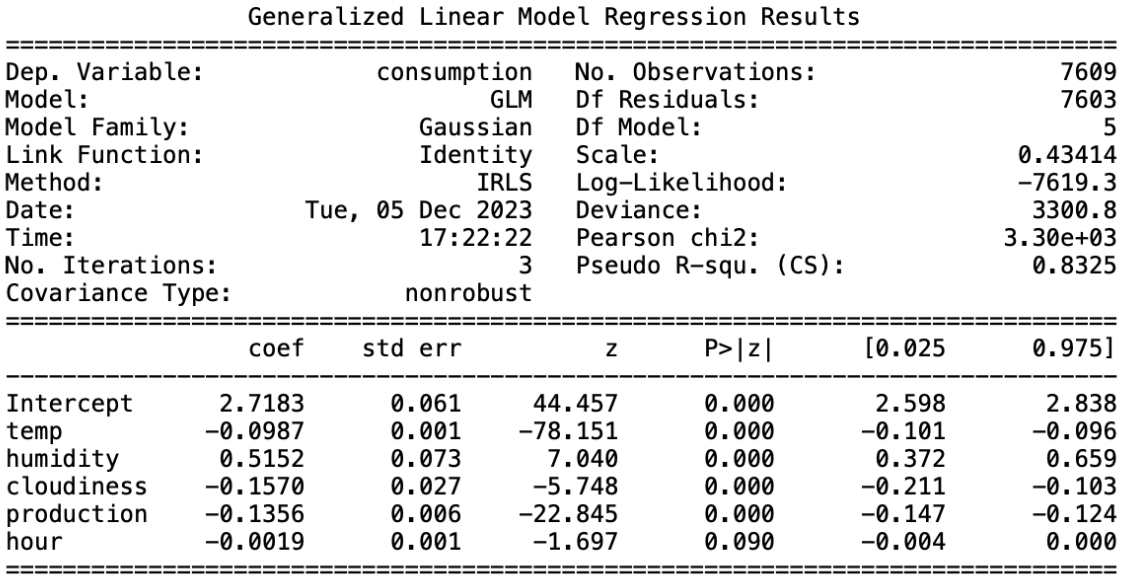
```
                 Generalized Linear Model Regression Results                  
==============================================================================
Dep. Variable:            consumption   No. Observations:                 7609
Model:                            GLM   Df Residuals:                     7603
Model Family:                Gaussian   Df Model:                            5
Link Function:               Identity   Scale:                        0.066640
Method:                          IRLS   Log-Likelihood:                -489.39
Date:                Tue, 05 Dec 2023   Deviance:                       506.66
Time:                        17:20:09   Pearson chi2:                     507.
No. Iterations:                     3   Pseudo R-squ. (CS):             0.9909
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      2.1174      0.024     88.859      0.000       2.071       2.164
temp          -0.0769      0.000   -157.848      0.000      -0.078      -0.076
humidity      -0.2235      0.028     -7.847      0.000      -0.279      -0.168
cloudiness     0.0504      0.010      4.812      0.000       0.030       0.071
production    -0.0217      0.003     -8.427      0.000      -0.027      -0.017
hour           0.0171      0.000     38.565      0.000       0.016       0.018
==============================================================================
```

Figure 31: Summary of the linear model of the third cluster.



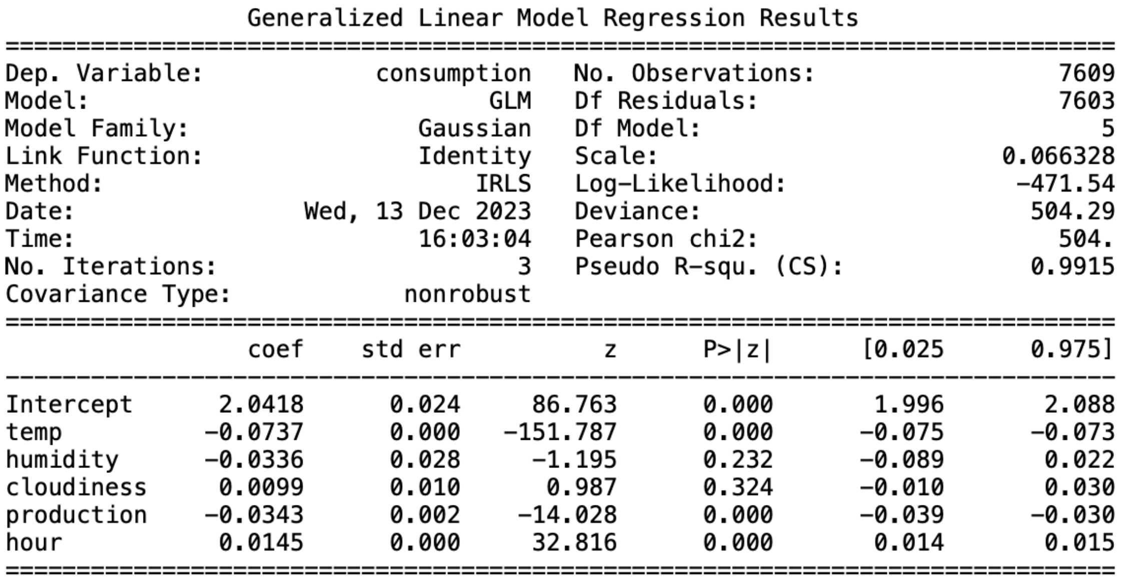
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:            consumption   No. Observations:                 7609
Model:                            GLM   Df Residuals:                     7603
Model Family:                Gaussian   Df Model:                            5
Link Function:               Identity   Scale:                         0.16491
Method:                          IRLS   Log-Likelihood:                -3936.6
Date:                Tue, 05 Dec 2023   Deviance:                       1253.8
Time:                        17:21:19   Pearson chi2:                 1.25e+03
No. Iterations:                     3   Pseudo R-squ. (CS):             0.9875
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      3.0066      0.035     85.240      0.000       2.937       3.076
temp          -0.1179      0.001   -157.693      0.000      -0.119      -0.116
humidity      -0.2084      0.043     -4.850      0.000      -0.293      -0.124
cloudiness     0.0867      0.015      5.606      0.000       0.056       0.117
production    -0.0005      0.002     -0.240      0.810      -0.005       0.004
hour           0.0271      0.001     39.620      0.000       0.026       0.028
==============================================================================
```

Figure 32: Summary of the linear model of the fourth cluster.



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:            consumption   No. Observations:                 7609
Model:                            GLM   Df Residuals:                     7603
Model Family:                Gaussian   Df Model:                            5
Link Function:               Identity   Scale:                         0.43414
Method:                          IRLS   Log-Likelihood:                -7619.3
Date:                Tue, 05 Dec 2023   Deviance:                       3300.8
Time:                        17:22:22   Pearson chi2:                 3.30e+03
No. Iterations:                     3   Pseudo R-squ. (CS):             0.8325
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      2.7183      0.061     44.457      0.000       2.598       2.838
temp          -0.0987      0.001    -78.151      0.000      -0.101      -0.096
humidity       0.5152      0.073      7.040      0.000       0.372       0.659
cloudiness    -0.1570      0.027     -5.748      0.000      -0.211      -0.103
production    -0.1356      0.006    -22.845      0.000      -0.147      -0.124
hour          -0.0019      0.001     -1.697      0.090      -0.004       0.000
==============================================================================
```

Figure 33: Summary of the linear model of the fifth cluster.



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:            consumption   No. Observations:                 7609
Model:                            GLM   Df Residuals:                     7603
Model Family:                Gaussian   Df Model:                            5
Link Function:               Identity   Scale:                        0.066328
Method:                          IRLS   Log-Likelihood:                -471.54
Date:                Wed, 13 Dec 2023   Deviance:                       504.29
Time:                        16:03:04   Pearson chi2:                     504.
No. Iterations:                     3   Pseudo R-squ. (CS):             0.9915
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      2.0418      0.024     86.763      0.000       1.996       2.088
temp          -0.0737      0.000   -151.787      0.000      -0.075      -0.073
humidity      -0.0336      0.028     -1.195      0.232      -0.089       0.022
cloudiness     0.0099      0.010      0.987      0.324      -0.010       0.030
production    -0.0343      0.002    -14.028      0.000      -0.039      -0.030
hour           0.0145      0.000     32.816      0.000       0.014       0.015
==============================================================================
```

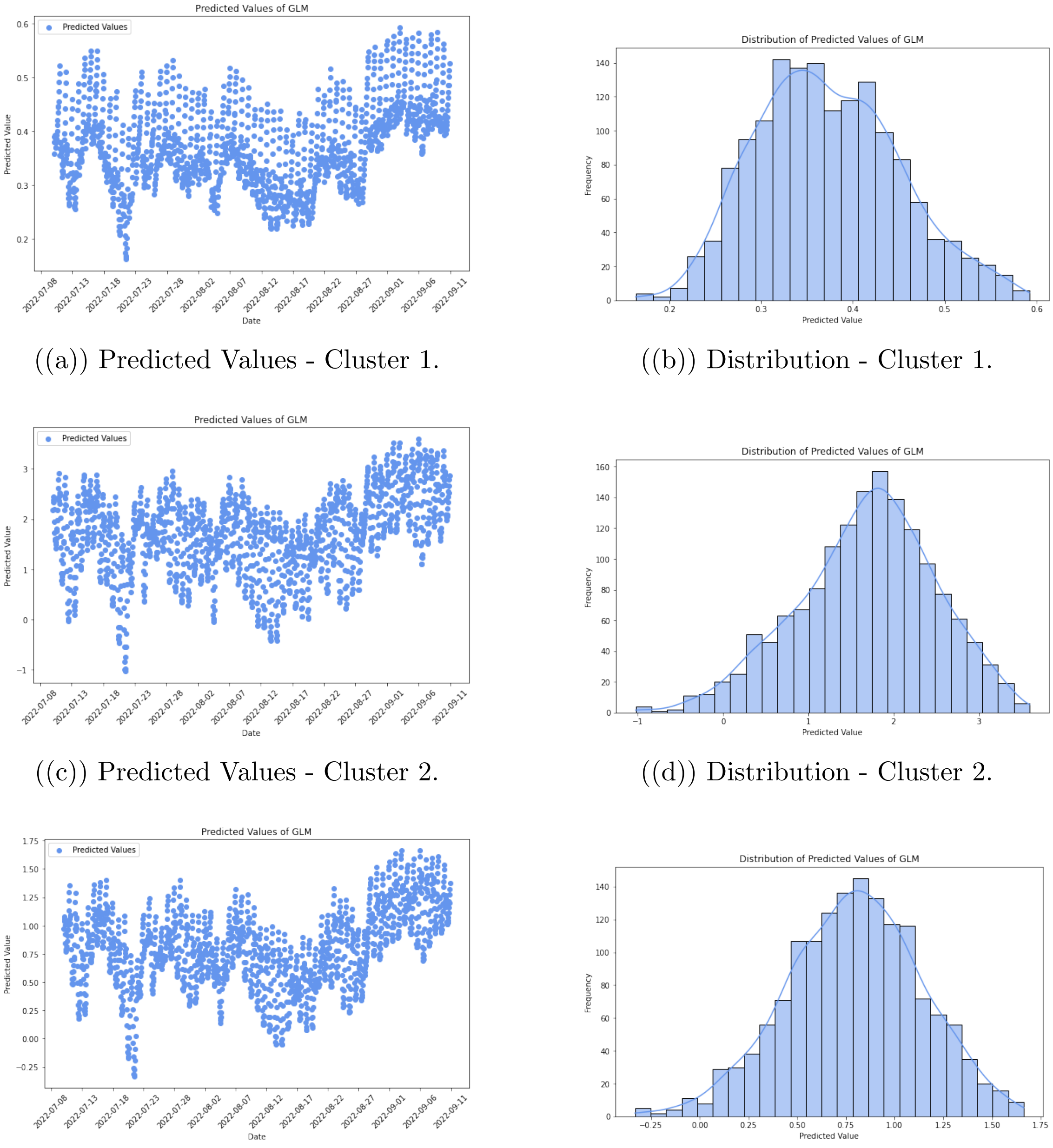
Figure 34: Summary of the linear model of the entire data as one cluster.

At first glance, it is evident that, across all LM result summaries, humidity has the most pronounced impact on the response variable, while the hour demonstrates a relatively lower impact. However, it is crucial to consider how these explanatory variables are presented. Given our intuitive expectation that humidity would have a smaller impact than temperature, it is essential to delve deeper into this comparison and explore the difference in range between these variables. Temperature is known to vary considerably within a range of +/- 30 (and potentially beyond), and can therefore have a large impact of the outcome. In contrast, both humidity and cloudiness exhibit values between 0 and 1, with a difference of 0.2 (0, 0.2, 0.4, ..., 1). Consequently, there is a much "larger jump" for humidity to increase by 1 compared to temperature, emphasizing the need to consider the scale and range of variables when interpreting their impact on the model's response.

Another notable observation is that temperature consistently exhibits a negative coefficient in all the linear models. This can be interpreted to mean that a higher temperature tends to a decreased consumption. Conversely, during colder weather, individuals are more likely to utilize more electricity to heat their households, thereby increasing consumption. Humidity, which has the highest coefficient across the different models, varies between being positive and negative values. This implies that within distinct clusters, an increase in humidity can either decrease or increase consumption. The same applies to cloudiness, although it has a lower coefficient. Naturally, the production value is negative because when the customers generate their own electricity, they tend to consume less electricity from the company. The hour coefficient is positive for all summaries excepts for cluster 5. For the rest of the linear models that implies that consumption tends to increase hourly during the day. For a more in-depth exploration of the relationships between all variables, see Appendix A for the covariance matrices derived from the linear models.

Figure 35(a) and 35(b) shows for cluster 1 the predicted values in the test time interval and their distribution, respectively. Afterwards, the remaining output for the clusters are presented in Figure 35 and 36. The plots reveal that the predictions for cluster 1 stand out when compared to the other clusters. This is noticeable in the wider spread of predicted values and a less pointed distribution. This observation aligns with the linear models, where cluster 1 exhibited a model with relatively small coefficients for both temperature and humidity. Cluster 5 also stands out compared to the other clusters, not only because of a more widespread pattern in predicted values but also due to a more skewed distribution. This could be attributed to the fact that cluster 5 is the only cluster with a negative hour coefficient.

Figure 37 and 38 depict the residual plots for the clusters, and Figure 38(c) illustrates a residual plot of the fitted model of the GLM with identity function as link function for all the data treated as one cluster. This plot illustrates the relationship between fitted values and residuals, showing the fitted values from the training set on the x-axis and the residuals, represented as $y_i - \hat{y}_i$, on the y-axis. In this representation, both cluster 1 and 5 stand out. Cluster 1 exhibits smaller residuals compared to the others, while the residual plot of cluster 5 has a more pronounced tilt to one side. The visual representation of residuals against the fitted models reveals a no ideal pattern. Nevertheless, in this thesis, the main focus has been on the predictions and the resulting value of the MAE.



((a)) Predicted Values - Cluster 1.

((b)) Distribution - Cluster 1.

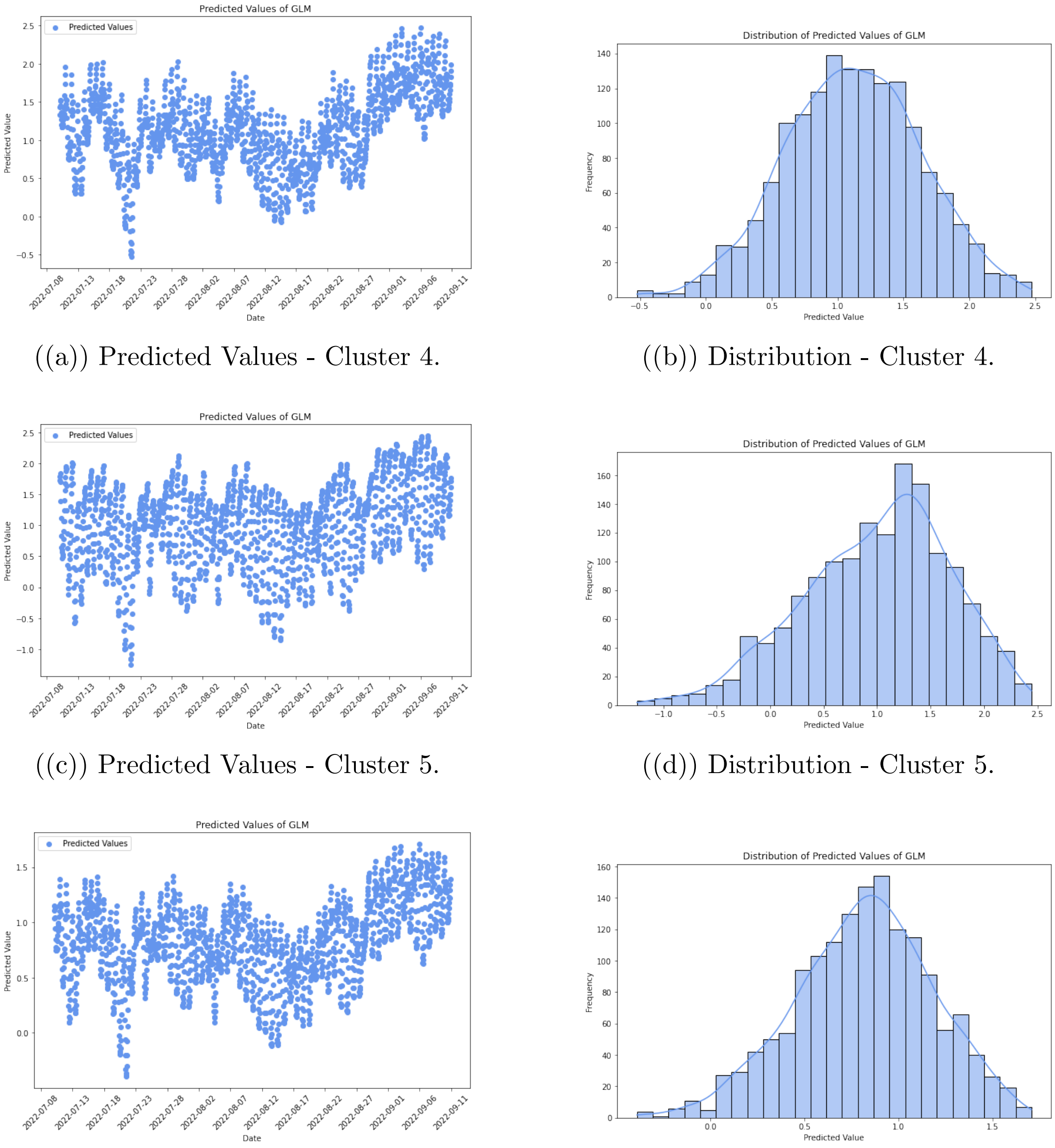
((c)) Predicted Values - Cluster 2.

((d)) Distribution - Cluster 2.

((e)) Predicted Values - Cluster 3.

((f)) Distribution - Cluster 3.

Figure 35: The left-sided plots show the predicted consumption values of one household for each hour in the test time interval for the clusters 1-3. The right-sided plots visualize the distribution of the predicted hour consumption values in the test time interval to the clusters.



((a)) Predicted Values - Cluster 4.

((b)) Distribution - Cluster 4.

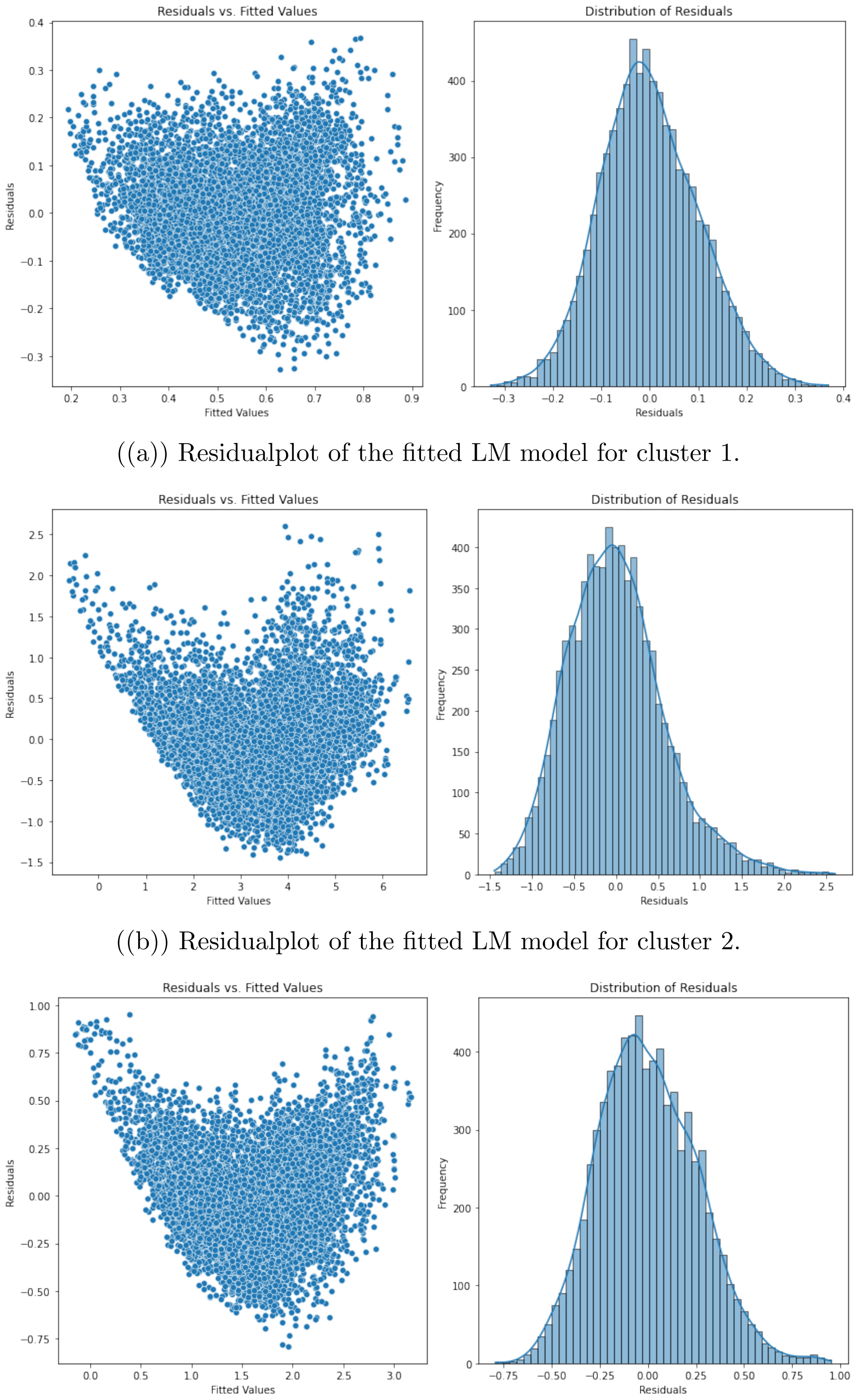
((c)) Predicted Values - Cluster 5.

((d)) Distribution - Cluster 5.

((e)) Predicted Values - All as 1 cluster.

((f)) Distribution - All as 1 cluster.

Figure 36: The left-sided plot shows the predicted consumption values of one household for each hour in the test time interval for cluster 4, 5 and the entire data set as one cluster. The right-sided plot visualizes the distribution of the predicted hour consumption values in the test time interval for the clusters 4, 5 and the entire data set as one cluster.

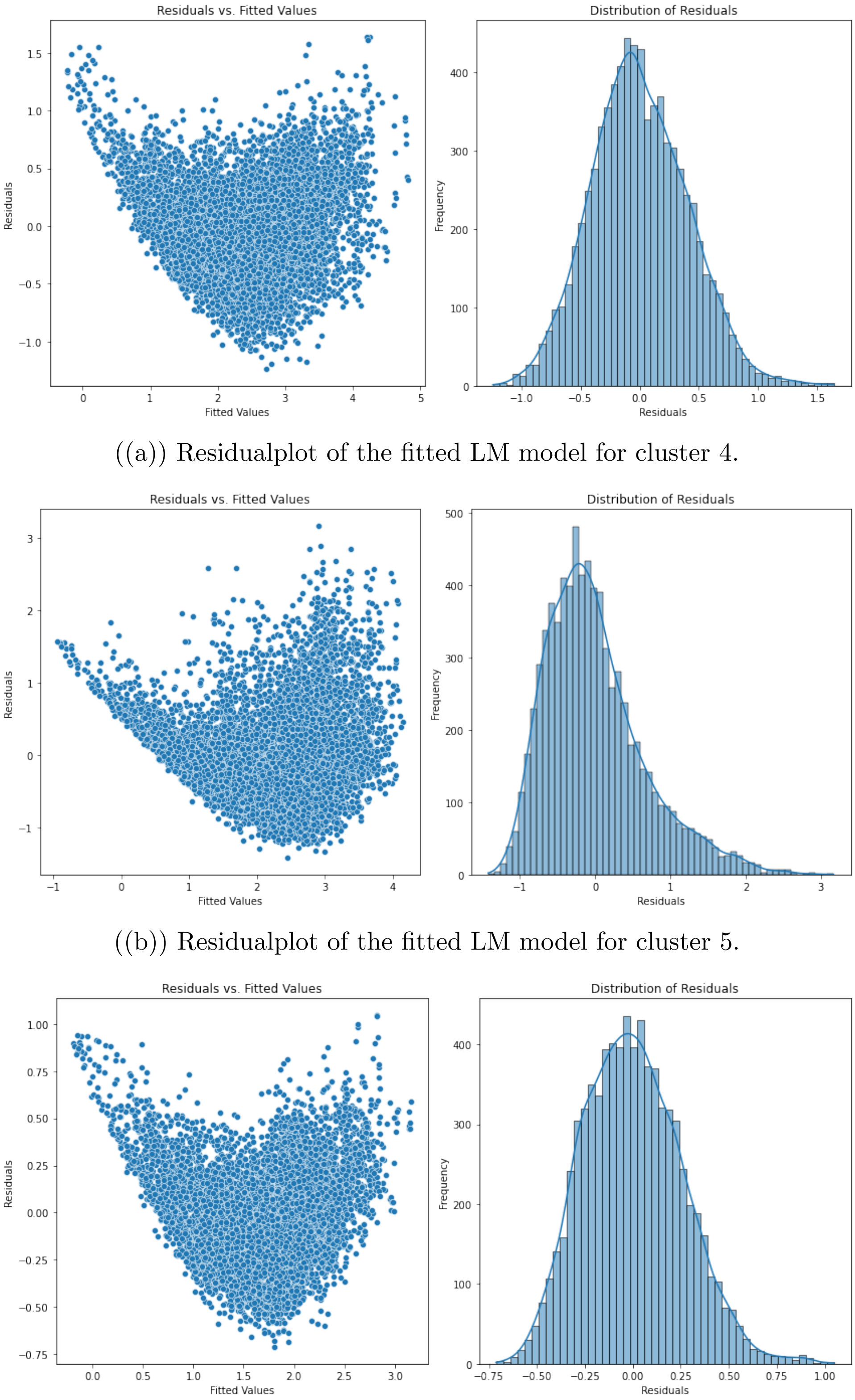


((a)) Residualplot of the fitted LM model for cluster 1.

((b)) Residualplot of the fitted LM model for cluster 2.

((c)) Residualplot of the fitted LM model for cluster 3.

Figure 37: The left-sided plot shows the residualplot of the fitted GLM model with the identity function as link function for cluster 1-3. The right-sided plot visualizes the distribution of the residuals, $r = y_i - \hat{y}_i$, for the corresponding clusters.



((a)) Residualplot of the fitted LM model for cluster 4.

((b)) Residualplot of the fitted LM model for cluster 5.

((c)) Residualplot of the fitted LM model for all the data as one cluster.

Figure 38: The left-sided plot shows the residualplot of the fitted GLM model with the identity function as link function for cluster 4, 5 and the entire data set as one cluster. The right-sided plot visualizes the distribution of the residuals, $r = y_i - \hat{y}_i$, for the corresponding clusters.

5.5 Comparization of MAE of the Different Approaches

After computing the MAE for the predicted consumption across the entire data set, a comparison was made with the MAE of the predicted consumption within clusters, including both old and new customers. A lower MAE is generally indicative of better performance. However, the observed difference is not notably significant. Despite the clustered and classified data exhibiting a lower error value compared to treating all data as a single cluster, a substantial improvement was not found. It is worth noting that while the LM may provide accurate predictions for certain clusters, its performance may vary across others, and that the new homes might not align well with the predictions of their assigned clusters. Although the results may not be groundbreaking, there is a marginal enhancement in the clustered and classified data that could be valuable for Tibber.

Figure 39 shows the predicted consumption values for the five clusters together with the actual consumption values throughout the tested time interval. The illustration reveals that the predicted values exhibit a notably larger amplitude but remain centered around the baseline of the actual consumption. Despite the model's divergence from the actual consumption, it consistently extends both above and below, resulting in an acceptable MAE, which aligns with our primary focus. Additionally, it is crucial to note that the employed LM is relatively straightforward and not intended to serve as an actual prediction model. Instead, its purpose is to check whether our clusters including classification, yield a better performance compared to a single cluster for all the data.

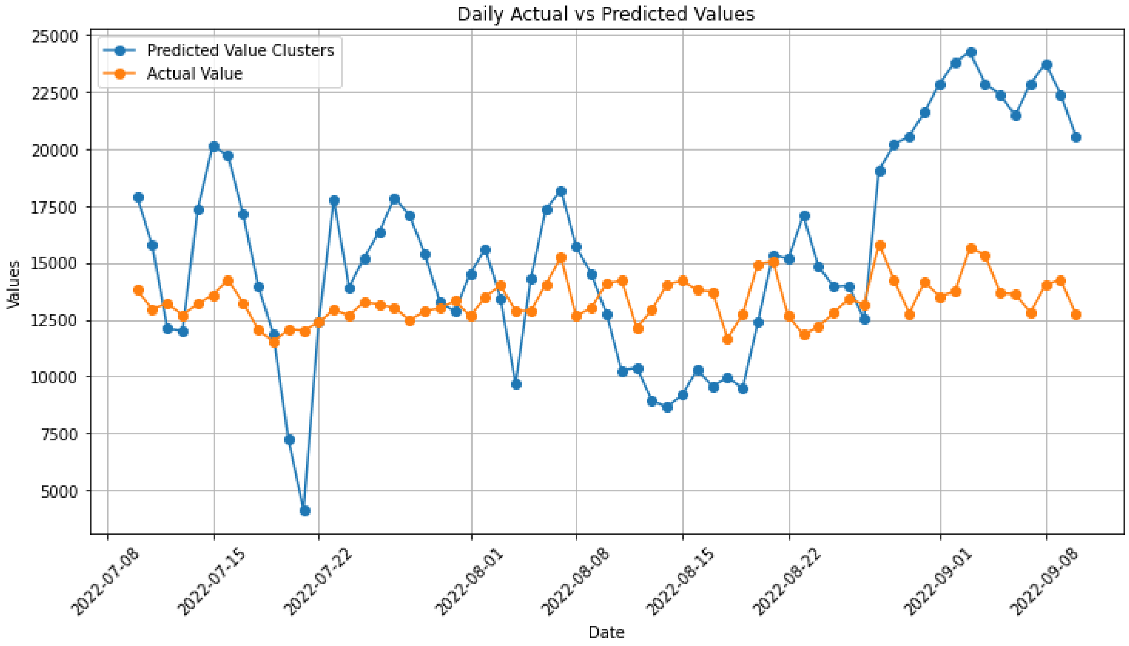


Figure 39: A plot of the predicted consumption values together with the actual consumption values in the test time interval when the five clusters were considered. The x-axis represents the days over two months, while the y-axis shows the total consumption of all households combined.

Figure 40 shows the predicted consumption values together with the actual consumption values throughout the tested time interval, but this time for the entire data set as one cluster. Since our main goal was to evaluate whether multiple clusters led to better predictions than a single-cluster approach, this comparison has been presented. An interesting observation is the near identical consumption patterns. The most noticeable difference lies in the y-axis, where the one-cluster data tends to overpredict more and have a larger amplitude span. While we expected more distinct differences in trends, this seems reasonable given that both approaches seem to have a tendency to overpredict relative to actual values and utilize a relatively simple GLM with identity function as link function. Figure 41 illustrates the disparity between the predicted values of the two approaches, revealing a observable difference that could impact the MAE. However, these variations are not substantial when compared to the aggregated overall consumption for a day, thereby resulting in minimal changes to the patterns observed in Figure 39 and 40. As illustrated in these figures, daily consumption averages around 13 000 kWh, with one of the lowest peaks at 5 000 kWh. Meanwhile, Figure 41 demonstrates that the difference between the one-cluster approach and the five-cluster approach generally remains around 500 kWh.

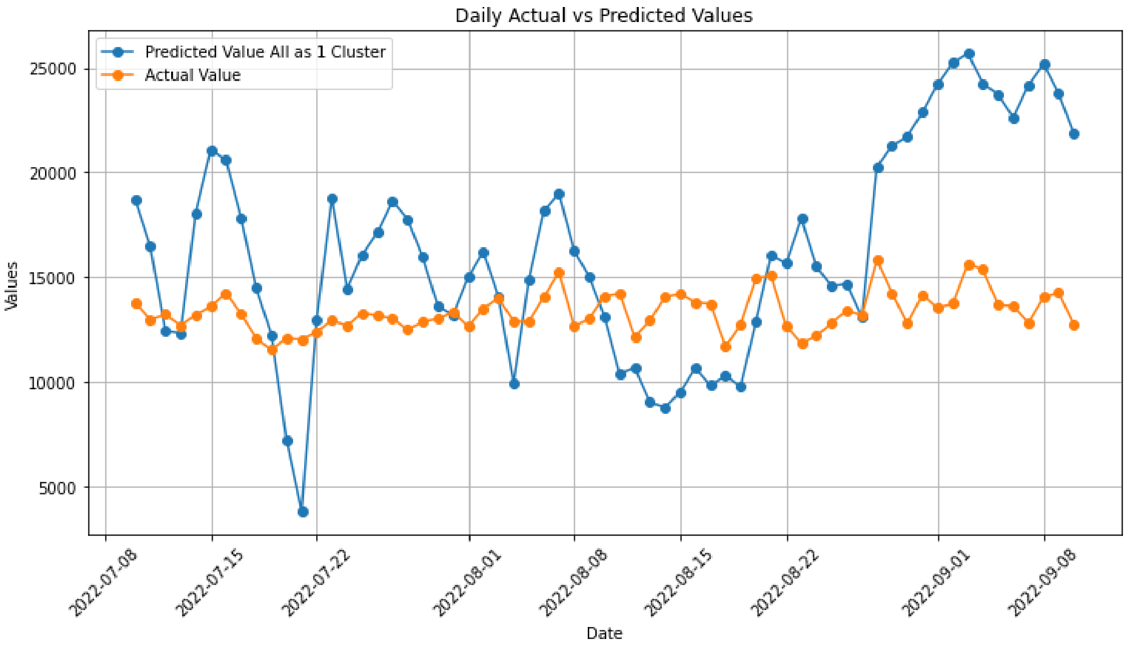


Figure 40: A plot of the predicted consumption values together with the actual consumption values in the test time interval when the entire data as one cluster was considered. The x-axis represents the days over two months, while the y-axis shows the total consumption of all households combined.

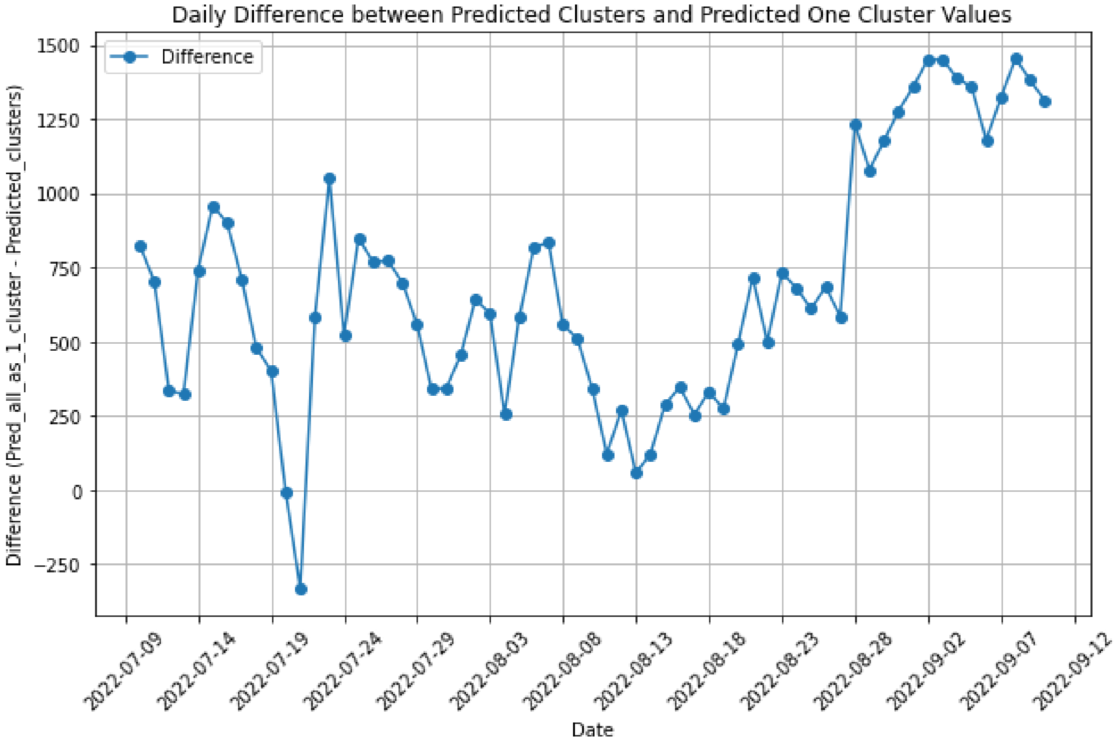


Figure 41: A plot of the difference in value between the predicted value for as one cluster and the predicted value from the five clusters. The x-axis represents the days over two months, while the y-axis shows the difference in consumption of the two clustering approaches.

Figure 42 illustrates the predicted value treating all data as a single cluster, the predicted values for the five clusters and the actual values. The plot represents hourly data within an interval, as it is easier for interpretation. It is evident that, for this timestamp, the forecast aligns with the actual consumption pattern but tends to overpredict. It is reasonable for linear models to show a wave-like pattern in daily predictions, given that the model includes temperature as a variable, and temperature will naturally have variations throughout the day. The disparity between the two predicted forecasts is not substantial, considering their reliance on relatively simple linear models. However, the forecast utilizing multiple clusters demonstrates a slightly improved prediction. This pertains to the aggregated values at those hours. In contrast, Figure 43 showcases the mean consumption for individual households, providing a more accurate representation of hourly consumption for a home.

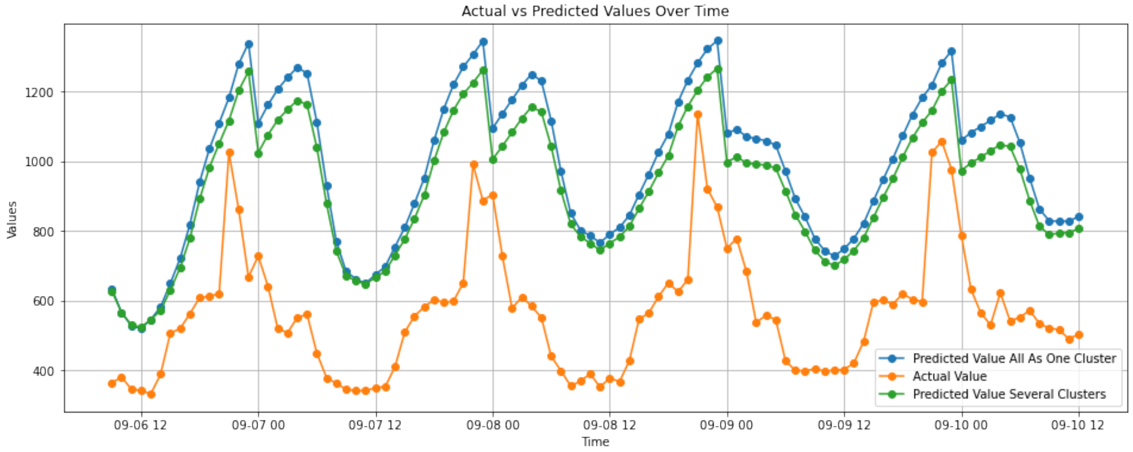


Figure 42: Plot of the value of the predicted value for clusters, the predicted value for all data as one cluster and the actual value. The x-axis represents the hours over four days, while the y-axis shows the total consumption of all households combined for each hour.

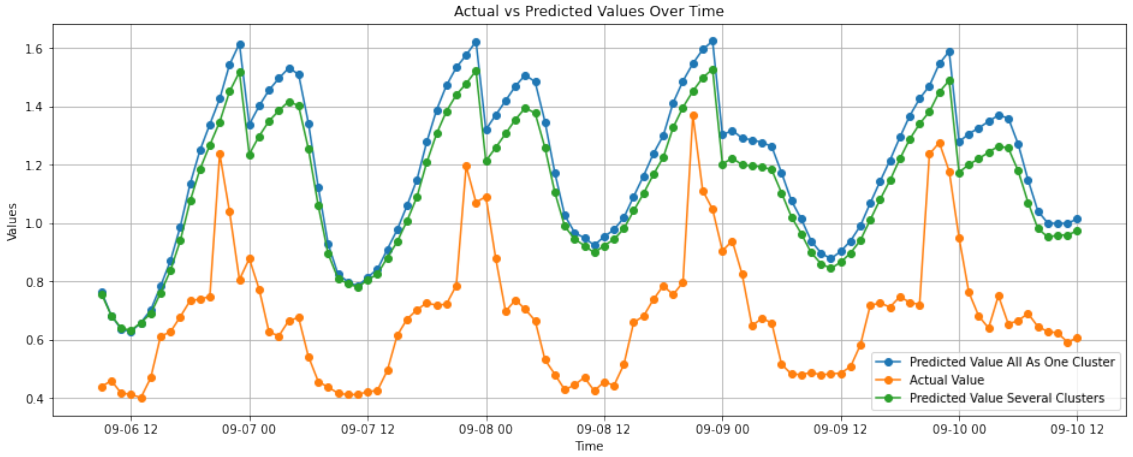


Figure 43: Plot of the value of the predicted value for clusters, the predicted value for all data as one cluster and the actual value. The x-axis represents the hours over four days, while the y-axis shows the mean consumption of one household for each hour.

The final results are presented in Table 3. Notably, the newly classified households exhibit a lower MAE and MAPE compared to considering the entire data set as a single cluster. This aligns with our expectations, signifying the beneficial impact of clusters for consumption prediction.

As outlined in Section 4.1, we explained three distinct approaches. In Table 3, approach 2 is characterized as "consumption of old clusters," being the five clusters formed from the entire data set. A surprising observation is that these old clusters predicted a higher consumption than the data set combining old homes and new homes. One might expect that providing the hierarchical method with more data would lead to better-fitted clusters. However, our observations in this case did not align with this expectation. One thinkable explanation could be the presence of outliers in the data set, affecting the optimality of the clusters. Figure 22(a) and 27(a) both highlight the relatively small size of cluster 4, suggesting a need for more effective handling of this cluster. Another consideration is the potential advantage of fitting households based on home_size and home_annual_consumption rather than the hourly consumption pattern. Alternatively, clustering homes based on their properties and subsequently constructing prediction models for these clusters might yield more effective results.

Table 3: Results Linear Model

Variable Name	***Value***
Number of households	829
Number of new households	248
Consumption of the clustered and classified data	977476.352 kWh
Consumption of the all data in one cluster	1021067.467 kWh
Consumption of old clusters	1016490.277 kWh
Actual Consumption	840133.558 kWh
MAE of the clustered and classified data	165.673 kWh
MAE of the all data in one cluster data	218.256 kWh
MAE of the old clusters data	212.734 kWh
MAPE of the clustered and classified data	0.319
MAPE of the all data in one cluster data	0.407

5.5.1 Extended Model

In the final phase of this master's thesis, an additional approach was explored, involving the incorporation of a lagged consumption variable into the linear model. This means incorporating the variables temperature, humidity, cloudiness, production, and hour, along with this new variable representing the consumption from the previous day. Since Tibber gets the historical data over time for their customers, it becomes possible to include yesterday's consumption starting from the second day of their subscription. Therefore, the lagged consumption variable in the model represents the consumption for the exact hour we are predicting, but from the day 24 hours prior. This was achieved using the pandas function shift(24) applied to the consumption data. However, to address the issue of obtaining NaN values for the first day with the shift function, the consumption values for that day were subtracted. Consequently, this adjustment resulted in a slightly lower overall consumption, as depicted in Table 4.

Figure 44 and 45 exhibit distinct patterns compared to Figure 39 and 40. Moreover, these figures depict smaller consumption values than the earlier models.

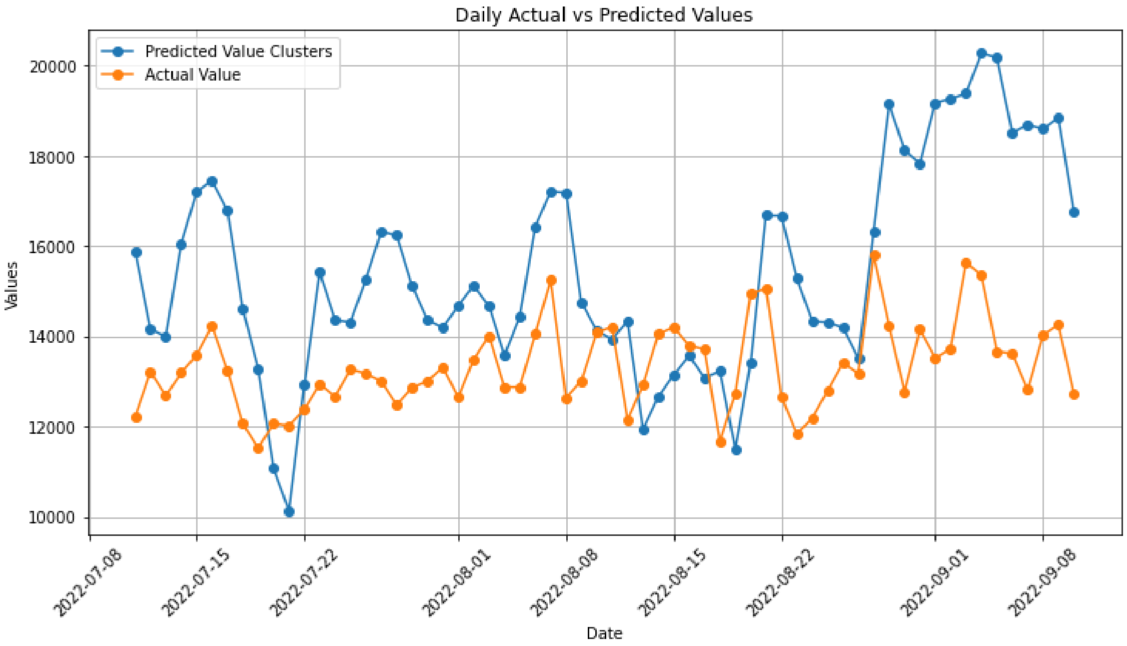


Figure 44: A plot of the predicted consumption values together with the actual consumption values in the test time interval when the five cluster with the modified model were considered. The x-axis represents the days over two months, while the y-axis shows the total consumption of all households combined.

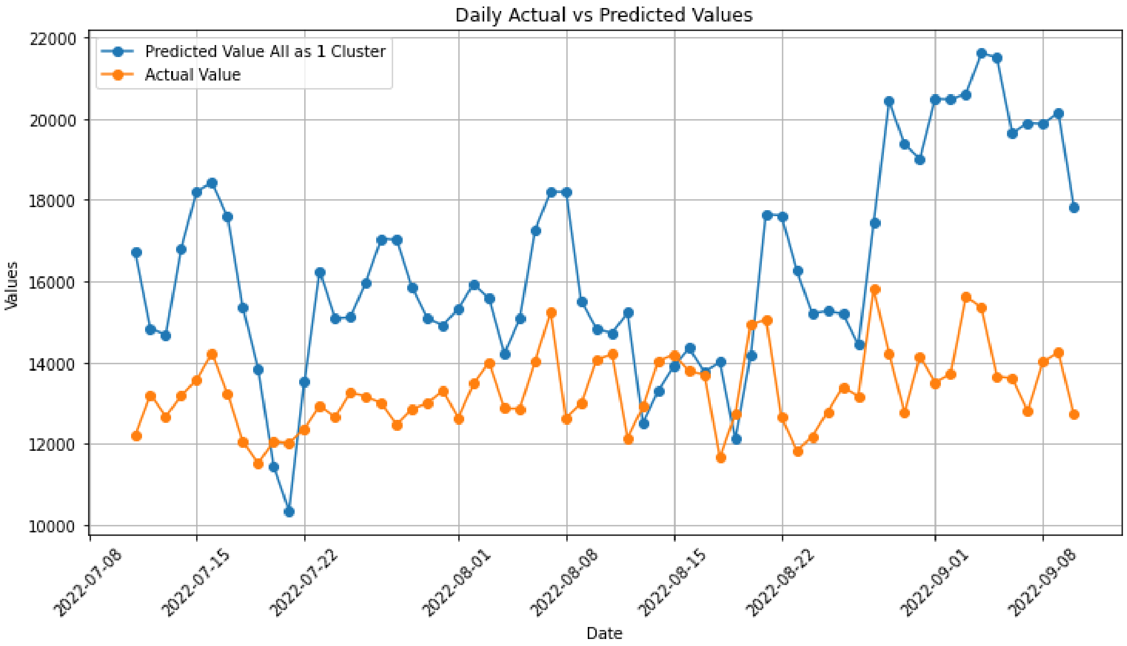


Figure 45: A plot of the predicted consumption values together with the actual consumption values in the test time interval when the entire data as one cluster with the modified model was considered. The x-axis represents the days over two months, while the y-axis shows the total consumption of all households combined.

In Figure 46, the difference in predicted values between the one-cluster approach and the five-cluster approach is illustrated. A notable difference observed in comparison to Figure 41 is a subtle change in pattern, and notably, there is no negative difference. Furthermore, the modified model appears to exhibit slightly higher differences, indicating more variation in predictions between the one-cluster and five-cluster approach achieved with the modified model.

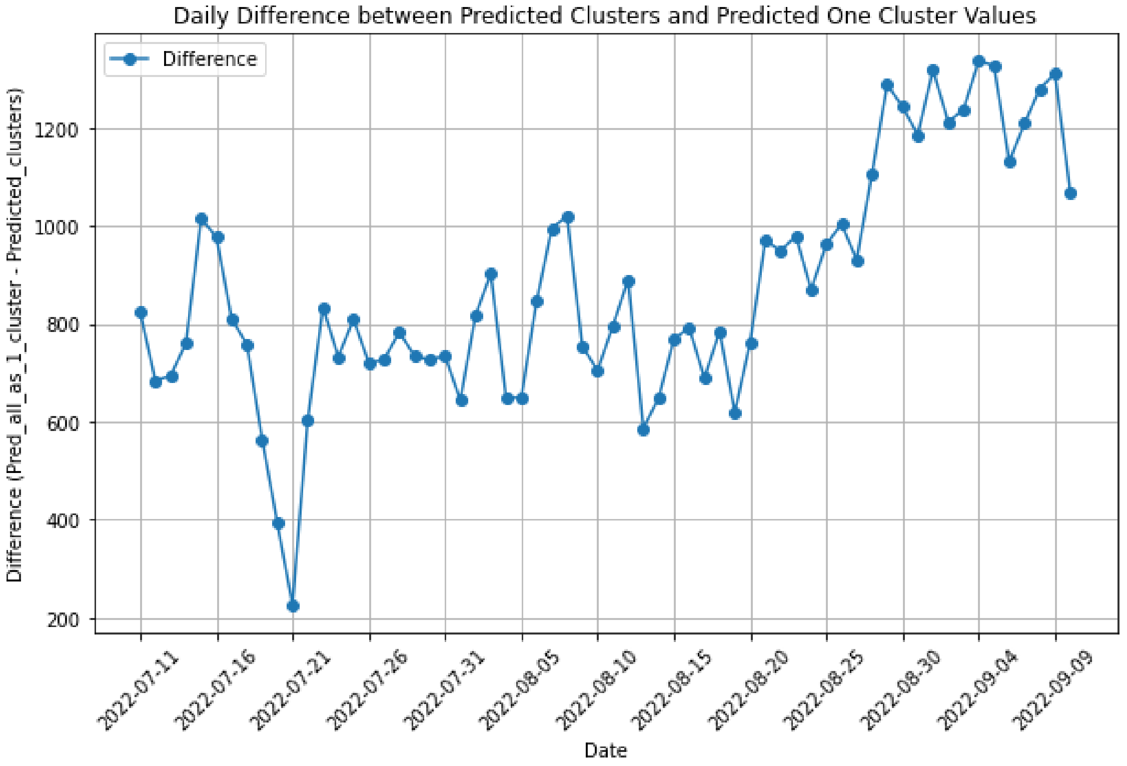


Figure 46: A plot of the difference in value between the predicted value for as one cluster and the predicted value from the five clusters with the modified model. The x-axis represents the days over two months, while the y-axis shows the difference in consumption of the two clustering approaches.

Figure 47 illustrates the predicted value treating all data as a single cluster, the predicted values for the five clusters and the actual values. The time interval aligns with the same x-axis as in Figure 43 and it seems as the extended linear model is better a fit for this timestamp. The modified model also follows a consumption pattern more closely aligned with the actual consumption. This is logical as the model incorporates temperature as earlier, but also has information about yesterday's consumption. The LM results of the modified model, provided in Appendix B, reveal that the consumption_lag variable has a sizable coefficient for all the clusters, indicating a substantial influence on the linear models.

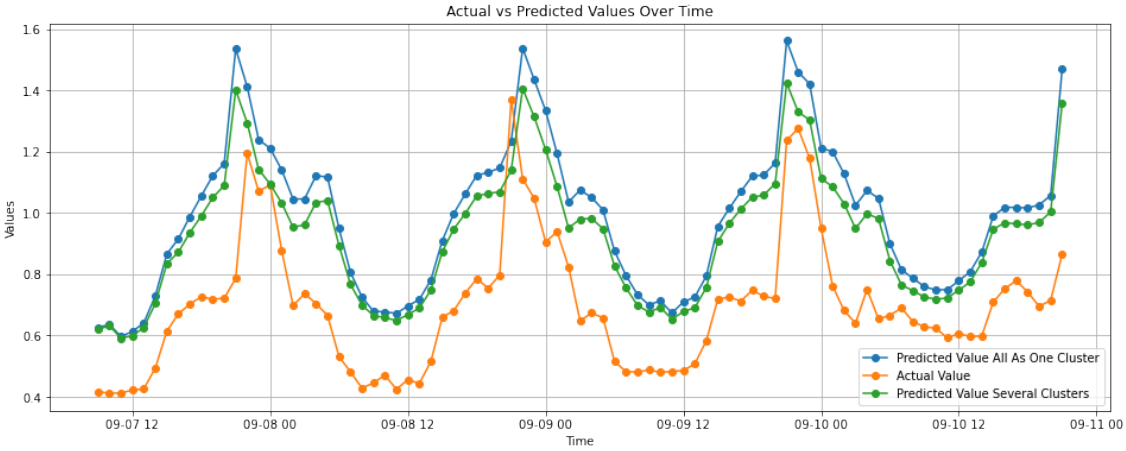


Figure 47: Plot of the value of the predicted value for clusters, the predicted value for all data as one cluster and the actual value with the modified model. The x-axis represents the hours over four days, while the y-axis shows the mean consumption of one household for each hour.

Table 4 displays the values obtained for the modified linear model. We observe a slight improvement in the MAE and MAPE for the five clustered data, incorporating the new homes. As one day is excluded, the consumption values in the rows are lower compared to those in Table 3. It is noteworthy that, for this model, the MAE and MAPE increased when considering all the data as a single cluster.

Table 4: Results Modified Model

Variable Name	***Value***
Consumption of the clustered and classified data	957960.644 kWh
Consumption of the all data in one cluster	1012297.217 kWh
Actual Consumption	825597.386 kWh
MAE of the clustered and classified data	159.666 kWh
MAE of the all data in one cluster data	225.211 kWh
MAPE of the clustered and classified data	0.311
MAPE of the all data in one cluster data	0.422

6 Concluding Remarks

In this section, we will draw conclusions from the thesis and examine potential paths for future research.

6.1 Conclusion

In this thesis, an analysis of the customer data provided by Tibber has been conducted. The training set, comprising 70% of the provided dataset has been clustered by their hourly mean consumption for a day from the time series data set. Then, the test set, representing new customers, has been classified into these established clusters based on their home properties. The chosen home properties variables for classifying new customers were home_size and home_annual_consumption, both standardized. This choice was made because these variables demonstrated to influence the assignment of groups the most. Ultimately, the selected classification model was LDA, given its consistent and high-performance results across various tests.

In our final trial, 248 households were randomly selected as new homes out of the 829 available households. The prediction was obtained using the five clusters prediction models, multiplying the results by the number of homes in each cluster, including both the already grouped homes and the newly classified ones. The MAE for the five clusters including the new homes was documented at 165.7 kWh. This indicate an improvement compared to the MAE of the entire data set treated as one cluster, which was 218.3 kWh. The prediction outcome from the clustered and classified households resulted in a MAPE of 0.319. Towards the end of the thesis, a variable was added into our linear model, being a 24-hour prior consumption variable. This modification led to an improvement for the five clusters, incorporating both old and new homes, in the linear model. In this case, the MAE resulted in 159.7 kWh and a MAPE of 0.311. Once again, the MAE was lower than the one-cluster approach for the entire dataset. This emphasizes that in both cases, whether with the initial linear model or the modified one, the approach with multiple clusters performed better.

The discoveries of this master's thesis propose that the results obtained from classifying new households could be beneficial for Tibber in enhancing their customers' consumption predictions. While there may not be an overwhelmingly large difference between the several clusters and a single cluster, and the predictions may not align perfectly with the actual consumption, the observed improvement is noteworthy and could be of interest. Certainly, for Tibber's benefit, it is advisable to conduct additional tests and implement a testing phase to determine whether this information is genuinely valuable and effective. It is essential to recognize that the current findings represent only a subset of their customer base, and the results may vary based on the country or price area.

6.2 Recommendations for Future Work

The master thesis could have delved deeper into various aspects or considered additional factors that might be of interest for future research.

While the classification methods focused solely on SE3 among the examined variables, future research could benefit from exploring other regions within Sweden and investigating additional price areas within Tibber's extensive customer base, considering that they have customers in other countries as well. Given the size of Tibber's customer base and the amount of available data, there is potential to broaden the scope of analysis.

Moreover, the master thesis used the same clustering methods employed earlier in the project, incorporating hierarchical clustering. However, there are numerous alternative clustering techniques that could have been explored. Diversifying the clustering approaches may offer new insights into the data. In the context of clustering, the primary focus was on the hourly consumption as the key response variable. Nevertheless, other variables could have been considered, and exploring different clustering methods beyond consumption alone could enhance the analysis. Additionally, clustering over a year of data was performed, but investigating seasonal clustering might have provided more informative results, considering the potential impact of seasonal variations.

There are, as always in data collecting, missing values. For new customers we might not have all the information used in our analysis. In some rare occations, the customer do not have their yearly consumption, which played a signifcant role in this analysis. Therefore, there could have been more testing with different approaches, considering the occasional unavailability of desired customer information.

Finally, the extended model presented in the end of the results could have been further developed to achieve greater optimization. Since this model was only tested in the concluding phase of the master's thesis, there was insufficient time to explore its full potential. In our current implementation, we incorporated only the 24-hour prior consumption into the model. However, in practice, all hourly consumption data from the initial subscription period could have been integrated into the model. This could involve utilizing more advanced time series models, such as Autoregressive (AR), Moving Average (MA), or a combination of both (ARMA). While the primary focus of this study was not on identifying the best prediction model but on investigating the potential benefits of predicting in clusters, delving deeper into model optimization could be interesting for future research.

Bibliography

[1] *A Map of the Overview of the Nord Pool Market Coupling.* 2023. URL: https://www.researchgate.net/figure/A-map-of-the-overview-of-the-Nord-Pool-market-coupling_fig1_348486579 (visited on 3rd Nov. 2023).

[2] Jason Brownlee. *Power Transforms with scikit-learn.* URL: https://machinelearningmastery.com/power-transforms-with-scikit-learn/ (visited on 12th Oct. 2023).

[3] Claudia Czado. 'Generalized linear models with applications'. unpublished lecture notes, can be given if requested. 2022.

[4] *Data from the power system.* en. Nov. 2023. URL: https://www.statnett.no/en/for-stakeholders-in-the-power-industry/data-from-the-power-system/ (visited on 15th Dec. 2023).

[5] Sachin Date. *What Are Dummy Variables And How To Use Them In A Regression Model.* en. July 2022. URL: https://timeseriesreasoning.com/contents/dummy-variables-in-a-regression-model/ (visited on 15th Nov. 2023).

[6] *Definitive Guide to Hierarchical Clustering with Python and Scikit-Learn.* 2022. URL: https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/ (visited on 1st Nov. 2022).

[7] Moussa Doumbia. *Elbow Method in Supervised Learning(Optimal K Value).* en. Aug. 2019. URL: https://medium.com/@moussadoumbia_90919/elbow-method-in-supervised-learning-optimal-k-value-99d425f229e7 (visited on 9th Dec. 2023).

[8] Ludwig Fahrmeir, Thomas Kneib, Stefan Lang and Brian Marx. *Regression: Models, Methods and Applications.* en. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. ISBN: 978-3-642-34332-2 978-3-642-34333-9. DOI: 10.1007/978-3-642-34333-9. URL: https://link.springer.com/10.1007/978-3-642-34333-9 (visited on 6th Dec. 2023).

[9] *Fig. 6. Support Vector Machine visualization.* en. URL: https://www.researchgate.net/figure/Support-Vector-Machine-visualization_fig5_332248436 (visited on 9th Dec. 2023).

[10] GeeksforGeeks. *Support Vector Machine Algorithm.* URL: %5Curl%7Bhttps://www.geeksforgeeks.org/support-vector-machine-algorithm/%7D (visited on 9th Nov. 2023).

[11] Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. *An Introduction to Statistical Learning.* Springer, 2013.

[12] Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis - Second Edition.* Prentice-Hall, 1988.

[13] *Mean Absolute Percentage Error (MAPE).* 2021. URL: https://www.statisticshowto.com/mean-absolute-percentage-error-mape/ (visited on 5th Dec. 2023).

[14] *Pandas – Filling NaN in Categorical data.* en-US. Section: Python. Apr. 2021. URL: https://www.geeksforgeeks.org/pandas-filling-nan-in-categorical-data/ (visited on 14th Nov. 2023).

[15] Sruthi E. R. *Understand Random Forest Algorithms With Examples (Updated 2023)*. en. June 2021. URL: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/ (visited on 9th Dec. 2023).

[16] Astrid Solheim. *Clustering households after their electrical consumption*. 2023.

[17] *Standardized linear regression*. URL: https://www.statlect.com/fundamentals-of-statistics/linear-regression-with-standardized-variables# (visited on 9th Dec. 2023).

[18] *step: Choose a model by AIC in a Stepwise Algorithm*. en. URL: https://rdrr.io/r/stats/step.html (visited on 16th Nov. 2023).

[19] Thomas Strohmer. *Clustering*. URL: https://www.math.ucdavis.edu/~strohmer/courses/180BigData/180lecture_kmeans.pdf (visited on 31st Oct. 2022).

[20] *Strømprisprognose Tibber*. nb-NO. URL: https://tibber.com/no,%20https://tibber.com/no/stromprisprognose (visited on 15th Dec. 2023).

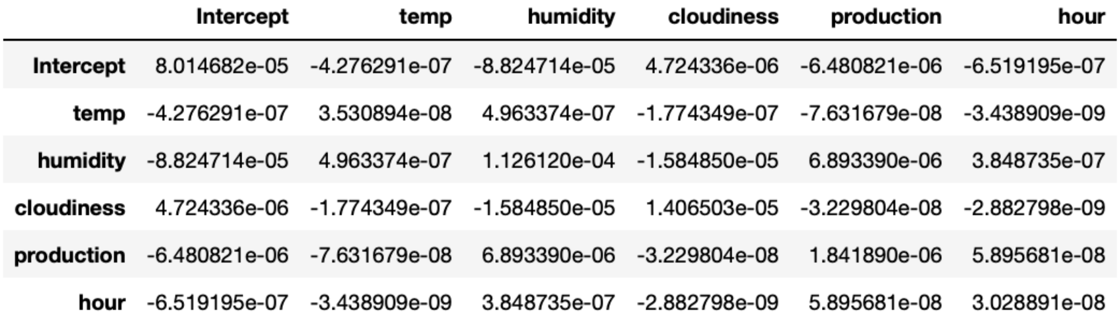
[21] Tibber. *Smart Heating*. 2022. URL: https://tibber.com/en/magazine/power-hacks/smart-heating (visited on 2nd Nov. 2023).

[22] *Tibber - Smart Energy Company*. URL: %5Curl%7Bhttps://tibber.com/no%7D (visited on 7th Nov. 2023).

[23] Walter Zucchini. *Kernel Density Estimation*. 2003. URL: http://staff.ustc.edu.cn/~zwp/teach/Math-Stat/kernel.pdf.

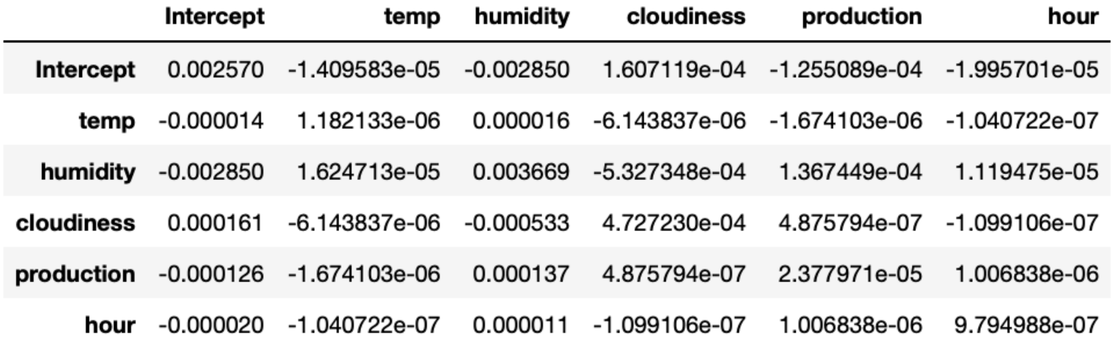
Appendix

A Covariance matrices from LM Results



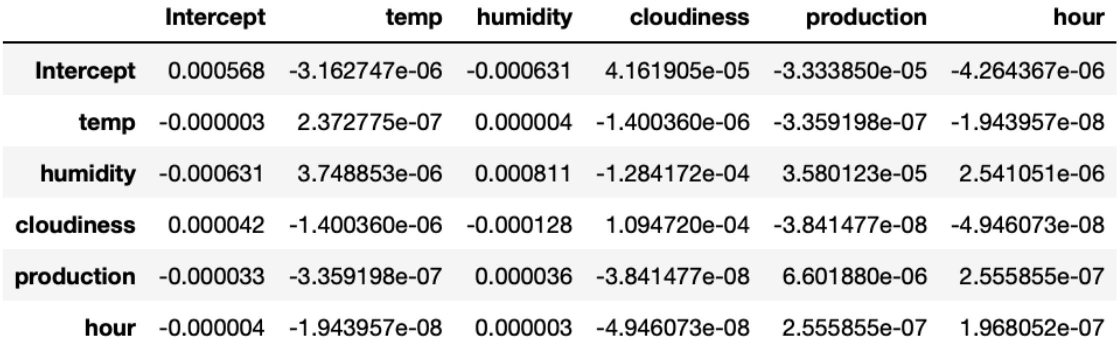
	Intercept	temp	humidity	cloudiness	production	hour
Intercept	8.014682e-05	-4.276291e-07	-8.824714e-05	4.724336e-06	-6.480821e-06	-6.519195e-07
temp	-4.276291e-07	3.530894e-08	4.963374e-07	-1.774349e-07	-7.631679e-08	-3.438909e-09
humidity	-8.824714e-05	4.963374e-07	1.126120e-04	-1.584850e-05	6.893390e-06	3.848735e-07
cloudiness	4.724336e-06	-1.774349e-07	-1.584850e-05	1.406503e-05	-3.229804e-08	-2.882798e-09
production	-6.480821e-06	-7.631679e-08	6.893390e-06	-3.229804e-08	1.841890e-06	5.895681e-08
hour	-6.519195e-07	-3.438909e-09	3.848735e-07	-2.882798e-09	5.895681e-08	3.028891e-08

Cluster 1



	Intercept	temp	humidity	cloudiness	production	hour
Intercept	0.002570	-1.409583e-05	-0.002850	1.607119e-04	-1.255089e-04	-1.995701e-05
temp	-0.000014	1.182133e-06	0.000016	-6.143837e-06	-1.674103e-06	-1.040722e-07
humidity	-0.002850	1.624713e-05	0.003669	-5.327348e-04	1.367449e-04	1.119475e-05
cloudiness	0.000161	-6.143837e-06	-0.000533	4.727230e-04	4.875794e-07	-1.099106e-07
production	-0.000126	-1.674103e-06	0.000137	4.875794e-07	2.377971e-05	1.006838e-06
hour	-0.000020	-1.040722e-07	0.000011	-1.099106e-07	1.006838e-06	9.794988e-07

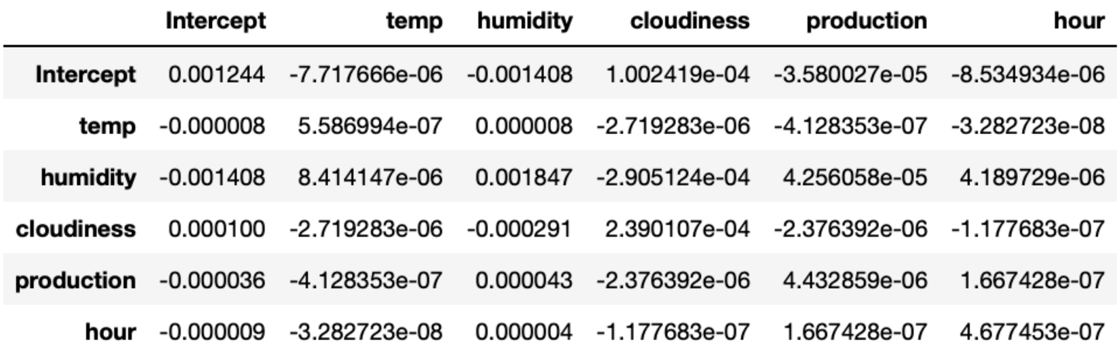
Cluster 2



	Intercept	temp	humidity	cloudiness	production	hour
Intercept	0.000568	-3.162747e-06	-0.000631	4.161905e-05	-3.333850e-05	-4.264367e-06
temp	-0.000003	2.372775e-07	0.000004	-1.400360e-06	-3.359198e-07	-1.943957e-08
humidity	-0.000631	3.748853e-06	0.000811	-1.284172e-04	3.580123e-05	2.541051e-06
cloudiness	0.000042	-1.400360e-06	-0.000128	1.094720e-04	-3.841477e-08	-4.946073e-08
production	-0.000033	-3.359198e-07	0.000036	-3.841477e-08	6.601880e-06	2.555855e-07
hour	-0.000004	-1.943957e-08	0.000003	-4.946073e-08	2.555855e-07	1.968052e-07

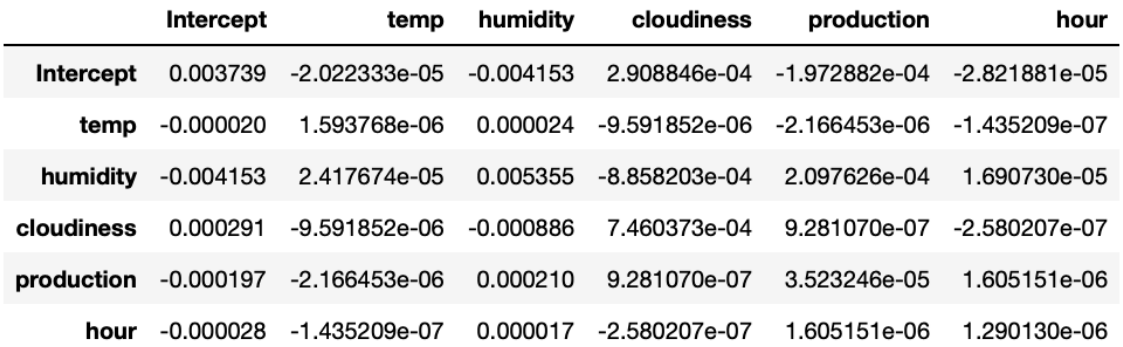
Cluster 3

Figure 48: Covariance matrices from the results of the GLM with identity function as link function for the clusters 1-3.



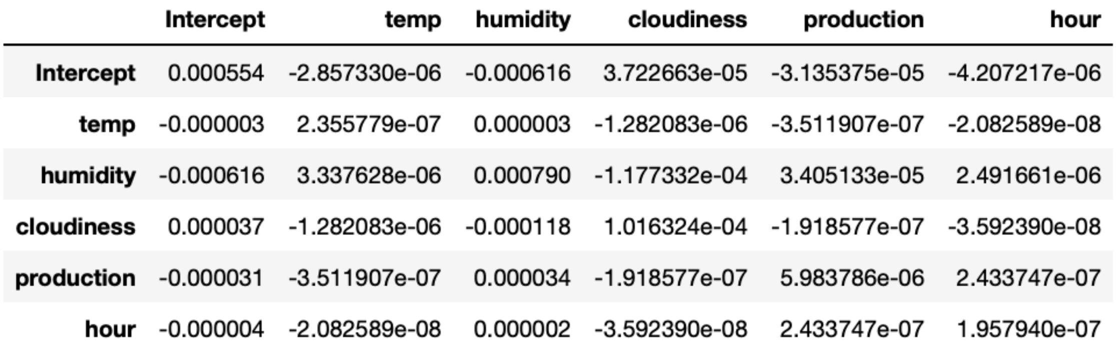
	Intercept	temp	humidity	cloudiness	production	hour
Intercept	0.001244	-7.717666e-06	-0.001408	1.002419e-04	-3.580027e-05	-8.534934e-06
temp	-0.000008	5.586994e-07	0.000008	-2.719283e-06	-4.128353e-07	-3.282723e-08
humidity	-0.001408	8.414147e-06	0.001847	-2.905124e-04	4.256058e-05	4.189729e-06
cloudiness	0.000100	-2.719283e-06	-0.000291	2.390107e-04	-2.376392e-06	-1.177683e-07
production	-0.000036	-4.128353e-07	0.000043	-2.376392e-06	4.432859e-06	1.667428e-07
hour	-0.000009	-3.282723e-08	0.000004	-1.177683e-07	1.667428e-07	4.677453e-07

Cluster 4



	Intercept	temp	humidity	cloudiness	production	hour
Intercept	0.003739	-2.022333e-05	-0.004153	2.908846e-04	-1.972882e-04	-2.821881e-05
temp	-0.000020	1.593768e-06	0.000024	-9.591852e-06	-2.166453e-06	-1.435209e-07
humidity	-0.004153	2.417674e-05	0.005355	-8.858203e-04	2.097626e-04	1.690730e-05
cloudiness	0.000291	-9.591852e-06	-0.000886	7.460373e-04	9.281070e-07	-2.580207e-07
production	-0.000197	-2.166453e-06	0.000210	9.281070e-07	3.523246e-05	1.605151e-06
hour	-0.000028	-1.435209e-07	0.000017	-2.580207e-07	1.605151e-06	1.290130e-06

Cluster 5

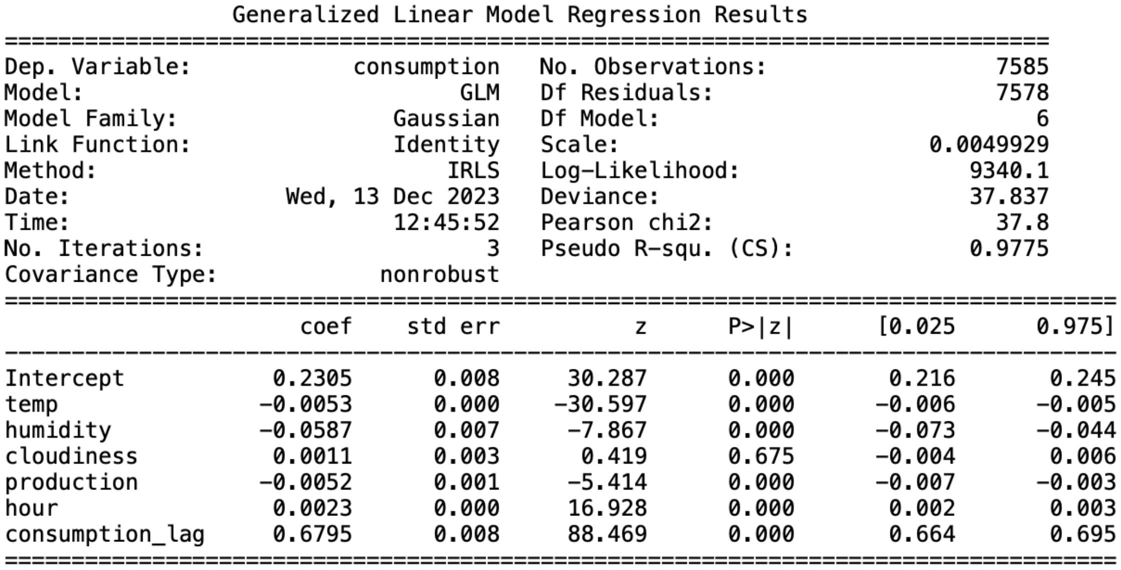


	Intercept	temp	humidity	cloudiness	production	hour
Intercept	0.000554	-2.857330e-06	-0.000616	3.722663e-05	-3.135375e-05	-4.207217e-06
temp	-0.000003	2.355779e-07	0.000003	-1.282083e-06	-3.511907e-07	-2.082589e-08
humidity	-0.000616	3.337628e-06	0.000790	-1.177332e-04	3.405133e-05	2.491661e-06
cloudiness	0.000037	-1.282083e-06	-0.000118	1.016324e-04	-1.918577e-07	-3.592390e-08
production	-0.000031	-3.511907e-07	0.000034	-1.918577e-07	5.983786e-06	2.433747e-07
hour	-0.000004	-2.082589e-08	0.000002	-3.592390e-08	2.433747e-07	1.957940e-07

All the data as one cluster

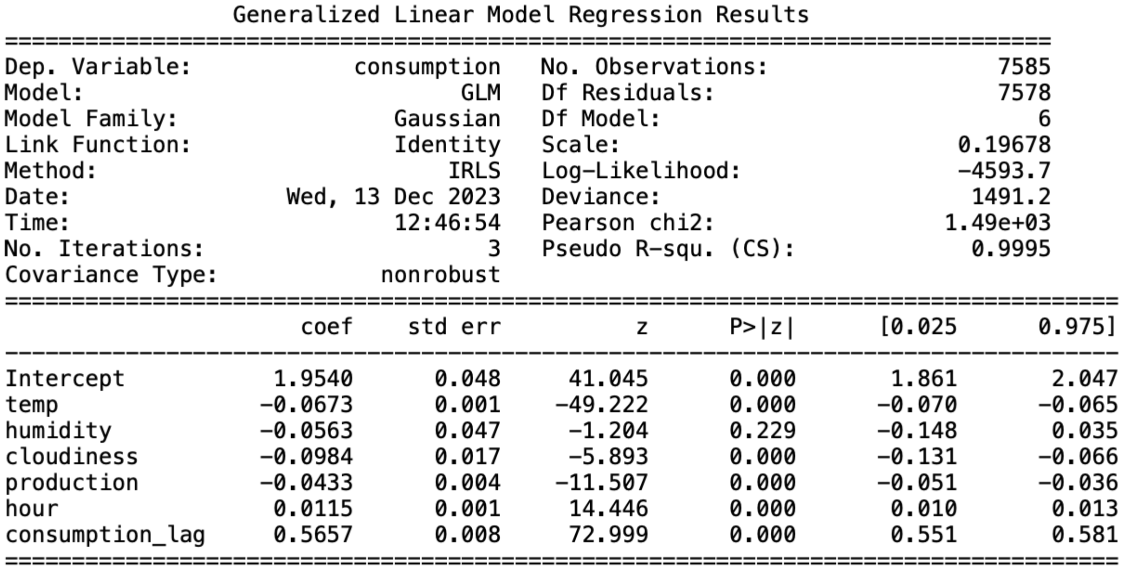
Figure 49: Covariance matrices from the results of the GLM with identity function as link function for cluster 4 and the entire data as one cluster.

B LM Results from the Modified Model



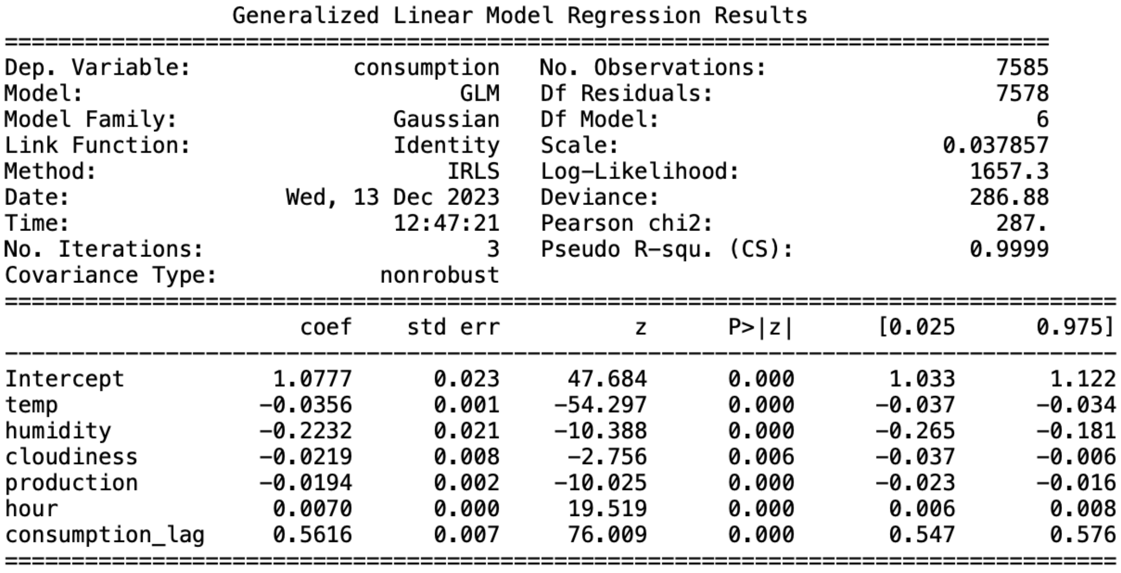
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:            consumption   No. Observations:                 7585
Model:                            GLM   Df Residuals:                     7578
Model Family:                Gaussian   Df Model:                            6
Link Function:               Identity   Scale:                       0.0049929
Method:                          IRLS   Log-Likelihood:                 9340.1
Date:                Wed, 13 Dec 2023   Deviance:                       37.837
Time:                        12:45:52   Pearson chi2:                     37.8
No. Iterations:                     3   Pseudo R-squ. (CS):             0.9775
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------
Intercept           0.2305      0.008     30.287      0.000       0.216       0.245
temp               -0.0053      0.000    -30.597      0.000      -0.006      -0.005
humidity           -0.0587      0.007     -7.867      0.000      -0.073      -0.044
cloudiness          0.0011      0.003      0.419      0.675      -0.004       0.006
production         -0.0052      0.001     -5.414      0.000      -0.007      -0.003
hour                0.0023      0.000     16.928      0.000       0.002       0.003
consumption_lag     0.6795      0.008     88.469      0.000       0.664       0.695
===================================================================================
```

Figure 50: Summary of the modified linear model of the first cluster.



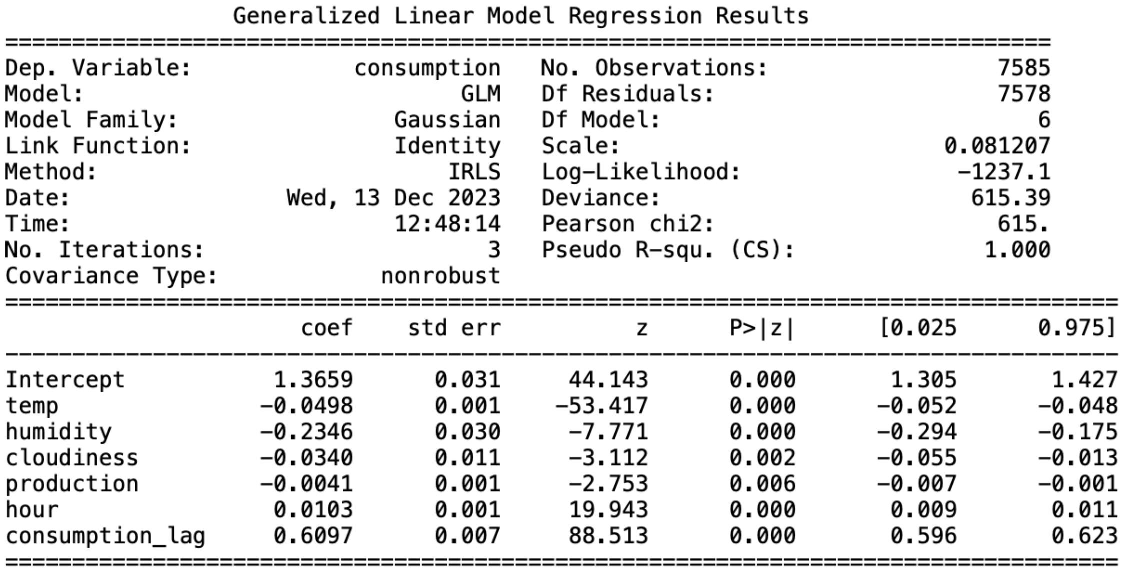
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:            consumption   No. Observations:                 7585
Model:                            GLM   Df Residuals:                     7578
Model Family:                Gaussian   Df Model:                            6
Link Function:               Identity   Scale:                         0.19678
Method:                          IRLS   Log-Likelihood:                -4593.7
Date:                Wed, 13 Dec 2023   Deviance:                       1491.2
Time:                        12:46:54   Pearson chi2:                 1.49e+03
No. Iterations:                     3   Pseudo R-squ. (CS):             0.9995
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------
Intercept           1.9540      0.048     41.045      0.000       1.861       2.047
temp               -0.0673      0.001    -49.222      0.000      -0.070      -0.065
humidity           -0.0563      0.047     -1.204      0.229      -0.148       0.035
cloudiness         -0.0984      0.017     -5.893      0.000      -0.131      -0.066
production         -0.0433      0.004    -11.507      0.000      -0.051      -0.036
hour                0.0115      0.001     14.446      0.000       0.010       0.013
consumption_lag     0.5657      0.008     72.999      0.000       0.551       0.581
===================================================================================
```

Figure 51: Summary of the modified linear model of the second cluster.



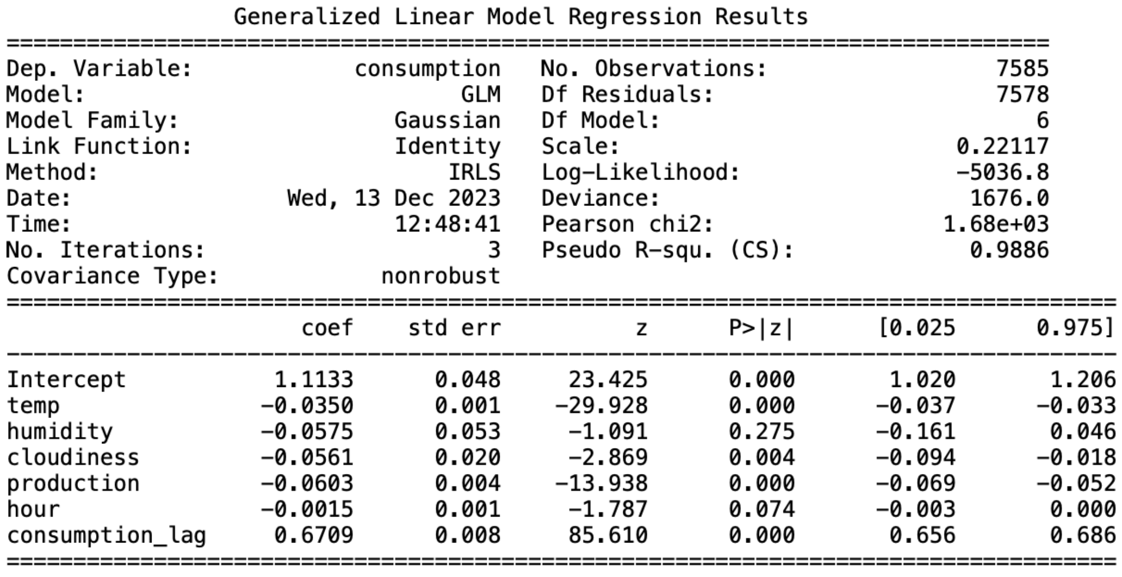
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:            consumption   No. Observations:                 7585
Model:                            GLM   Df Residuals:                     7578
Model Family:                Gaussian   Df Model:                            6
Link Function:               Identity   Scale:                        0.037857
Method:                          IRLS   Log-Likelihood:                 1657.3
Date:                Wed, 13 Dec 2023   Deviance:                       286.88
Time:                        12:47:21   Pearson chi2:                     287.
No. Iterations:                     3   Pseudo R-squ. (CS):             0.9999
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------
Intercept           1.0777      0.023     47.684      0.000       1.033       1.122
temp               -0.0356      0.001    -54.297      0.000      -0.037      -0.034
humidity           -0.2232      0.021    -10.388      0.000      -0.265      -0.181
cloudiness         -0.0219      0.008     -2.756      0.006      -0.037      -0.006
production         -0.0194      0.002    -10.025      0.000      -0.023      -0.016
hour                0.0070      0.000     19.519      0.000       0.006       0.008
consumption_lag     0.5616      0.007     76.009      0.000       0.547       0.576
===================================================================================
```

Figure 52: Summary of the modified linear model of the third cluster.



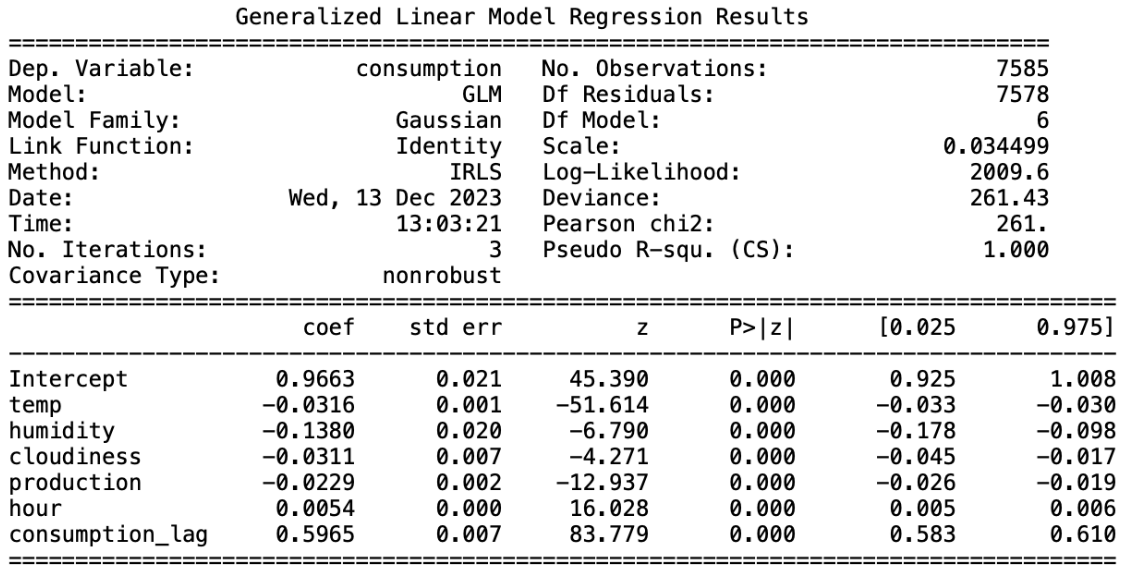
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:            consumption   No. Observations:                 7585
Model:                            GLM   Df Residuals:                     7578
Model Family:                Gaussian   Df Model:                            6
Link Function:               Identity   Scale:                        0.081207
Method:                          IRLS   Log-Likelihood:                -1237.1
Date:                Wed, 13 Dec 2023   Deviance:                       615.39
Time:                        12:48:14   Pearson chi2:                     615.
No. Iterations:                     3   Pseudo R-squ. (CS):              1.000
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------
Intercept           1.3659      0.031     44.143      0.000       1.305       1.427
temp               -0.0498      0.001    -53.417      0.000      -0.052      -0.048
humidity           -0.2346      0.030     -7.771      0.000      -0.294      -0.175
cloudiness         -0.0340      0.011     -3.112      0.002      -0.055      -0.013
production         -0.0041      0.001     -2.753      0.006      -0.007      -0.001
hour                0.0103      0.001     19.943      0.000       0.009       0.011
consumption_lag     0.6097      0.007     88.513      0.000       0.596       0.623
===================================================================================
```

Figure 53: Summary of the modified linear model of the fourth cluster.



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:            consumption   No. Observations:                 7585
Model:                            GLM   Df Residuals:                     7578
Model Family:                Gaussian   Df Model:                            6
Link Function:               Identity   Scale:                         0.22117
Method:                          IRLS   Log-Likelihood:                -5036.8
Date:                Wed, 13 Dec 2023   Deviance:                       1676.0
Time:                        12:48:41   Pearson chi2:                 1.68e+03
No. Iterations:                     3   Pseudo R-squ. (CS):             0.9886
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------
Intercept           1.1133      0.048     23.425      0.000       1.020       1.206
temp               -0.0350      0.001    -29.928      0.000      -0.037      -0.033
humidity           -0.0575      0.053     -1.091      0.275      -0.161       0.046
cloudiness         -0.0561      0.020     -2.869      0.004      -0.094      -0.018
production         -0.0603      0.004    -13.938      0.000      -0.069      -0.052
hour               -0.0015      0.001     -1.787      0.074      -0.003       0.000
consumption_lag     0.6709      0.008     85.610      0.000       0.656       0.686
===================================================================================
```

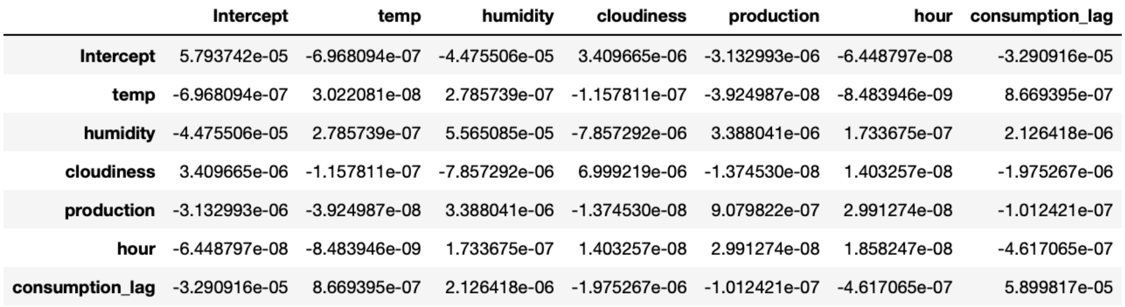
Figure 54: Summary of the modified linear model of the fifth cluster.



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:            consumption   No. Observations:                 7585
Model:                            GLM   Df Residuals:                     7578
Model Family:                Gaussian   Df Model:                            6
Link Function:               Identity   Scale:                        0.034499
Method:                          IRLS   Log-Likelihood:                 2009.6
Date:                Wed, 13 Dec 2023   Deviance:                       261.43
Time:                        13:03:21   Pearson chi2:                     261.
No. Iterations:                     3   Pseudo R-squ. (CS):              1.000
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------
Intercept           0.9663      0.021     45.390      0.000       0.925       1.008
temp               -0.0316      0.001    -51.614      0.000      -0.033      -0.030
humidity           -0.1380      0.020     -6.790      0.000      -0.178      -0.098
cloudiness         -0.0311      0.007     -4.271      0.000      -0.045      -0.017
production         -0.0229      0.002    -12.937      0.000      -0.026      -0.019
hour                0.0054      0.000     16.028      0.000       0.005       0.006
consumption_lag     0.5965      0.007     83.779      0.000       0.583       0.610
===================================================================================
```

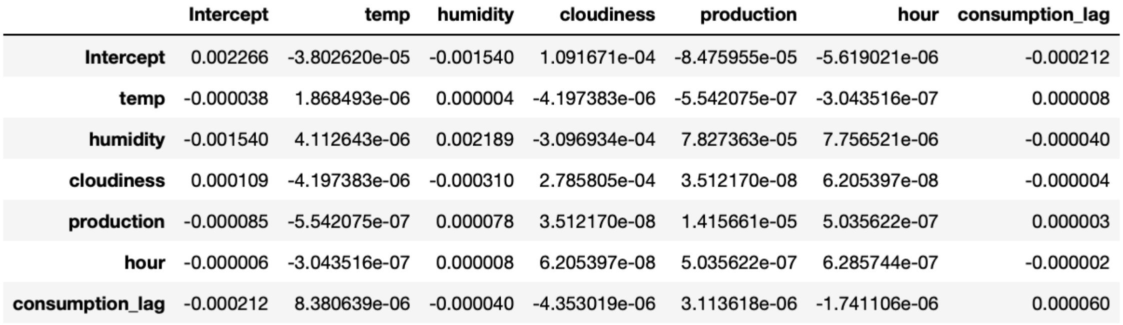
Figure 55: Summary of the modified linear model of the entire data as one cluster.

C Covariance matrices from Modified LM Results



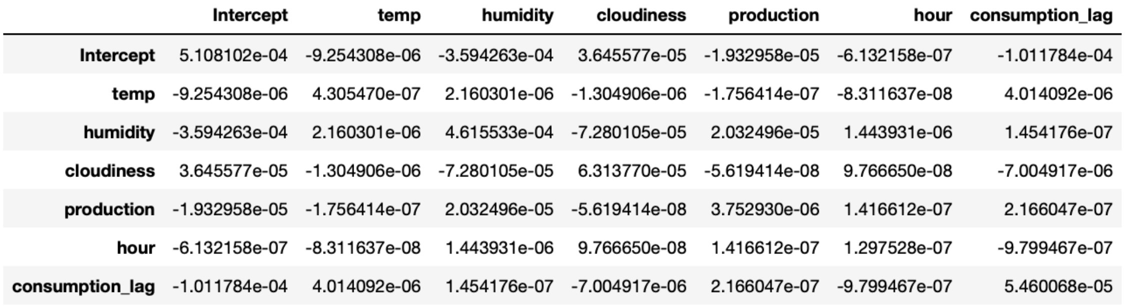
	Intercept	temp	humidity	cloudiness	production	hour	consumption_lag
Intercept	5.793742e-05	-6.968094e-07	-4.475506e-05	3.409665e-06	-3.132993e-06	-6.448797e-08	-3.290916e-05
temp	-6.968094e-07	3.022081e-08	2.785739e-07	-1.157811e-07	-3.924987e-08	-8.483946e-09	8.669395e-07
humidity	-4.475506e-05	2.785739e-07	5.565085e-05	-7.857292e-06	3.388041e-06	1.733675e-07	2.126418e-06
cloudiness	3.409665e-06	-1.157811e-07	-7.857292e-06	6.999219e-06	-1.374530e-08	1.403257e-08	-1.975267e-06
production	-3.132993e-06	-3.924987e-08	3.388041e-06	-1.374530e-08	9.079822e-07	2.991274e-08	-1.012421e-07
hour	-6.448797e-08	-8.483946e-09	1.733675e-07	1.403257e-08	2.991274e-08	1.858247e-08	-4.617065e-07
consumption_lag	-3.290916e-05	8.669395e-07	2.126418e-06	-1.975267e-06	-1.012421e-07	-4.617065e-07	5.899817e-05

Cluster 1



	Intercept	temp	humidity	cloudiness	production	hour	consumption_lag
Intercept	0.002266	-3.802620e-05	-0.001540	1.091671e-04	-8.475955e-05	-5.619021e-06	-0.000212
temp	-0.000038	1.868493e-06	0.000004	-4.197383e-06	-5.542075e-07	-3.043516e-07	0.000008
humidity	-0.001540	4.112643e-06	0.002189	-3.096934e-04	7.827363e-05	7.756521e-06	-0.000040
cloudiness	0.000109	-4.197383e-06	-0.000310	2.785805e-04	3.512170e-08	6.205397e-08	-0.000004
production	-0.000085	-5.542075e-07	0.000078	3.512170e-08	1.415661e-05	5.035622e-07	0.000003
hour	-0.000006	-3.043516e-07	0.000008	6.205397e-08	5.035622e-07	6.285744e-07	-0.000002
consumption_lag	-0.000212	8.380639e-06	-0.000040	-4.353019e-06	3.113618e-06	-1.741106e-06	0.000060

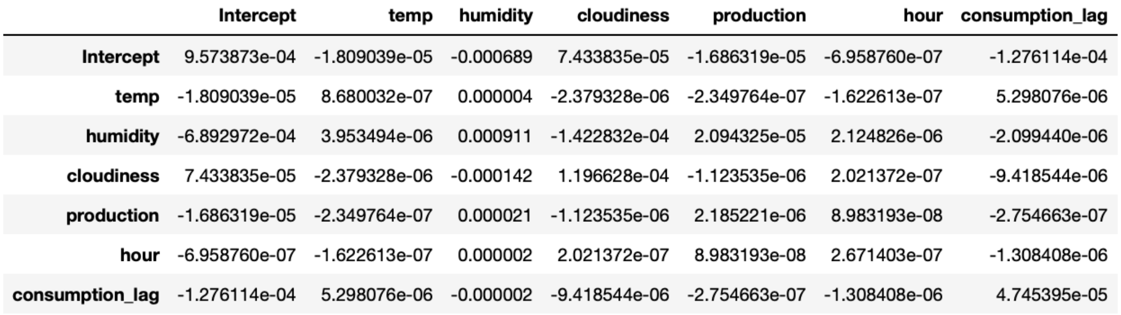
Cluster 2



	Intercept	temp	humidity	cloudiness	production	hour	consumption_lag
Intercept	5.108102e-04	-9.254308e-06	-3.594263e-04	3.645577e-05	-1.932958e-05	-6.132158e-07	-1.011784e-04
temp	-9.254308e-06	4.305470e-07	2.160301e-06	-1.304906e-06	-1.756414e-07	-8.311637e-08	4.014092e-06
humidity	-3.594263e-04	2.160301e-06	4.615533e-04	-7.280105e-05	2.032496e-05	1.443931e-06	1.454176e-07
cloudiness	3.645577e-05	-1.304906e-06	-7.280105e-05	6.313770e-05	-5.619414e-08	9.766650e-08	-7.004917e-06
production	-1.932958e-05	-1.756414e-07	2.032496e-05	-5.619414e-08	3.752930e-06	1.416612e-07	2.166047e-07
hour	-6.132158e-07	-8.311637e-08	1.443931e-06	9.766650e-08	1.416612e-07	1.297528e-07	-9.799467e-07
consumption_lag	-1.011784e-04	4.014092e-06	1.454176e-07	-7.004917e-06	2.166047e-07	-9.799467e-07	5.460068e-05

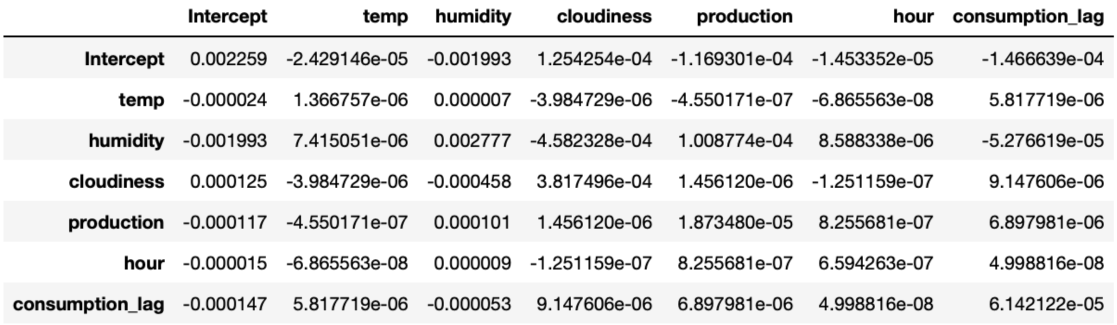
Cluster 3

Figure 56: Covariance matrices from the results of the modified GLM with identity function as link function for the clusters 1-3.



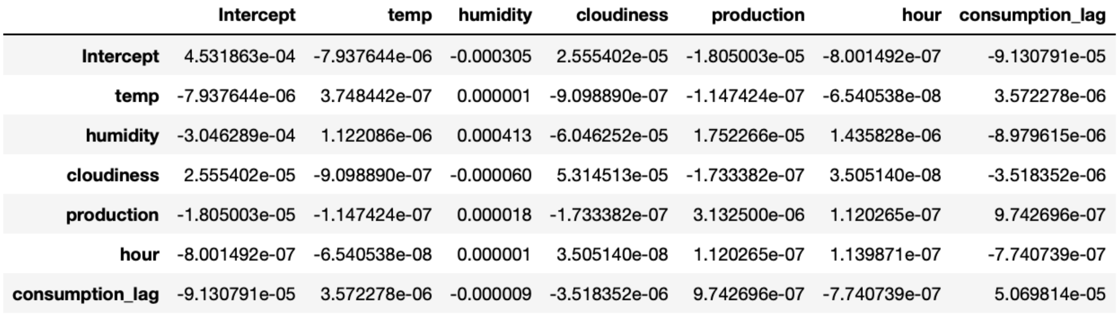
	Intercept	temp	humidity	cloudiness	production	hour	consumption_lag
Intercept	9.573873e-04	-1.809039e-05	-0.000689	7.433835e-05	-1.686319e-05	-6.958760e-07	-1.276114e-04
temp	-1.809039e-05	8.680032e-07	0.000004	-2.379328e-06	-2.349764e-07	-1.622613e-07	5.298076e-06
humidity	-6.892972e-04	3.953494e-06	0.000911	-1.422832e-04	2.094325e-05	2.124826e-06	-2.099440e-06
cloudiness	7.433835e-05	-2.379328e-06	-0.000142	1.196628e-04	-1.123535e-06	2.021372e-07	-9.418544e-06
production	-1.686319e-05	-2.349764e-07	0.000021	-1.123535e-06	2.185221e-06	8.983193e-08	-2.754663e-07
hour	-6.958760e-07	-1.622613e-07	0.000002	2.021372e-07	8.983193e-08	2.671403e-07	-1.308408e-06
consumption_lag	-1.276114e-04	5.298076e-06	-0.000002	-9.418544e-06	-2.754663e-07	-1.308408e-06	4.745395e-05

Cluster 4



	Intercept	temp	humidity	cloudiness	production	hour	consumption_lag
Intercept	0.002259	-2.429146e-05	-0.001993	1.254254e-04	-1.169301e-04	-1.453352e-05	-1.466639e-04
temp	-0.000024	1.366757e-06	0.000007	-3.984729e-06	-4.550171e-07	-6.865563e-08	5.817719e-06
humidity	-0.001993	7.415051e-06	0.002777	-4.582328e-04	1.008774e-04	8.588338e-06	-5.276619e-05
cloudiness	0.000125	-3.984729e-06	-0.000458	3.817496e-04	1.456120e-06	-1.251159e-07	9.147606e-06
production	-0.000117	-4.550171e-07	0.000101	1.456120e-06	1.873480e-05	8.255681e-07	6.897981e-06
hour	-0.000015	-6.865563e-08	0.000009	-1.251159e-07	8.255681e-07	6.594263e-07	4.998816e-08
consumption_lag	-0.000147	5.817719e-06	-0.000053	9.147606e-06	6.897981e-06	4.998816e-08	6.142122e-05

Cluster 5



	Intercept	temp	humidity	cloudiness	production	hour	consumption_lag
Intercept	4.531863e-04	-7.937644e-06	-0.000305	2.555402e-05	-1.805003e-05	-8.001492e-07	-9.130791e-05
temp	-7.937644e-06	3.748442e-07	0.000001	-9.098890e-07	-1.147424e-07	-6.540538e-08	3.572278e-06
humidity	-3.046289e-04	1.122086e-06	0.000413	-6.046252e-05	1.752266e-05	1.435828e-06	-8.979615e-06
cloudiness	2.555402e-05	-9.098890e-07	-0.000060	5.314513e-05	-1.733382e-07	3.505140e-08	-3.518352e-06
production	-1.805003e-05	-1.147424e-07	0.000018	-1.733382e-07	3.132500e-06	1.120265e-07	9.742696e-07
hour	-8.001492e-07	-6.540538e-08	0.000001	3.505140e-08	1.120265e-07	1.139871e-07	-7.740739e-07
consumption_lag	-9.130791e-05	3.572278e-06	-0.000009	-3.518352e-06	9.742696e-07	-7.740739e-07	5.069814e-05

All the data as one cluster

Figure 57: Covariance matrices from the results of the modified GLM with identity function as link function for cluster 4 and the entire data as one cluster.

D Python Code Before Classification

```
#Homes considered as new customers.
HOMEIDS =
↪  random.sample(consumption_and_weather["new_homeid"].unique().tolist(),
↪  k = 250)

#Make train/test set
def fetch_train_test_for_clusters(homeid_and_cluster: pd.DataFrame,
↪  cluster_id: int, consumption_and_weather: pd.DataFrame)->
↪  tuple[pd.DataFrame, pd.DataFrame, pd.DataFrame, list]:
    cluster_homeids =
    ↪  homeid_and_cluster[homeid_and_cluster["cluster"] ==
    ↪  cluster_id].new_homeid
    cluster_consumption_and_weather =
    ↪  pd.merge(consumption_and_weather,cluster_homeids, on
    ↪  ="new_homeid")
    agg_cons_and_weather =
    ↪  cluster_consumption_and_weather.groupby("time").mean()
    agg_cons_and_weather = agg_cons_and_weather.drop(["index",
    ↪  "weatherseriesid","tempk"],axis=1)
    agg_cons_and_weather['production'] =
    ↪  agg_cons_and_weather['production'].fillna(0)
    train_df = agg_cons_and_weather[agg_cons_and_weather.index <
    ↪  pd.to_datetime('2022-07-10', format='%Y-%m-%d')]
    test_df = agg_cons_and_weather[agg_cons_and_weather.index >
    ↪  pd.to_datetime('2022-07-10', format='%Y-%m-%d')]
    train_df = train_df.copy()
    train_df["hour"] = train_df.index.hour
    test_df = test_df.copy()
    test_df["hour"] = test_df.index.hour
    return train_df, test_df, cluster_consumption_and_weather,
    ↪  cluster_homeids

#The prediction model
def fit_and_predict_with_glm(train: pd.DataFrame, test: pd.DataFrame)
↪  -> pd.Series:
    train["hour"] = train.index.hour
    test["hour"] = test.index.hour
    formula = 'consumption ~
    ↪  temp+humidity+cloudiness+production+hour'
    glm_model = smf.glm(formula = formula, data = train, family =
    ↪  sm.families.Gaussian())
    fitted_glm = glm_model.fit()
    glm_prediction = fitted_glm.predict(test)
    return glm_prediction, fitted_glm
```

```
#The extended prediction model
def fit_and_predict_glm_with_lagged_consumption(train: pd.DataFrame,
↪   test: pd.DataFrame, lags: int = 24) -> pd.Series:
    train["hour"] = train.index.hour
    test["hour"] = test.index.hour

    train["consumption_lag"] = train['consumption'].shift(lags)
    test["consumption_lag"] = test['consumption'].shift(lags)

    train = train.dropna()
    test = test.dropna()

    formula_with_lag = 'consumption ~ temp + humidity + cloudiness +
    ↪   production + hour + consumption_lag'

    glm_model_with_lag = smf.glm(formula=formula_with_lag,
    ↪   data=train, family=sm.families.Gaussian())
    fitted_glm_with_lag = glm_model_with_lag.fit()

    glm_prediction_with_lag = fitted_glm_with_lag.predict(test)

    return glm_prediction_with_lag, fitted_glm_with_lag

#Hierarchical clustering
def make_clusters(consumption_and_weather: pd.DataFrame) ->
↪   pd.DataFrame:

    consumption_and_weather_vol2 =
    ↪   clean_data_dataframe(consumption_and_weather[consumption_and_weather.time
    ↪   < pd.to_datetime('2022-07-10', format='%Y-%m-%d')])
    mean_hourly_consumption_per_home =
    ↪   table_form(consumption_and_weather_vol2)

    plt.figure(figsize=(10, 7))
    plt.title("Customers Dendrogram")

    selected_data = mean_hourly_consumption_per_home[:1000].iloc[:,
    ↪   1:25]
    clusters = shc.linkage(selected_data,
                method='ward',
                metric="euclidean")
    shc.dendrogram(Z=clusters)
    plt.show()

    mean_hourly_consumption_per_home["cluster"] = fcluster(clusters,
    ↪   5, criterion="maxclust")
    homeids_and_cluster =
     ↪  mean_hourly_consumption_per_home[["new_homeid","cluster"]]
    return homeids_and_cluster

#Remove "new customers"
consumption_and_weather_without_HOMEIDS =
 ↪  consumption_and_weather[~consumption_and_weather['new_homeid'].isin(HOMEIDS)]

#All the data clustered. Approach 2
clusters = make_clusters(consumption_and_weather)

#Clustering the data without considered new households. Approach 3

cluster_without_HOMEIDS =
 ↪  make_clusters(consumption_and_weather_without_HOMEIDS)
```

E R Code Classification

```
#Make a dataframe of the desired variables
small_keep <- c("home_size","home_annual_consumption",
↪ "cluster","new_homeid")
smaller_keep <- c("home_size","home_annual_consumption","new_homeid")

smaller_frame <- new_homeids_frame[small_keep]
prop_frame <- properties_new_homes[smaller_keep]

#Test labels
find_old_cluster <- find_old_cluster[c("new_homeid", "cluster")]

#Standarization
smaller_frame$home_size <- scale(smaller_frame$home_size)
smaller_frame$home_annual_consumption <-
↪ scale(smaller_frame$home_annual_consumption)

prop_frame$home_size <- scale(prop_frame$home_size)
prop_frame$home_annual_consumption <-
↪ scale(prop_frame$home_annual_consumption)

#Train/Test set
train_frame <- subset(smaller_frame, select = -new_homeid)

train_set <- train_frame[, ! colnames(train_frame) %in% "cluster"]
train_labels <- train_frame$cluster
test_set <- subset(prop_frame, select = -new_homeid)
test_labels <- find_old_cluster$cluster

#Classification models

#K-nearest neighbour
knn_model <- knn(train = train_set, test = test_set, cl =
↪ train_labels, k = 18)

#Linear Discriminant Analysis
lda_model <- lda(train_labels ~ ., data = cbind(train_set,
↪ train_labels))
lda_predictions <- predict(lda_model, newdata = test_set)

#Random Forest
rf_model <- randomForest(train_labels ~ ., data = train_set, ntree=
↪ 40, mtry = sqrt(ncol(train_set)))
rf_predictions <- predict(rf_model, newdata = test_set)

#KDC
```

```
nb1 <- NaiveBayes(as.factor(train_labels) ~.,data=train_set,
 ↪  usekernel=T)
p1 <- predict(nb1, test_set)

#Quadratic Discriminant Analysis
qda_model <- qda(train_labels ~ ., data = cbind(train_set,
 ↪  train_labels))
qda_predictions <- predict(qda_model, newdata = test_set)

#Support Vector Maschine
svm_model <- svm(train_labels ~ ., data = train_set, kernel =
 ↪  "radial")
svm_predictions <- predict(svm_model, newdata = test_set)

#Accuracy Rate
accuracy_lda <- sum(lda_predictions$class == test_labels) /
 ↪  length(test_labels)

#Confusion Matrix
table(true = test_labels, predict = lda_predictions$class)

#Save predictions
new_frames = prop_frame
new_frames$new_cluster <- lda_predictions$class
```

F Python Code After Classification

```
#Change fit_and_predict_with_glm with
↪ fit_and_predict_glm_with_lagged_consumption to get the modified
↪ model

#Prediction of consumption
def calculate_metrics(cluster_data: pd.DataFrame, train:
↪ pd.DataFrame, test: pd.DataFrame, consumption_and_weather:
↪ pd.DataFrame) -> tuple[float, float, float, list, list]:
#Splitting the data depending on cluster
    one_train, one_test, one_data, one_home =
     ↪ fetch_train_test_for_clusters(cluster_data, 1,
     ↪ consumption_and_weather)
    two_train, two_test, two_data, two_home =
     ↪ fetch_train_test_for_clusters(cluster_data, 2,
     ↪ consumption_and_weather)
    three_train, three_test, three_data, three_home =
     ↪ fetch_train_test_for_clusters(cluster_data, 3,
     ↪ consumption_and_weather)
    four_train, four_test, four_data, four_home =
     ↪ fetch_train_test_for_clusters(cluster_data, 4,
     ↪ consumption_and_weather)
    five_train, five_test, five_data, five_home =
     ↪ fetch_train_test_for_clusters(cluster_data, 5,
     ↪ consumption_and_weather)

#Keep the prediction for each cluster
    glm_models_clusters = [fit_and_predict_with_glm(one_train,
     ↪ one_test)[0],fit_and_predict_with_glm(two_train,
     ↪ two_test)[0],fit_and_predict_with_glm(three_train,
     ↪ three_test)[0],fit_and_predict_with_glm(four_train,
     ↪ four_test)[0],fit_and_predict_with_glm(five_train,
     ↪ five_test)[0]]

#Keep the fitted model for each cluster
    glm_fits_clusters = [fit_and_predict_with_glm(one_train,
     ↪ one_test)[1],fit_and_predict_with_glm(two_train,
     ↪ two_test)[1],fit_and_predict_with_glm(three_train,
     ↪ three_test)[1],fit_and_predict_with_glm(four_train,
     ↪ four_test)[1],fit_and_predict_with_glm(five_train,
     ↪ five_test)[1]]

#Find the consumption in several clusters
    consumption_of_clusters =
     ↪ (sum(fit_and_predict_with_glm(one_train, one_test)) *
     ↪ one_home.count() +
                     sum(fit_and_predict_with_glm(two_train,
                      ↪  two_test)) * two_home.count() +
                     sum(fit_and_predict_with_glm(three_train,
                      ↪  three_test)) * three_home.count() +
                     sum(fit_and_predict_with_glm(four_train,
                      ↪  four_test)) * four_home.count() +
                     sum(fit_and_predict_with_glm(five_train,
                      ↪  five_test)) * five_home.count()
                     )

#Count homes
    number_of_homes = one_home.count() + two_home.count() +
     ↪  three_home.count() + four_home.count() + five_home.count()

#Find consumption of all data as one single cluster
    consumption_of_all_as_one_cluster =
     ↪  sum(fit_and_predict_with_glm(train,test))*number_of_homes

#Find actual consumption
    actual_consumption = (
        one_data[one_data.time > pd.to_datetime('2022-07-10',
         ↪  format='%Y-%m-%d')].consumption.sum() +
        two_data[two_data.time > pd.to_datetime('2022-07-10',
         ↪  format='%Y-%m-%d')].consumption.sum() +
        three_data[three_data.time > pd.to_datetime('2022-07-10',
         ↪  format='%Y-%m-%d')].consumption.sum() +
        four_data[four_data.time > pd.to_datetime('2022-07-10',
         ↪  format='%Y-%m-%d')].consumption.sum() +
        five_data[five_data.time > pd.to_datetime('2022-07-10',
         ↪  format='%Y-%m-%d')].consumption.sum()
    )

    return consumption_of_clusters,
     ↪  consumption_of_all_as_one_cluster, actual_consumption,
     ↪  glm_models_clusters, glm_fits_clusters

#Calculating difference
def calculate_mae(actual_consumption: float,
 ↪  consumption_of_all_as_one_cluster: float,
 ↪  consumption_of_clusters: float) -> tuple[float, float]:
    only_one_cluster_mae = mean_absolute_error([actual_consumption],
     ↪  [consumption_of_all_as_one_cluster])/829
    clusters_mae = mean_absolute_error([actual_consumption],
     ↪  [consumption_of_clusters])/829
    return only_one_cluster_mae, clusters_mae

#To find the predicted consumption with the old prediction models
 ↪  with the old and new homes included
```

```
clus_and_class_cons_clusters =
↪  (sum(clustered_few_homes_glm_models[0]) *
↪  clustered_and_classified_homes[clustered_and_classified_homes["cluster"]
↪  == 1].count()[0] +

sum(clustered_few_homes_glm_models[1]) *
↪  clustered_and_classified_homes[clustered_and_classified_homes["cluster"]
↪  == 2].count()[0] +

sum(clustered_few_homes_glm_models[2]) *
↪  clustered_and_classified_homes[clustered_and_classified_homes["cluster"]
↪  == 3].count()[0] +

sum(clustered_few_homes_glm_models[3]) *
↪  clustered_and_classified_homes[clustered_and_classified_homes["cluster"]
↪  == 4].count()[0] +

sum(clustered_few_homes_glm_models[4]) *
↪  clustered_and_classified_homes[clustered_and_classified_homes["cluster"]
↪  == 5].count()[0])

#Dataframes for plotting

single_cluster_vs_actual_consumption = pd.DataFrame({"time" :
↪  mean_consumption_all.index, "Predicted Value" :
↪  mean_consumption_all*829, "Actual Value":
↪  actual_consumption_frame.consumption})

several_clusters_vs_actual_consumption = pd.DataFrame({"time" :
↪  mean_consonsumption_cluster.index, "Predicted Value" :
↪  mean_consonsumption_cluster, "Actual Value" :
↪  actual_consumption_frame.consumption})


#Calculating MAE, MAPE and plotting

def hourly_plot(single_cluster_vs_actual_consumption,
several_clusters_vs_actual_consumption):
    plt.figure(figsize=(16, 6))

    mae1 = sum(single_cluster_vs_actual_consumption['Predicted
    ↪  Value']/829 - single_cluster_vs_actual_consumption['Actual
    ↪  Value']/829)

    mae2 = sum(several_clusters_vs_actual_consumption['Predicted
    ↪  Value']/829 - several_clusters_vs_actual_consumption['Actual
    ↪  Value']/829)
```

```
mape1 = (sum((single_cluster_vs_actual_consumption['Predicted
↪  Value'] - single_cluster_vs_actual_consumption['Actual
↪  Value'])/single_cluster_vs_actual_consumption['Actual
↪  Value'])) / 829

mape2 = (sum((several_clusters_vs_actual_consumption['Predicted
↪  Value'] - several_clusters_vs_actual_consumption['Actual
↪  Value'])/several_clusters_vs_actual_consumption['Actual
↪  Value'])) / 829

print(mae1, mae2, mape1, mape2)

plt.plot(single_cluster_vs_actual_consumption['time'],
↪  single_cluster_vs_actual_consumption['Predicted Value'],
↪  marker='o', label='Predicted Value All As One Cluster')

plt.plot(single_cluster_vs_actual_consumption['time'],
↪  single_cluster_vs_actual_consumption['Actual Value'],
↪  marker='o', label='Actual Value')

plt.plot(several_clusters_vs_actual_consumption['time'],
↪  several_clusters_vs_actual_consumption['Predicted Value'],
↪  marker='o', label='Predicted Value Several Clusters')

plt.xlabel('Time')
plt.ylabel('Values')
plt.title('Actual vs Predicted Values Over Time')
plt.legend()
plt.grid(True)
plt.show()
```

NTNU
Norwegian University of
Science and Technology

tibber