Assessment 3

Group Project Report

Submitted by

Yiu Tong, Chiu – z5039191

The University of New South Wales (UNSW)

Master of Data Science Course
(ZZSC9020 – Data Science Project)

Hexamester 2, 2024

Personal Statement

Formal Acknowledgment of Limited Participation on Assessment 3

I am writing to sincerely acknowledge the negative impact that my recent professional commitments have had on my ability to participate in our group projects for the course.

Since **April**, I have been deeply involved in an exhaustive project due diligence assignment that has taken me from Hong Kong to Vancouver. This business trip is anticipated to continue for the next year, and it has significantly limited my capacity to engage consistently and effectively with our team.

I profoundly regret any inconvenience and challenges that my reduced availability may have caused, particularly for my teammates, **Keith Chuang** and **Rupesh Gandhi**. They have had to shoulder this substantial project that was originally intended to be managed by four members. Therefore, I fully acknowledge and take responsibility for my lack of communication, and the increased workload that has consequently been imposed on them. I extend my heartfelt apologies to all stakeholders for the difficulties and extra workload that have emerged from this situation.

Based on previous email communication with Dr. Singh, I was suggested to complete my own version of Assessment 3 for evaluation separated from the project of Team Echo. Prior to my relocation, I was involved to parts of the Preliminary Project Planning and the Literature Review Section for Assessment 3, evidenced by Microsoft Teams records and Participation in 2 Consultation Meetings. These have established a foundational base for my own independent project work, and certain sections of this project report comes from the works and insights I developed before April. While Keith and Rupesh of course have since been tasked with the majority of further completing their version project.

I truly appreciate the understanding from the faculty and apologize again for the inconvenience caused regarding this issue.

Thank you.

Best regards,

Yiu Tong, CHIU (Davis)

Table of Contents

1 Abstract

Endgame Economics, my client, emphasizes the critical need for accurate electricity demand forecasting in New South Wales (NSW). Planning models play a vital role in policy formulation and grid stability. The dynamic landscape, influenced by weather patterns, economic activity, government policies, and demographic shifts, impacts energy consumption. Various forecasting methods, spanning statistical to machine learning approaches, have been explored to address this challenge.

Following factors influence the forecasting method selection:

- **Historical Data Available:** The quality and quantity of historical data impact the choice.
- **Forecasting Horizon:** Different methods suit short-term, medium-term, or long-term forecasts**.**
- **Type of Model:** Conventional or AI-based models.
- **Forecasting Accuracy:** The desired level of accuracy guides the selection.

Forecast models are grouped into two broad categories conventional models and AI-based models.

Conventional models are classified into:

- **Time Series Models**: Capture historical patterns to forecast the future.
- **Regression Models**: Model relationships between variables.

AI-based models are broadly classified into:

- **Deep Learning (DL) Models**: Employ feedforward neural networks for electricity demand forecasting.

Researchers have developed numerous hybrid approaches that combine various models to cater to specific needs. My comprehensive review focused on forecasting electricity consumption, particularly using the mean Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) as a key metric to assess accuracy. This review established a reference range for MAE and RMSE to evaluate the precision of forecasts and highlighted the effectiveness of time-series analysis as the superior method for predicting electricity consumption, outperforming other techniques.

2   Introduction

The task of forecasting electricity consumption in New South Wales (NSW) is geared towards tackling the issue of rising energy costs. The goal is to provide precise forecasts that allow consumers to make well-informed decisions about their energy use and expenses, thus promoting affordability and accessibility in the NSW energy market. This initiative uses historical data combined with various factors such as population growth, infrastructure development, economic trends, and governmental policies to create sophisticated forecasting models specific to NSW's unique energy requirements.

Significant research has been conducted on energy consumption forecasting, with models categorized into conventional, AI-based, and hybrid approaches. Key studies by (Daut, et al., 2017), (Bourdeau, Zhai, Nefzaoui, Guo, & Chatellier, 2019), (Connor, Chan, & Laforge, 2012)[1], and (Ahmad, Chen, Yabin, & Jiangyu, 2018) offer insights into the advantages, limitations, and applications of these models. These studies examine aspects such as the scope of predictions, data attributes, model applications, data pre-processing techniques, types of buildings, specific energy uses targeted, and evaluation of accuracy. This comprehensive analysis aids in developing effective models tailored to meet NSW's energy needs, contributing to better energy management and reduced costs.

Historically, energy consumption forecasts relied on conventional models like time series analysis, regression models, and gray models before the advent of AI. Recent research indicates that with careful parameter tuning, these traditional models can rival AI models in accuracy (Wei N. , Li, Duan, Liu, & Zeng, 2019). They typically establish clear correlations between energy use and factors like temperature and population (Debnath & Mourshed, 2018).

On the other hand, AI-based models such as artificial neural networks (ANN) and support vector regression (SVR) learn patterns from historical data and are particularly adept at addressing nonlinear issues and performing short-term forecasts. Despite their advantages, there is a noted scarcity of comprehensive review studies that analyze the various aspects of these forecasting models, including the types of models, multivariate considerations, forecasting horizons, and their accuracy across different time frames (Wei N. , Li, Duan, Liu, & Zeng, 2019).

3 Literature Review

3.1 Methodology

Stages		Outcomes
Preparation	Scope of Review	Understand and obtain insights from historically relevant methodologies for forecasting modelling of electricity demand in NSW
	Search Criteria	Keyword based search focus on energy consumption and prediction models across the globe
	Literature Identification	1. Science Direct 2. Google Scholar
Retrieval	Targeted Literature Search	Search by using. 1. "Electricity demand forecasting in NSW" 2. "Energy consumption trends NSW" 3. "Renewal energy impact on NSW grid"
	Gather Results	Papers read and reviewed for the literature review
Select	Inclusion Criteria	1. **Geographical Relevance:** Focus on or include data relevant to Australia or comparable regions with similar energy consumption patterns, economic structure, and climate etc. 2. **Forecasting Models:** Discuss, develop, or evaluate forecasting models for electricity demand, prefer an applicable regional scale like NSW. 3. **Rigorous Statistical Methodology:** Described clearly the forecasting methods, data used, and statistical techniques with validation evidence or testing of the models. 4. **Reliable Data Sources:** Data from credible sources such as the Australian Energy Market Operator (AEMO) or other recognized entities. All other papers were excluded from the review
	Quality Assessment	1. **Study Design:** Robust forecasting models that have been tested against actual data. 2. **Methodology:** Clearly defined and replicable methodology for data analysis and forecast modelling. 3. **Sample Size:** Adequate sample size to support the conclusions. 4. **Data Source:** Data from reliable sources such as government databases or reputable market operators such as AEMO. 5. **Data Analysis Technique:** Proper data preparation technique such as handling of missing data and outliers, data transformations, feature engineering, etc. 6. **Modelling Technique:** Appropriate and sophisticated statistical or machine learning techniques with justifiable assumptions. 7. **Discussion:** Acknowledgement of study's limitations and discussion of potential impact on the findings. 8. **Reproducibility:** Results that have been reproduced by other researchers is a strong indicator of quality and is highly preferred.
Synthesis	Findings	Synthesise relevant key insights from qualified literatures for the objective of electricity demand forecasting modelling for NSW.
	Insights	1. Suggest methodologies for data preparation, model variation. 2. Project assumptions for modelling 3. Selection of variables
	Forecasting model	Based on the literature review, data provided and pros-and-cons between conventional and AI based models for energy forecasting, I focussed on building a conventional model.
	Analysing Results	Analysed and coded multiple models including linear models, and post that ARIMA, SARIMA to cover seasonality.

Table 1: Methodology of Literature Reviews

3.2 Modelling Methods

3.2.1 Conventional Models

In energy consumption forecasting, conventional models like time series, regression models are standard. Researchers have sought to improve their accuracy by refining their structures or integrating them with advanced methods. This has led to the creation of various enhanced conventional models and hybrid combinations.

3.2.1.1 Time Series models

Time series (TS) models are forecasting methods that utilize historical demand data without requiring external variables, as highlighted by (A. Azadeh, 2015) . These models, such as the autoregressive (AR) and moving-average (MA) models, are commonly employed in energy consumption forecasting due to their simplicity and effectiveness. They require only a small amount of historical data for construction, many TS approaches represent regression models since the predicted value is estimated based on one or more previous values.

(Xu & Wang, 2010) focussed on improving the MA model with a second order polynomial curve, producing the polynomial curve and MA combination projection model (PCMACP). Results demonstrate the PCMACP model's superior reliability, with a Mean Absolute Percentage Error (MAPE) lower than previous methods such as back propagation neural network (BPNN) and GM.

This category includes univariate time series models such as autoregressive moving average (ARMA) models, Popular other techniques are autoregressive integrated moving average (ARIMA) models for non-stationary time series (by introducing a lag), seasonal autoregressive integrated moving average (SARIMA) models for seasonality and ARMA models for exogenous variables (ARMAX). A typical multivariate TS method is vector auto-regression and includes smoothing models and ARCH techniques.

3.2.1.2 Regression models (RMs)

(Thatcher, 2007) A linear regression model for predicting changes to electricity demand load duration curves (LDCs) in response to climate change in Australia. The model modifies a linear regression approach to incorporate intraday variability at 30-minute intervals, using regional electricity demand data and climate variables like temperature and humidity. Data from the Australian National Electricity Market (NEM) for four states—New South Wales (NSW), Victoria (VIC), Queensland (QLD), and South Australia (SA)—is utilized, alongside climate model datasets downscaled to 60 km resolution. Key variables include heating degree days (HDDs), cooling degree days (CDDs), and apparent temperature. The model's performance is validated against historical demand and climate data, with the adjusted R^2 metric indicating a close match between observed and predicted LDCs.

(H. Fan and I.F. MacGill and A.B. Sproul, 2015) utilized a bottom-up statistical linear regression model to forecast average daily electricity demand based on data from Australia's Smart Grid Smart City (SGSC) project. The model incorporated factors such as household demographics, behaviour, building and appliances, and climate impact, using a dataset that included half-hour interval electricity readings and survey data from 9903 households. Performance metrics for individual household predictions showed an adjusted R^2 value of 55%, indicating moderate accuracy, while aggregated zone-level predictions achieved a Mean Absolute Percentage Error (MAPE) of 3.9%, demonstrating good accuracy in forecasting average daily electricity demand.

(Radharani Panigrahi, 2022) The study presents a Machine Learning Categorical Boosting (ML CatBoost) Regressor model for predicting hourly aggregated electricity demand, utilizing data from the Electricity Reliability Council of Texas (ERCOT western). The model is praised for efficiently handling categorical features and reducing overfitting, showcasing superior performance over other machine learning models through various evaluation metrics. Key findings emphasize the model's accuracy in demand prediction, which is crucial for energy management and optimizing the contribution of clean energy sources, marking a significant advancement in the field of energy demand forecasting.

3.2.2 AI-based models

In the field of energy consumption forecasting, Artificial Neural Networks (ANN), Support Vector Regression (SVR), and Random Forest (RF) stand out as leading AI-based models. ANN and SVR are widely used and highly favored for addressing various forecasting challenges in this area.

3.2.2.1 DL Models

Recent advancements in deep learning (DL) models have led to their ability to produce forecasting predictions comparable to or superior to human experts, especially in domains like stock market index forecasting. Recurrent Neural Network (RNN) stands out as a key DL model for energy consumption forecasting, utilizing its internal state to process input sequences and improve prediction accuracy. (He W. , 2017) combined RNN with Convolutional Neural Network (CNN) to successfully forecast electricity load in a northern Chinese city, with CNN extracting crucial historical data features and RNN capturing dynamic load patterns.

Long Short-Term Memory (LSTM), a specialized form of RNN, (Graves, 2012), (Wei N. , Li, Duan, Liu, & Zeng, 2019) has been shown to outperform traditional models, addressing issues such as vanishing or exploding gradients in the back-propagation process. Additionally, integrating sequence-to-sequence architecture into LSTM models enables the estimation of loads for future time steps, enhancing prediction performance compared to other models such as Artificial Neural Networks (ANN), Support Vector Regression (SVR), and traditional RNNs.

4 Exploratory Data Analysis

This section initiates a comprehensive investigation into the consumption patterns of electricity within New South Wales (NSW). To achieve this, the analysis integrates a wide array of data sources, encompassing current electricity demand metrics, temperature fluctuations, predictive models of future electricity requirements, and additional pertinent external factors. The purpose of this detailed synthesis is to amalgamate these varied datasets to form a coherent understanding of the multifaceted influences that govern electricity demand in the region. By examining these elements in conjunction, I aim to delineate the interdependencies and isolate key drivers that significantly impact electricity consumption trends in NSW.

4.1 Project Objectives

Items	Description
Trends and seasonality	Identify the times and factors that cause fluctuations in electricity demand in NSW.
Air temperature and holidays	Explore the relationship between air temperature, holidays and electricity demand.
Modelling	Compare predictions with actual demand data to build a forecasting model for electricity demand in NSW.

Table 2: Project Objectives Description

4.2 Data Sources

Source	Data Type	Description
Project Share	Electricity Demand Data	Consists of detailed records of electricity demand in NSW, recorded in half-hourly intervals.
Project Share	Forecast Demand Data	Features predictions of electricity demand in NSW, provided in half-hourly increments.
Project Share	Air Temperature Data	Contains measurements of air temperature in NSW, noted for having inconsistent time intervals.
External	Holiday	Information on public holidays in Australia

Table 3: Data Source Description

4.3 Data Preparation

4.3.1 Data Cleaning

Data Type	Handling Technique
Null Values	Remove 579 row with missing values from Air Temperature data frame
Irrelevant Features	Remove irrelevant 'LOCATION' and 'REGIONID' from Air Temperature and Demand data frame.
Duplicated Data	Remove duplicated data by retaining only one record.

Table 4: Data Cleaning Techniques

4.3.2 Data Transformation

Type	Handling Technique
Datetime Object	Define a function to convert DATETIME feature into Datetime object type
Add Datetime Attributes	Define a function to add data frame columns year value, month value, day value, day of week and day of month from decomposition of Datetime feature.
Select Relevant Feature	Select relevant features DATETIME and TEMPERATURE only. Remove minute value other than 0 and 30 so that the timestamp matched the half hourly interval of Electricity Demand data frame.
Data Scaling	Since optimal 10 hours of day delays are determined, the Electricity Demand data frame is adjusted according.
Tsibble Object	Data frames are converted to tsibble object for Time Series Analysis and Forecasting.

Table 5: Data Transformation Technique

4.4 Descriptive Statistics

Summary Statistics of Electricity Demand

```
    DATETIME            TOTALDEMAND         REGIONID
Length:196513       Min.   : 5075      Length:196513
Class :character    1st Qu.: 7150      Class :character
Mode  :character    Median : 8053      Mode  :character
                    Mean   : 8113
                    3rd Qu.: 8959
                    Max.   :14580
```

The 'dfDemand' dataset provides information on the electricity demand in New South Wales (NSW) and contains 196,513 entries, detailed across three variables: 'DATETIME', 'TOTALDEMAND', and 'REGIONID'. Both 'DATETIME' and 'REGIONID' are character-type fields, likely storing textual data such as timestamps and regional identifiers, respectively.

The 'TOTALDEMAND' variable, which represents the electricity demand measured in some unit (likely megawatts, though the unit isn't specified), shows a range of values from a minimum of 5,075 to a maximum of 14,580. The distribution of 'TOTALDEMAND' is summarized with key quartile values: the first quartile is at 7,150, indicating that 25% of the demand values are below this point. The median demand is 8,053, which suggests that half of the demand values are less than 8,053, making it a central point of the data distribution. The mean demand, slightly higher at 8,113, is close to the median, suggesting a relatively symmetric distribution of demand values around the central tendency. The third quartile is at 8,959, indicating that 75% of the electricity demand values fall below this level.

Summary Statistics of Air Temperature Data

```
      LOCATION              DATETIME              TEMPERATURE
  Length:220326         Length:220326         Min.   :-1.30
  Class :character      Class :character      1st Qu.:13.40
  Mode  :character      Mode  :character      Median :17.70
                                              Mean   :17.42
                                              3rd Qu.:21.30
                                              Max.   :44.70
```

The 'dfTemp' dataset, which captures temperature readings in New South Wales (NSW), consists of 220,326 observations across three distinct variables: 'LOCATION', 'DATETIME', and 'TEMPERATURE'. Both the 'LOCATION' and 'DATETIME' variables are stored as character strings, indicating that they might contain textual data such as names of places and date-time stamps, respectively.

Focusing on the 'TEMPERATURE' variable, the summary statistics provide a comprehensive view of the temperature distribution. The temperatures range from a minimum of -1.3°C, indicating occasional below-freezing conditions, to a maximum of 44.7°C, suggesting extremely hot weather conditions at certain times. The first quartile (25th percentile) of the temperature is 13.4°C, which means that 25% of the temperature readings are below this value. The median temperature is 17.7°C, highlighting that half of the readings are below this temperature and the other half above, suggesting a moderate climate for much of the time.

The mean temperature is slightly lower than the median at 17.42°C, which could indicate a slight skew in the data towards cooler temperatures. The third quartile (75th percentile) is 21.3°C, indicating that 75% of the temperatures are below this value. The distribution of temperatures, with a wide range from frigid to extremely hot, reflects the varied climatic conditions experienced across NSW, possibly influenced by seasonal variations and differing geographical locations within the state.
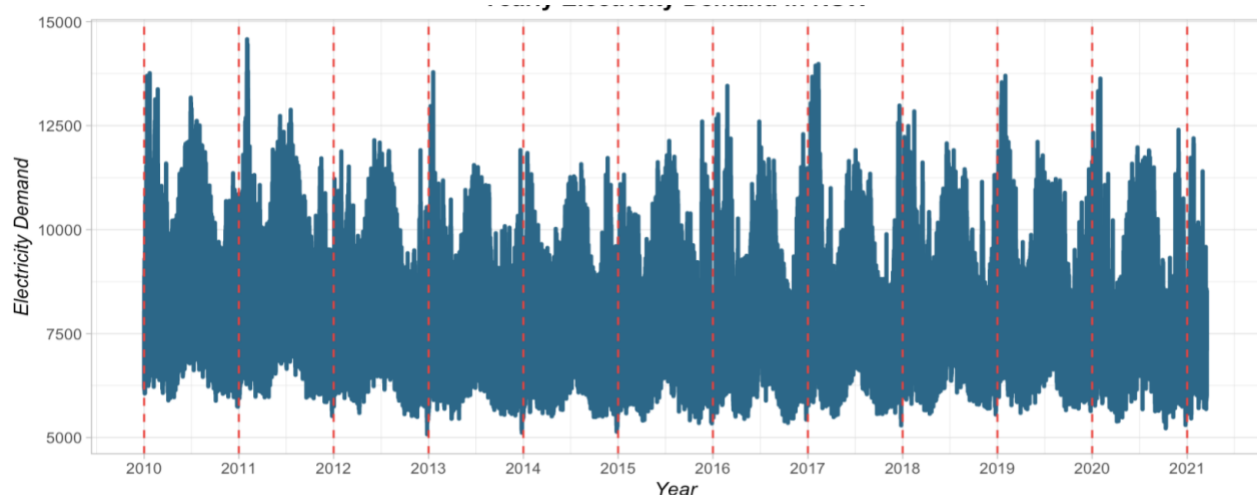
4.5 Univariate Analysis

Figure 1: Yearly Electricity Demand in NSW from 2010 to 2021

The time series plot shows the change of electricity demand in New South Wales over 11years. A clear seasonal pattern is observed, with peaks during winter and summer months, reflecting higher heating and cooling needs.
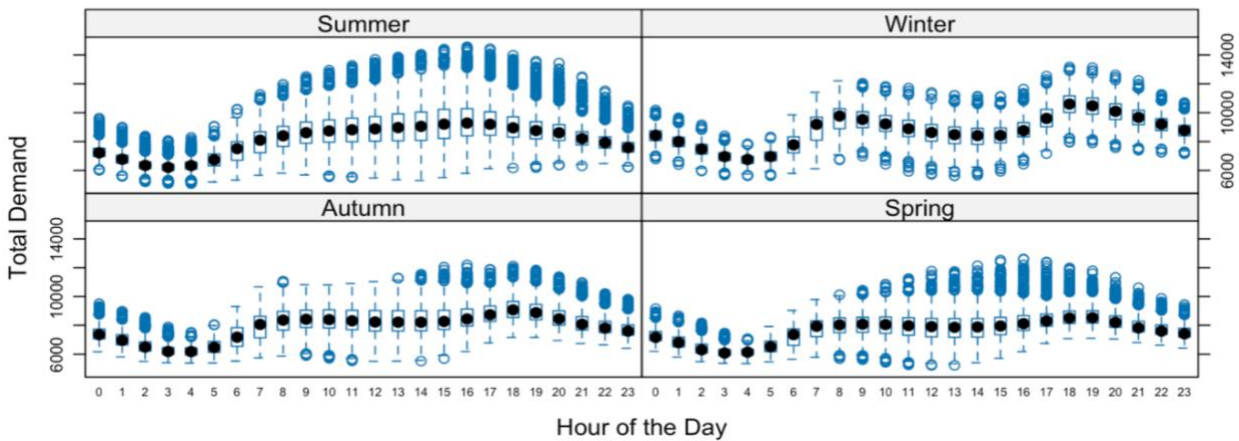

Figure 2: Hourly Electricity Demand by Season

Figure 2 shows the examination of the daily electricity demand patterns across different seasons. The box plots for each season reveal the following patterns:

Season	Description
Spring	Increases slightly from 1pm and stabilizes until midnight.
Summer	Peak demand from midday to afternoon, correlating with increased use of cooling systems.
Autumn	Relatively consistent with slight increases in the evening, reflecting typical residential consumption.
Winter	Distinct peaks in the early morning and evening, due to heating needs during the colder parts of the day.

Table 6: Interpretation of Seasonal Impact on Hourly Electricity Demand

The differentiation in daily patterns by season underscores the influence of weather-related factors on electricity consumption.

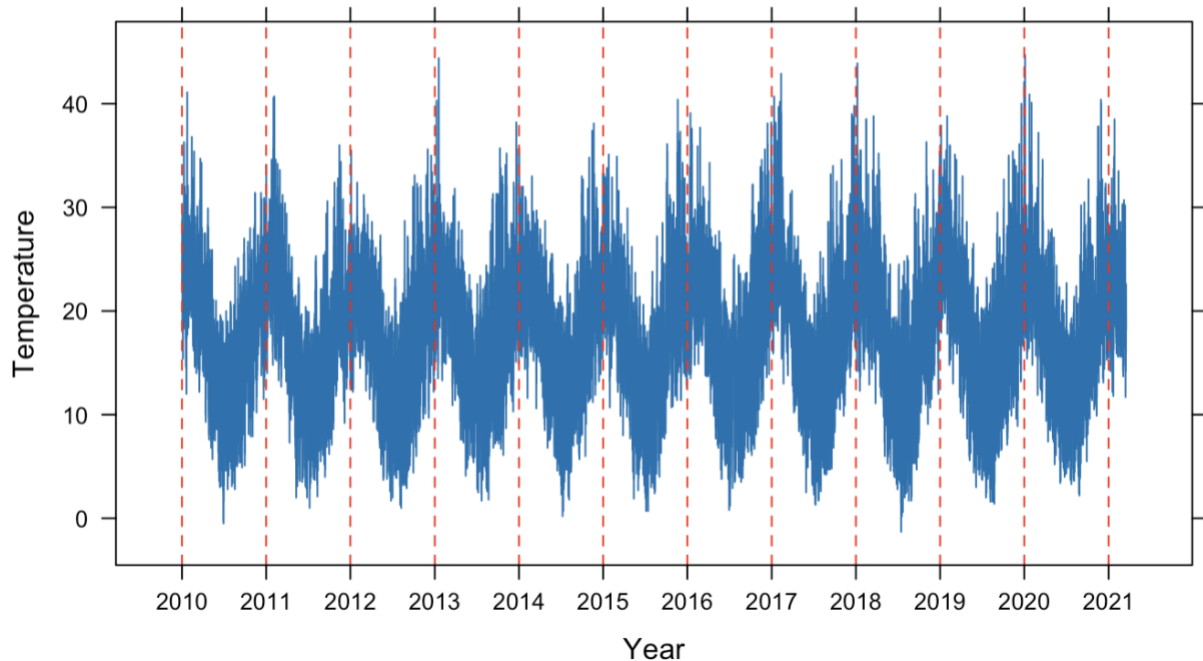**Yearly Temperature Vairation in NSW**

Figure 3: Yearly Temperature Variation in NSW

The line graph in Figure 3 showcasing temperature variations in New South Wales from 2010 to 2021 vividly illustrates a pronounced seasonal pattern, with temperature fluctuations that consistently repeat each year throughout the observed period. Each year within the dataset, temperatures in NSW ascend during the warmer months and descend during the cooler periods, showcasing a clear and predictable cycle of seasonal temperature changes. This regularity is marked by similar patterns emerging year after year, underscoring the stable and predictable nature of seasonal influences on the climate of NSW.
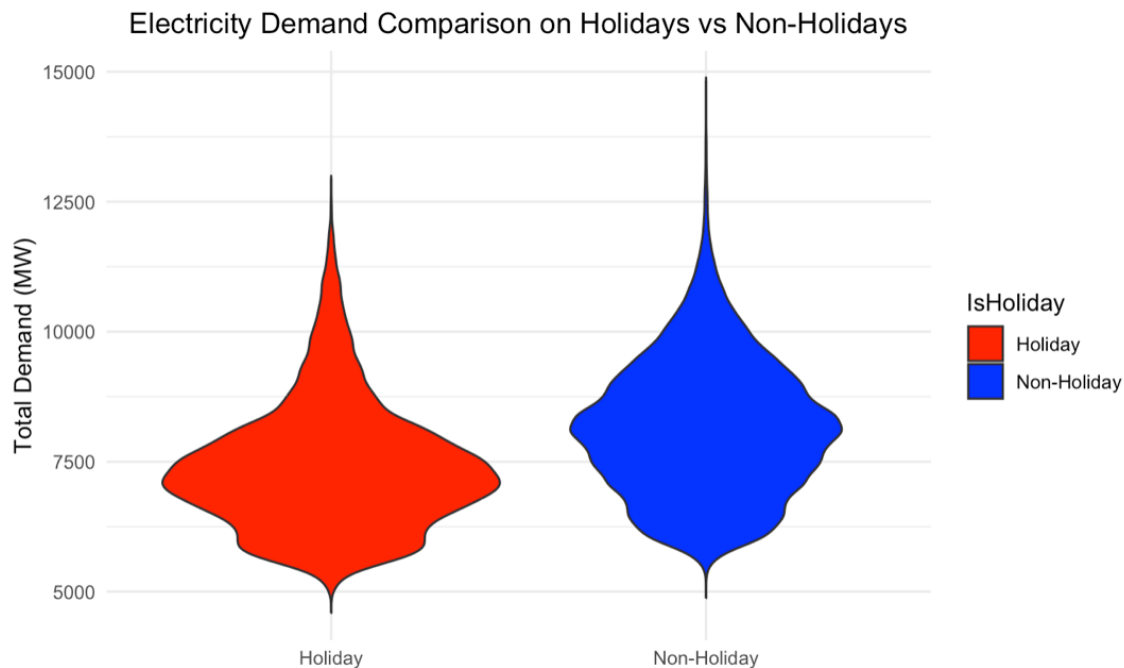
4.6 Bivariate Analysis

Figure 4: Electricity Demand on Holidays vs Non-Holidays

The violin plot of Figure 4 serves to delineate the variance in electricity demand between holiday and non-holiday periods. It reveals that the median electricity demand during holidays is generally lower than that during non-holidays. Additionally, the plot elucidates the comprehensive distribution of demand and identifies a few outliers that deviate from the typical range observed in both types of days. Notably, the maximum value of electricity demand recorded during non-holiday periods surpasses that observed during holidays. Moreover, the distribution of electricity demand on holidays is more tightly clustered around the median, specifically at around 7500 MW, indicating a more concentrated dispersion compared to non-holiday periods.

Electricity Demand vs Temperature

```
[1] 0.1490452

        Pearson's product-moment correlation

data:  dfMerged$TOTALDEMAND and dfMerged$TEMPERATURE
t = 66.721, df = 195945, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1447130 0.1533717
```

Pearson's Correlation (Electricity Demand vs Temperature)
Using Pearson's correlation to assess the relationship between electricity demand and temperature in NSW is well-suited due to both variables being continuous and the assumption that they share a linear relationship. The Pearson correlation analysis between electricity demand ('TOTALDEMAND') and temperature ('TEMPERATURE') using the dataset 'dfMerged' produced a correlation coefficient ('cor') of 0.149. This positive correlation indicates that as the temperature increases, there is a slight upward trend in electricity demand. The results of the Pearson's test were highly statistically

significant, as indicated by an extremely small p-value (less than 2.2e-16), strongly rejecting the null hypothesis that there is no correlation between these two variables.

The confidence interval for this correlation, ranging from 0.1447 to 0.1534, is narrow, suggesting a stable estimate of the correlation coefficient. This interval underscores that, while the correlation is relatively modest, it is consistently different from zero across different samples from the population.

The analysis confirms a statistically significant with positive relationship between temperature and electricity demand. This suggests that higher temperatures may lead to increased electricity usage, possibly due to greater use of cooling systems or other temperature-dependent activities.
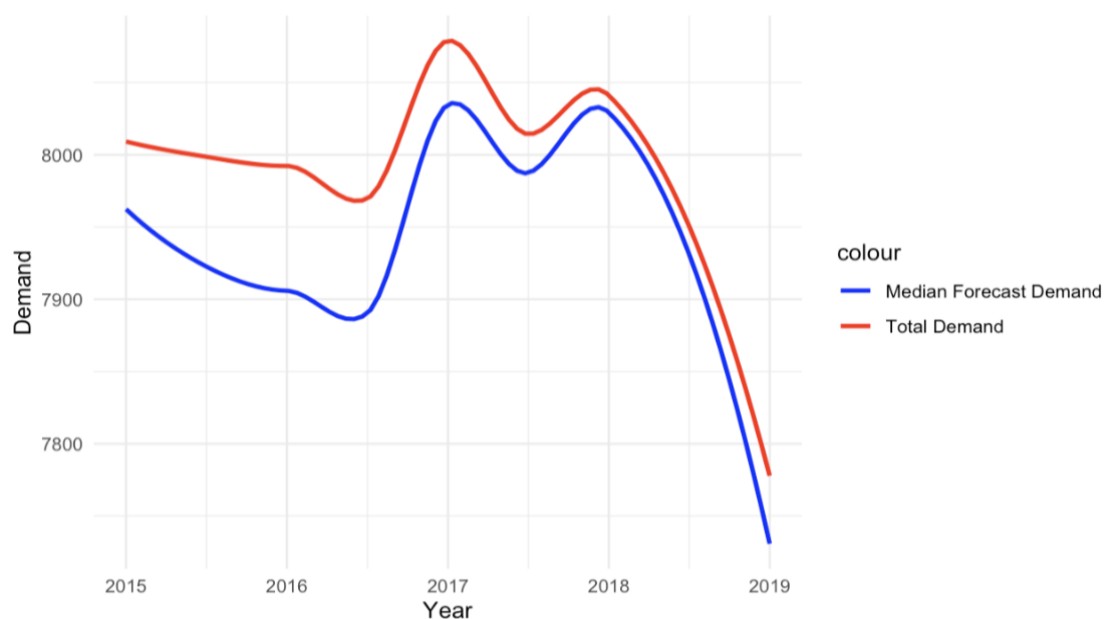

Figure 5: Forecast Electricity Demand vs Total Electricity Demand from 2015 to 2018

The 2015 to 2019 data of Median Forecast Demand and Actual Demand are selected, and smoothing technique are used for better visualization.

Figure 5 shows that the gap between the smoothed lines of Forecast Demand and Actual Demand narrow since 2017 and the discrepancy become stable starting from 2018 to 2019.

MAE & RMSE of Forecast Electricity Demand vs Actual Electricity Demand

```
"Mean Absolute Error (MAE): 1409.71203738599"
"Root Mean Squared Error (RMSE): 1771.09094554203"
```

MAE = 1409.71: This value suggests that on average, the predictions of the forecasting model deviate from the actual demand by about 1409.71 units.

RMSE = 1771.09: The value of 1771.09 indicates that while many of the predictions may be reasonably accurate, there are notable instances where the forecast deviates

significantly from the actual demand. This could be indicative of the model's sensitivity to data anomalies or its failure to capture extreme fluctuations due to external factors or rare events.

From **Literature Review**, various studies have explored the concept of introducing time delays into forecasting models to better align predictions with actual outcomes as follows:

Field	Findings
Weather Forecasting and Energy Demand	A study by Gajowniczek and Ząbkowski (2017) investigated the impact of temperature forecasts and their timing on energy demand predictions. The researchers found that adjusting forecast models to account for delays between temperature changes and peak energy demand significantly improved the accuracy of the forecasts.
Economic Forecasting	Economic Forecasting: In economics, the work by Clements and Hendry (1998) demonstrated that forecast models often need to incorporate information delays to effectively predict economic indicators. Their findings suggest that understanding and modelling the time lags between economic policy implementation and its impact on economic indicators can enhance forecasting accuracy.
Traffic Flow and Congestion Prediction	Traffic Flow and Congestion Prediction: Research by Vlahogianni, Karlaftis, and Golias (2014) on short-term traffic forecasting highlighted the necessity of including time delays to account for the non-linear and dynamic nature of traffic flow and congestion patterns. The study suggested that temporal adjustments could capture the delayed effects of traffic incidents on flow conditions.

Table 7: Literature Reviews of Time Delay Introduction

These studies collectively underline the importance of considering time delays in predictive models across various fields. Thus, it is reasonable to assume there could be time delay for the time-series forecasting modelling.

In optimizing electricity demand forecasts, the accuracy is assessed on the impact of adjusting forecast timing relative to actual demand, over a range from -12 to 12 hours (24 hours). For each tested delay, I calculated the root mean square error (RMSE) after aligning the forecasted data with the actual demand data.

Optimal Hours of Delay

	Delay <dbl>	RMSE <dbl>
3	−10	208.5899

The process revealed that a 10-hour delay between the forecasted and actual data minimized the RMSE to 208.5899, a significant improvement from the initial RMSE of 1771.09 seen without any delay. This finding indicates that precisely aligning forecast timing with actual electricity usage is critical for enhancing forecast reliability and accuracy.
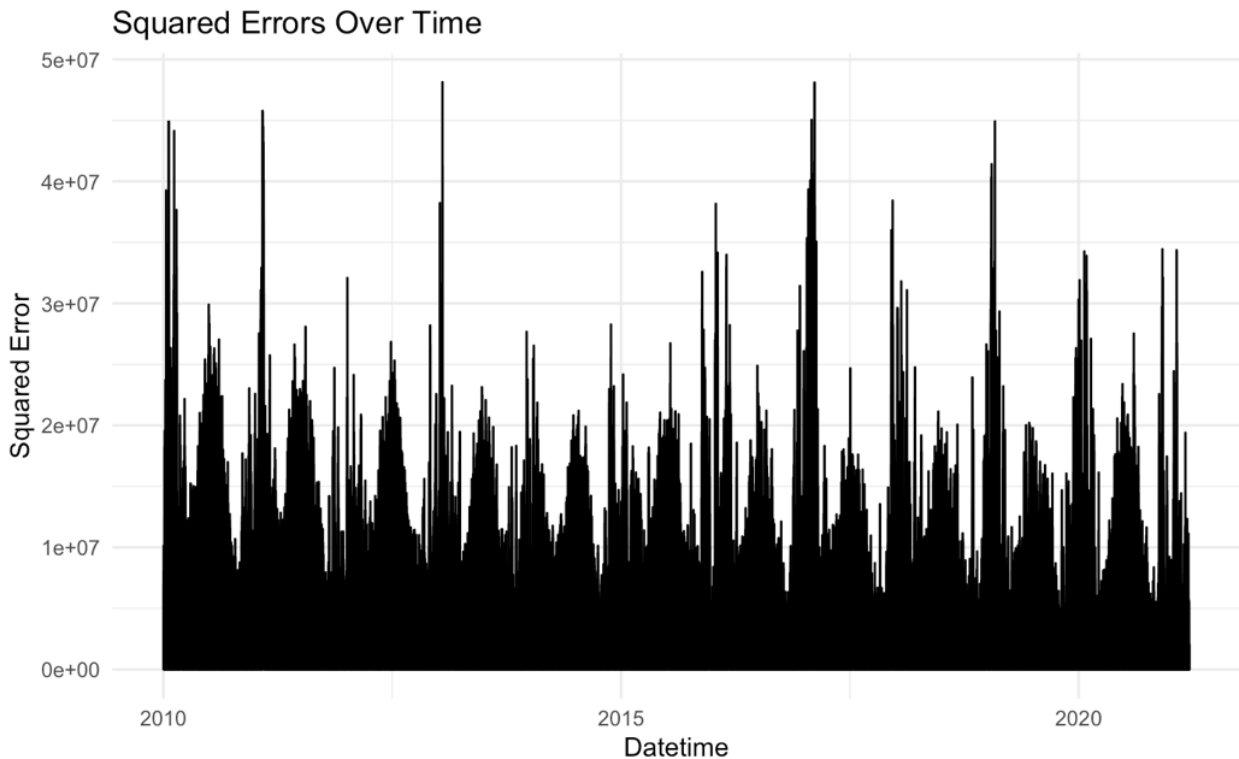

Figure 6: Squared Errors of Forecast Electricity Demand vs Actual Electricity Demand

In the analysis of the forecasted versus actual electricity demand over the course of a year, the squared errors depicted in Figure 6 reveal discernible patterns and seasonal variations. This figure, which plots these squared errors, serves as a quantitative measure of the discrepancies between predicted and actual demand values, providing insights into the accuracy of my forecasting model. The presence of strong patterns and seasonality in the data underscores the cyclical nature of electricity demand, which can be attributed to various factors such as climatic changes, industrial activity, and consumer behaviour throughout the year. The peaks in squared errors typically coincide with periods of high demand variability, suggesting that my model's performance varies seasonally. These findings indicate a need for model adjustments or the incorporation of more dynamic variables to enhance forecast accuracy during these critical periods. This analysis not only helps in understanding the model's limitations but also directs future efforts towards improving the precision of electricity demand forecasts.
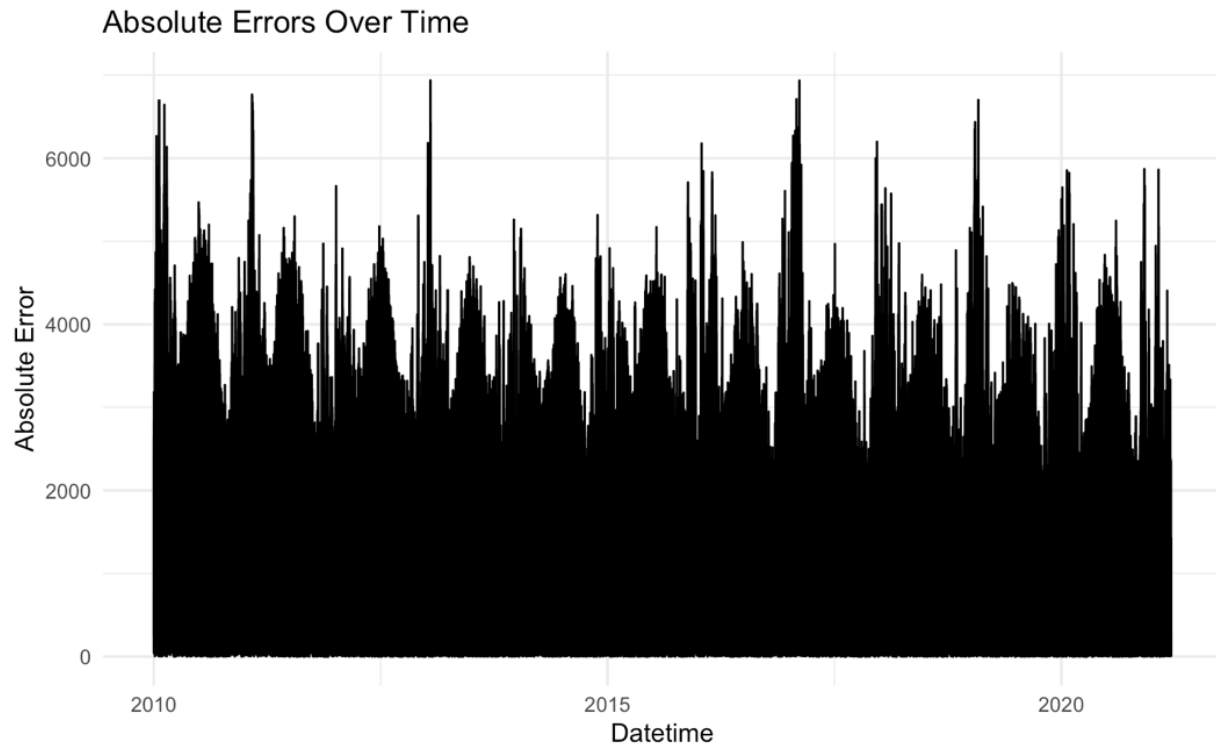
Absolute Errors Over Time

Figure 7: Absolute Errors of Forecast Electricity Demand vs Actual Electricity Demand over Year

The analysis of forecasted versus actual electricity demand over the year, presented in squared errors (Figure 6) and absolute errors (Figure 7), shows similar seasonal patterns and variations. Both figures illustrate the cyclical nature of electricity demand and reveal peaks in errors during periods of high demand variability. This consistency across both error types highlights seasonal fluctuations in the model's performance, indicating a need for model adjustments to enhance forecast accuracy.
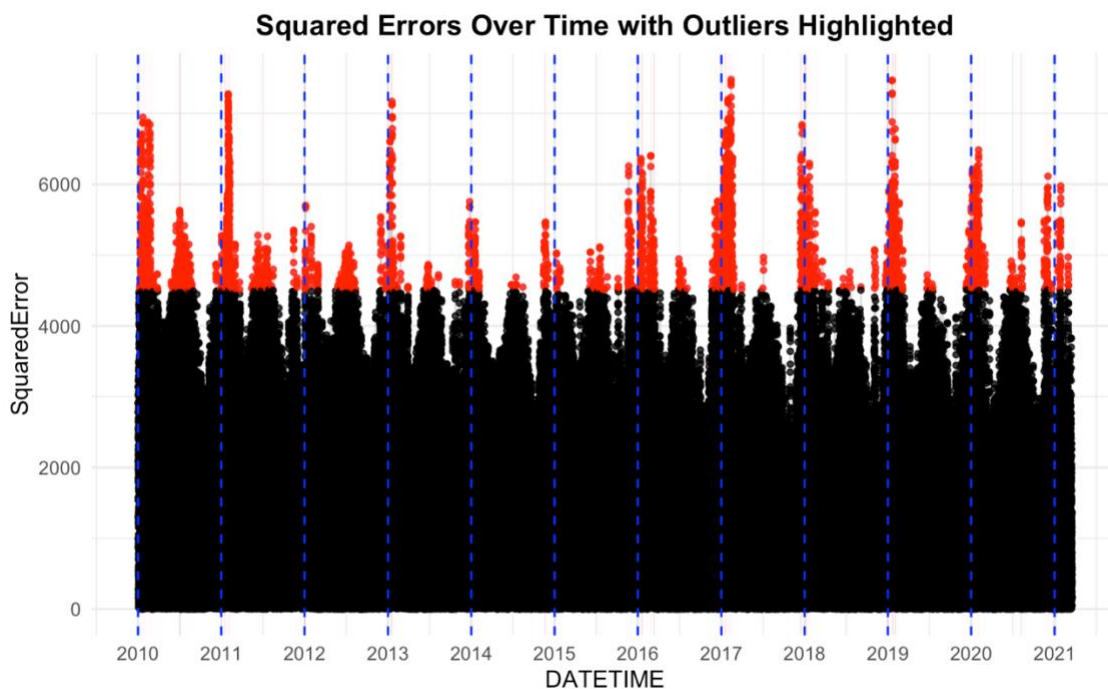

Squared Errors Over Time with Outliers Highlighted

Figure 8: Squared Errors Over Time

In the analysis of the dataset, it has been observed that outliers, defined as values exceeding a threshold of 4500, predominantly manifest around the transitions between years. This pattern indicates a potential seasonal or systematic influence related to annual cycles. The identification of such outliers was conducted through the application of robust statistical methods, where outliers were determined based on their deviation from the median forecast demand by more than 1.5 times the interquartile range (IQR), a common approach for outlier detection in time series data.

The concentration of extreme values at the beginning and end of each year could be attributed to several factors, such as operational shifts during holiday seasons, annual climatic variations, or end-of-year economic activities that typically influence the parameters being measured. This recurring pattern warrants further investigation to determine the underlying causes and to assess their impact on the overall dataset.

4.7 Correlation Heatmap

4.7.1 Electricity Demand vs Selected Variables
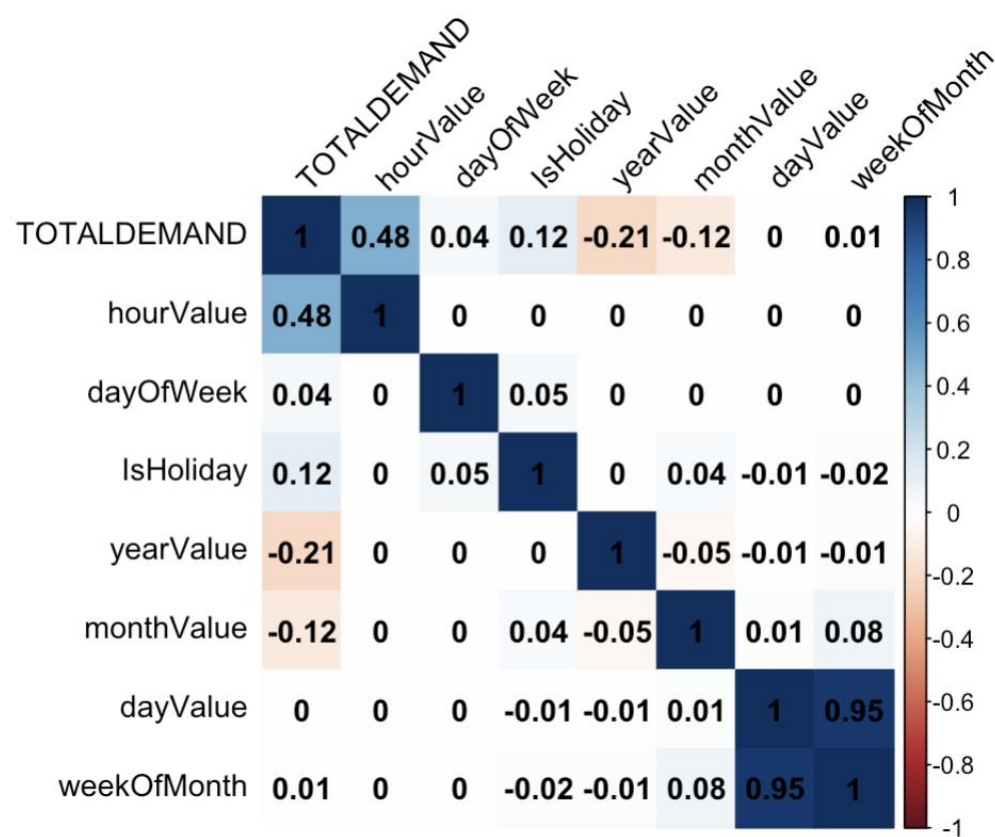

Figure 9: Correlation Heatmap of TOTALDEMAND

```
    hourValue     yearValue    monthValue     IsHoliday     dayOfWeek    weekOfMonth      dayValue
 0.476948947  -0.209345123  -0.122574678   0.103294225   0.043732681   0.005078356   0.004951010
```

The correlation analysis in Figure 9 conducted to assess the impact of Hour, Day, Month, Year, Day of Week, Week of Month and Holiday variables on TOTALDEMAND produced a series of insightful results. The variable hourValue exhibited the strongest positive correlation (0.476948947), highlighting that electricity demand significantly peaks during specific hours, likely related to routine daily activities. In contrast, yearValue showed a notable negative correlation (-0.209345123), indicating a gradual

reduction in electricity demand over the years, potentially due to advancements in energy-saving technologies or changes in consumer behavior.

Additionally, monthValue displayed a moderate negative correlation (-0.122574678), suggesting seasonal variations in demand, possibly driven by heating in winter and cooling in summer. The positive correlation with IsHoliday (0.103294225) suggests increased electricity usage during holidays, possibly due to more activities and gatherings. However, correlations with dayOfWeek (0.043732681), weekOfMonth (0.005078356), and dayValue (0.004951010) were relatively weak, indicating these factors have a minimal influence on demand fluctuations compared to the time of day or year.

In the development of the forecasting model, a criterion was established whereby only features exhibiting an absolute correlation of at least 0.1 with TOTALDEMAND were considered for inclusion. Consequently, the variables hourValue, yearValue, monthValue, and IsHoliday were selected based on this threshold.
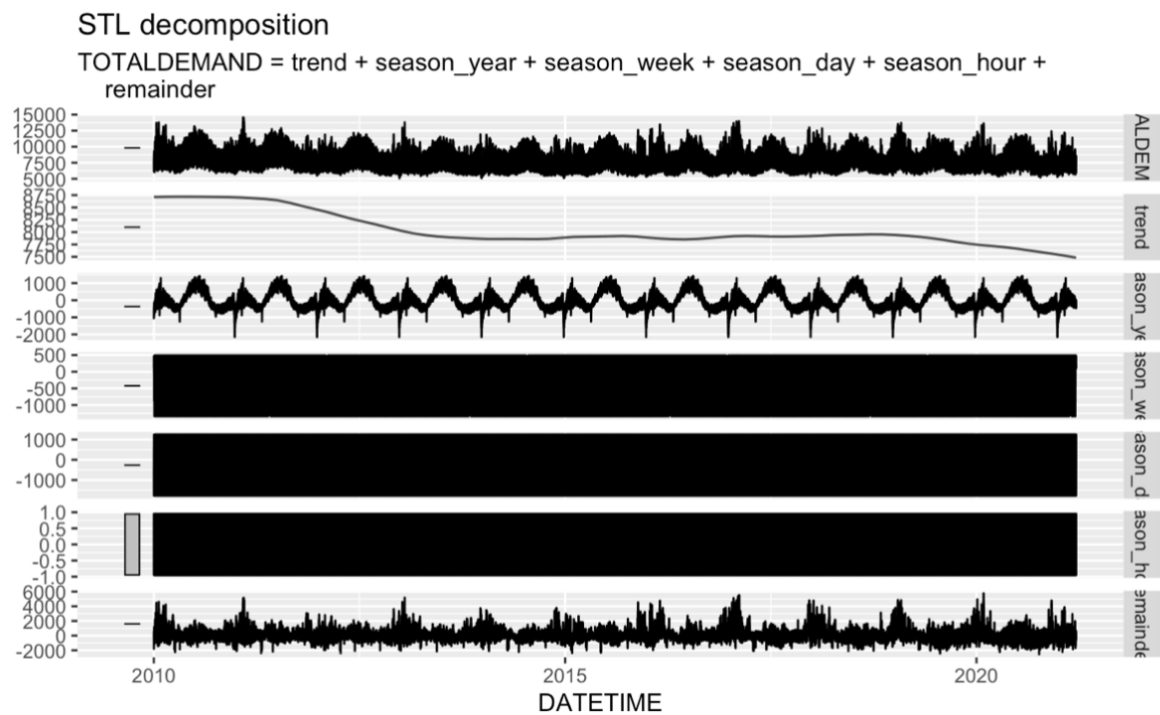
4.8 STL Decomposition

Figure 10: STL decomposition of Total Electricity Demand

The STL decomposition of the TOTALDEMAND time series has separated the data into several distinct components: trend, seasonal decompositions at different scales (yearly, weekly, daily, and hourly), and a remainder component.

Component	Description
Trend	Reflects long-term changes in TOTALDEMAND due to factors like population growth, economic development, changes in industry, or energy-efficient technologies. Useful for strategic planning and policy making.

Component	Description
Yearly Seasonality	Captures annual fluctuations in TOTALDEMAND, typically due to seasonal weather changes affecting electricity demand for heating or cooling.
Weekly Seasonality	Shows weekly patterns in TOTALDEMAND, differentiating between weekdays and weekends, influenced by business operations and consumer lifestyles.
Daily Seasonality	Illustrates intraday variations in demand, driven by daily human activities that increase electricity usage during morning and evening peaks.
Hourly Seasonality	Indicates hourly fluctuations within each day, possibly due to changes in human activity, commercial usage peaks, or time-of-use rates.
Remainder Component	Contains the unexplained variability in TOTALDEMAND, useful for identifying anomalies such as unexpected spikes or drops due to emergencies or unusual weather.

Table 8: Interpretation of STL Components
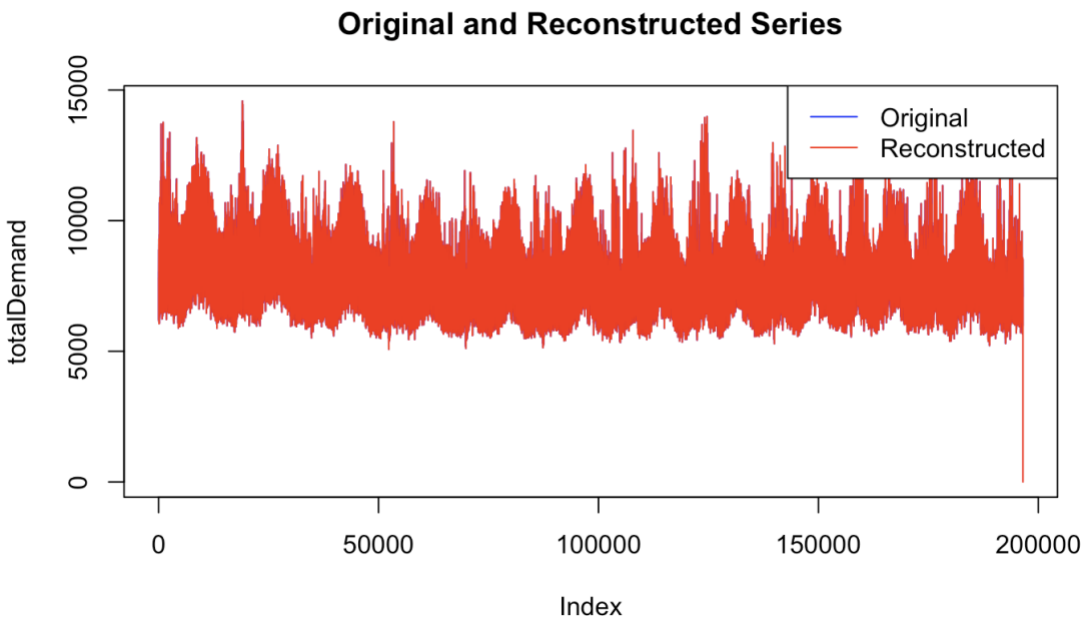
4.9 Wavelet Decomposition

Figure 11: Wavelet Decomposition of Total Electricity Demand

Figure 11 shows the wavelet decomposition of electricity demand data in NSW, using the "haar" wavelet, captures the primary dynamics of the time series, as evidenced by the significant overlap between the original and reconstructed series. However, the detection of an outlier at approximately index 190,000, where the reconstructed value abruptly drops to zero, highlights a potential anomaly or a limitation in handling specific data variations with the "haar" wavelet.

This finding underscores the potential shortcomings of the "haar" wavelet in modelling subtle changes in electricity demand, which is crucial for data preparation in forecasting

models such as ARIMA. The wavelet decomposition helps in achieving data stationarity by reducing noise and clarifying the main demand components, ensuring that forecasting models are trained on data that reflect true underlying patterns.

Furthermore, the extracted components—trend, seasonality, and residuals—offer valuable features for enhancing forecasting models, particularly in machine learning where multiple features can improve prediction accuracy. However, the analysis also recommends omitting the identified outlier to maintain the integrity and accuracy of the forecasting models.

4.10 Augmented Dickey-Fuller Test

```
data:   tsData
Dickey-Fuller = -2.1127, Lag order = 73, p-value = 0.5309
alternative hypothesis: stationary
```

Since the p-value is 0.5309, which is much greater than the typical significance level of 0.05. This means it fails to reject the null hypothesis that there is a unit root present in the series, implying the data likely contains some trend or seasonality that needs addressing.

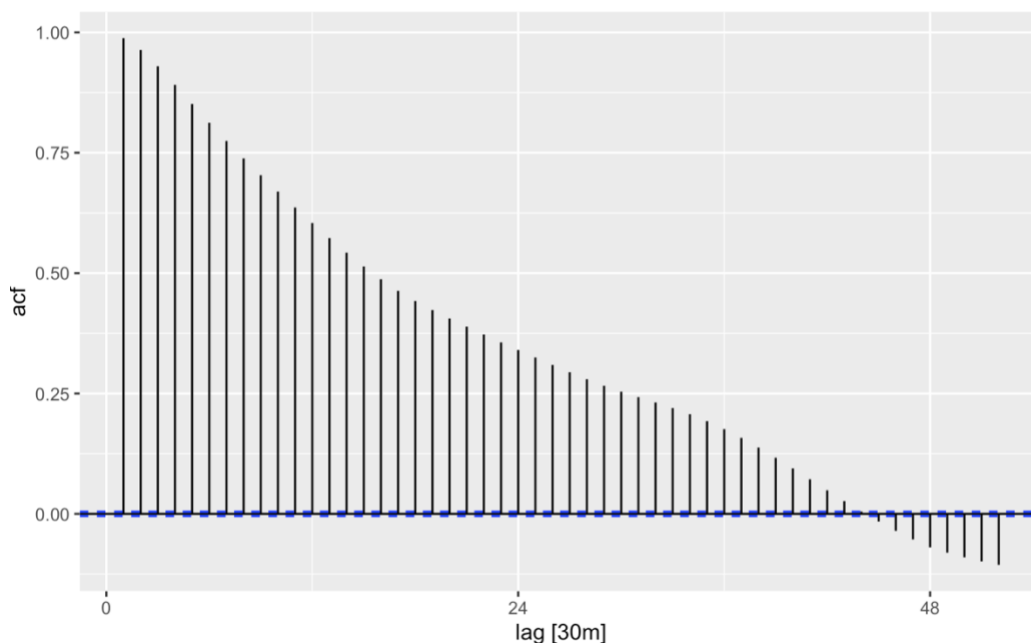## 4.11 First Differenced ACF Plot for Daily Seasonality

Figure 12: Daily Seasonally Differenced ACF plot of Demand

In the analysis of daily seasonality via the Autocorrelation Function (ACF) plot of differenced data, the observation of an ACF value of zero at lag 44 (equivalent to a 22-hour period) is particularly noteworthy. This result indicates no correlation between observations 22 hours apart, suggesting an absence of predictive continuity at this interval within the daily cycle. This finding could imply a periodic reset or shift in the underlying process, affecting the predictability and potentially informing adjustments in forecasting models. Understanding and exploring the reasons behind these zero

autocorrelations could provide valuable insights for refining time series analyses and developing targeted interventions.
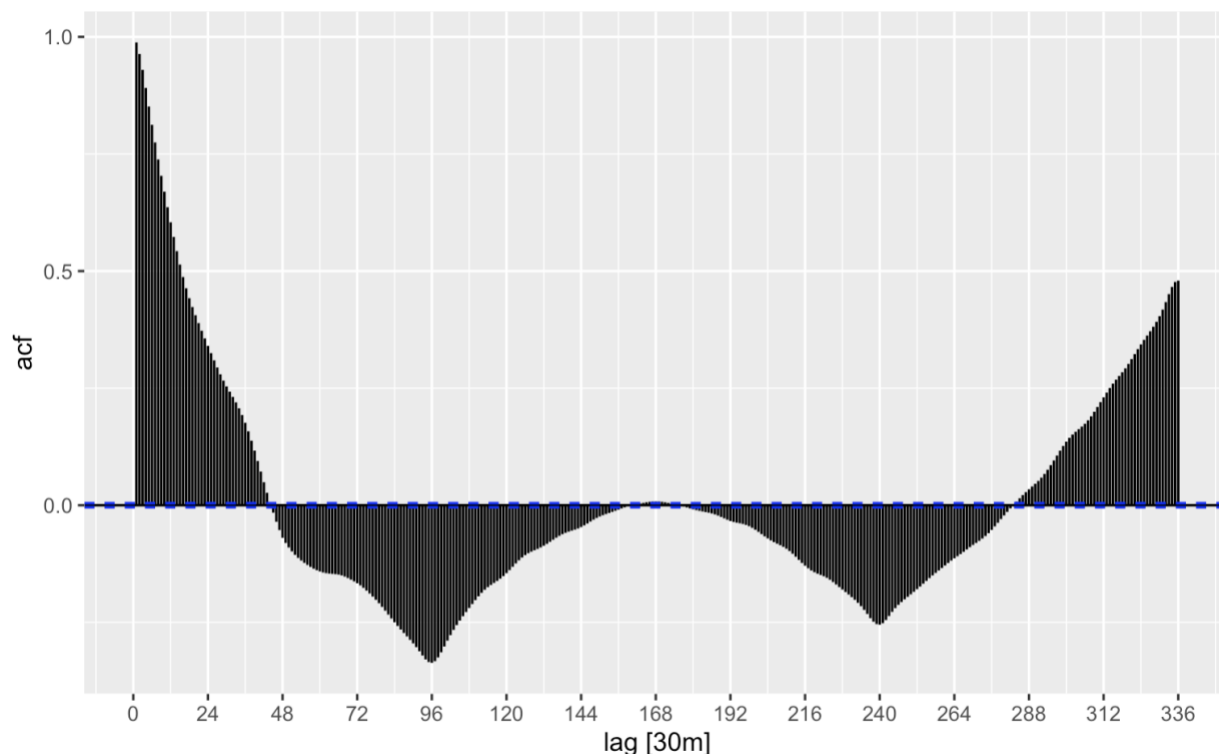

Figure 13: Daily and Weekly Seasonally Differenced ACF plot of Demand

In the analysis of the Autocorrelation Function (ACF) for seasonally differenced demand data, the ACF values on Figure 13 at lags of around 44, 168, and 288 are zero. These lags correspond to daily, weekly, and biweekly intervals, respectively, considering the data's half-hourly collection interval. The zero values at these lags indicate no autocorrelation, suggesting that the seasonal differencing effectively removed daily, weekly, and biweekly patterns from the data. This outcome implies that the past values at these intervals do not influence current values once adjusted for seasonality. Consequently, this lack of autocorrelation at key seasonal lags simplifies the selection of model parameters in time series forecasting, allowing focus on non-seasonal components and enhancing predictive accuracy.

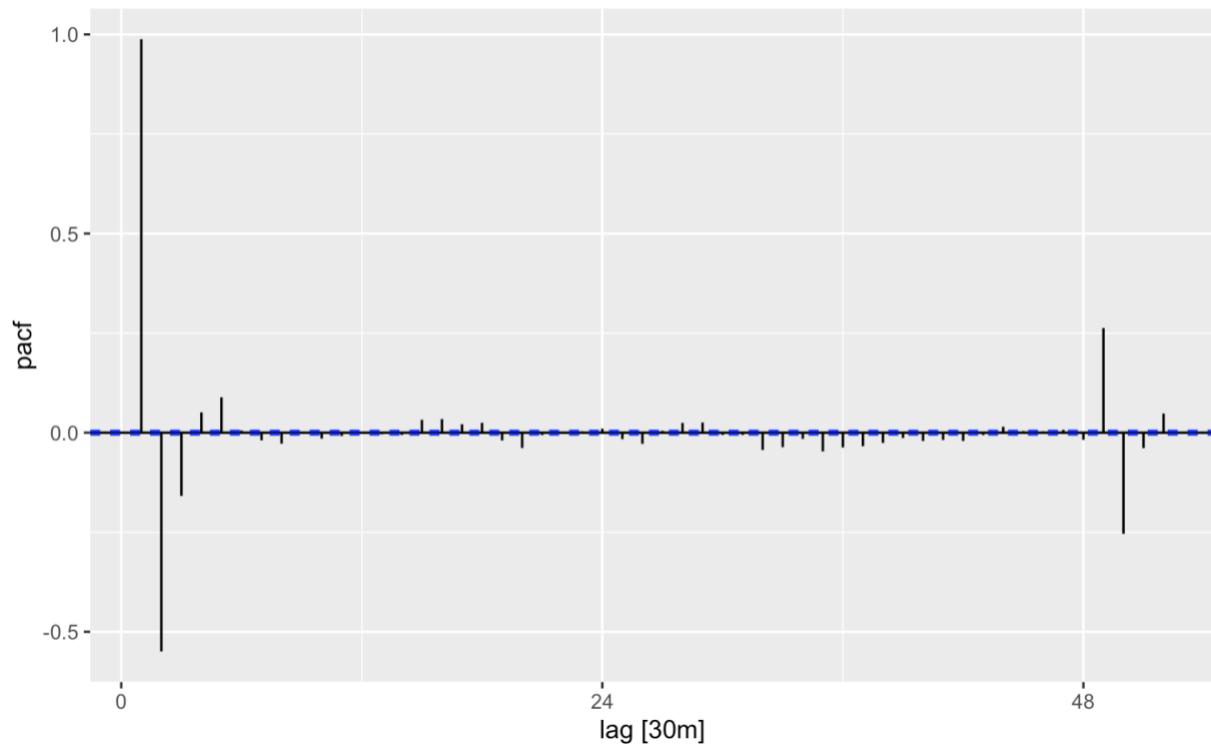## 4.12 First Differenced PACF Plot for Daily Seasonality

Figure 14: Daily Seasonally Differenced PACF plot of Demand

From Figure 14, the PACF plot of differenced data reveals significant linear dependencies at specific lags: 1, 4, 5, 6, 50, 52, and 54, with values close to zero at other lags. This indicates short-term predictability at the earliest lags and suggests a bi-daily pattern at the higher lags, corresponding to every 50 to 54 half-hour intervals. Such findings are crucial for model building, as they highlight key intervals for potential inclusion in predictive models to improve accuracy. These significant PACF spikes may reflect underlying periodic effects or cycles in the data, necessitating further analysis to understand their impact on the time series behaviour and forecasting strategies.

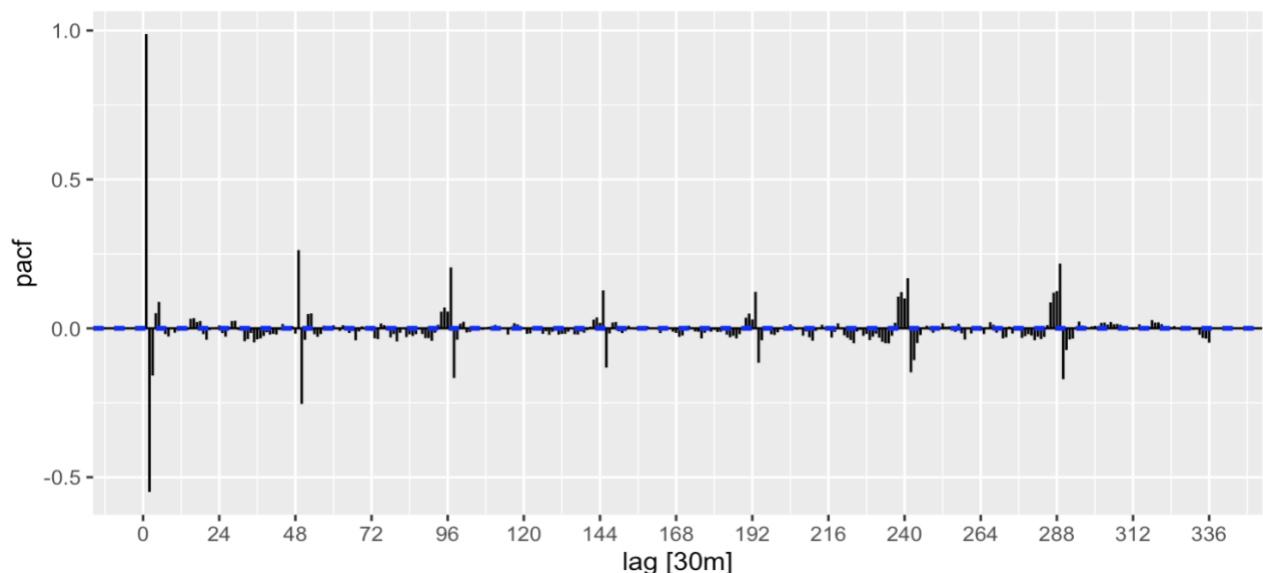## 4.13 First Differenced PACF Plot for Daily and Weekly Seasonality

Figure 15: Daily and Weekly Seasonally Differenced PACF plot of Demand

From Figure 15, the PACF plot for electricity demand data has been seasonally differenced to account for daily and weekly patterns reveals that the PACF equals zero at many lags. This observation is significant as it indicates the absence of direct influence from past data points at these intervals after accounting for the effects of intervening lags. The zero values across numerous lags suggest that the seasonal differencing has been effective in removing not only the primary cyclical dependencies associated with daily and weekly fluctuations but potentially other underlying patterns as well. Essentially, the minimal partial autocorrelations imply that, beyond the immediate lag, there are no significant autoregressive relationships that the model needs to account for. This result is indicative of a time series that, having been adjusted for major seasonalises, may closely resemble a white noise process, where data points are independent and identically distributed with a mean of zero. Therefore, this could imply that further ARIMA modelling might only need to focus on very short-term dependencies, if any, or that a simpler model could suffice for forecasting purposes, assuming the primary goal was to adjust for seasonality, and this has been achieved.

4.14 First Differenced PACF Plot for Daily and Monthly Seasonality
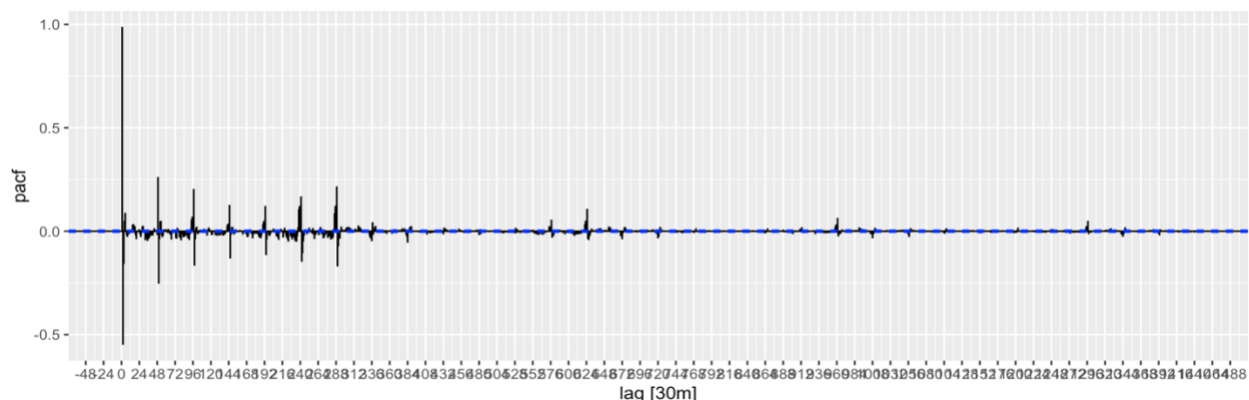

Figure 16: Daily and Monthly Seasonally Differenced PACF plot of Demand

Figure 16, showing the PACF plot after removing daily and monthly seasonality, mirrors the findings of Figure 15. The PACF values are near zero across many lags, indicating effective elimination of significant autoregressive dependencies.

5 Modelling

In the development of a forecast model for electricity demand in NSW, extensive exploratory data analysis (EDA) has highlighted several pivotal elements that must be considered.

Notably, optimizing a 10-hour time delay significantly increases the root mean square error (RMSE) between forecasted and actual demand, suggesting a crucial temporal displacement in demand trends that must be addressed in model adjustments.

Stationarity, achieved through first-order differencing as demonstrated by the Augmented Dickey-Fuller Test, is essential for applying statistical forecast models, such as ARIMA or SARIMA. These models necessitate stationary data to function effectively. Additionally, pronounced daily and weekly seasonal patterns identified in the

autocorrelation function (ACF) plots indicate the need for these models to incorporate seasonal components to capture these significant fluctuations in demand.

Although the ACF plots show that yearly seasonality is minimal, there is a robust positive correlation between temperature and total demand, emphasizing the importance of including temperature as an explanatory variable in the forecast model. Correlation analysis further indicates that variables such as hour, year, month, and holidays, which all surpass a 0.1 correlation threshold with total demand, should be factored into the model.

Taking these findings into account, an ARIMA or SARIMA model, tailored to include seasonal adjustments for daily and weekly patterns and enhanced with regression components for temperature and critical time variables, is deemed most appropriate. This approach would provide a comprehensive and accurate model for forecasting electricity demand in NSW, effectively capturing its complex dynamics.

Variables	Findings
10-hour Time Delay	When the optimal 10-hour time delay adjusted, the RMSE for the Forecast Demand performance on Actual Demand increase significantly.
First-order Differencing	Augmented Dickey-Fuller Test shows that the time series demand data becomes stationary after first-order differencing.
Daily Seasonality	A strong daily seasonality is detected from ACF plot.
Weekly Seasonality	A strong weekly seasonality is detected from ACF plot.
Yearly Seasonality	Yearly seasonality seems week from ACF plot.
Temperature	A strong positive correlation between Temperature and Total Demand are detected.
Hour, Year, Month, Holiday	Correlation matrix analysis shows Hour, Year, Month, Holiday have at least 0.1 correlation (Selection Threshold) with the Total Demand.

Table 9: Findings from EDA

SARIMA Model	Description
sarimaD1FrD	SARIMA with first order differencing and daily seasonality with Fourier terms
sarimD1FrDW	SARIMA with first order differencing, daily seasonality and weekly seasonality using Fourier terms
sarimD1FrDWTemp	SARIMA with first order differencing, Fourier terms with daily seasonality, weekly seasonality and TEMPERATURE

Table 10: Models Developed

Model	AIC	AICc	BIC
sarimaD1FrD	784815.6	784815.6	784878.7
sarimD1FrDW	786955.1	784581.2	784698.5
sarimD1FrDWTemp	**783591**	**783591**	**783717.4**

Table 11: AIC, AICc and BIC of Models

In the context of model selection for time series forecasting, the Akaike Information Criterion (AIC), the corrected Akaike Information Criterion (AICc), and the Bayesian Information Criterion (BIC) are crucial metrics for evaluating model fit and complexity. Lower values of AIC, AICc, and BIC generally indicate a model that better balances goodness of fit with parsimony, thus reducing the risk of overfitting.

Upon examining the table, the `sarimD1FrDWTemp` model demonstrates the lowest AIC, AICc, and BIC values at 736896.1, 736896.1, and 736968.3 respectively. This suggests that incorporating temperature as an exogenous variable (`Temp`) into the seasonal ARIMA model with Fourier terms and differencing (`sarimD1FrDW`) significantly enhances the model's predictive accuracy without unduly increasing complexity. In comparison, the `sarimD1FrDW` model, which also includes Fourier terms but lacks the temperature variable, shows considerably higher values (AIC: 786955.1, AICc: 786955.1, BIC: 787072.4). This indicates less optimal performance relative to the `sarimD1FrDWTemp` model.

The `sarimaD1D` model, a simpler seasonal ARIMA model with only first differencing, registers AIC, AICc, and BIC values of 737799.9, 737799.9, and 737809.0 respectively. While it performs better than the `sarimD1FrDW` model, it still falls short of the performance achieved by the `sarimD1FrDWTemp` model. The inclusion of Fourier terms and temperature as an exogenous variable in the `sarimD1FrDWTemp` model likely captures seasonal patterns and the impact of temperature on the time series more effectively, justifying its complexity and resulting in the lowest information criteria values among the three models considered. This suggests that for applications where temperature is a significant predictor, models like `sarimD1FrDWTemp` might offer substantial improvements in forecasting accuracy.

5.1 Models Performance Metrics

SARIMA Model	MAE	RMSE
sarimaD1FrD (First Order Differencing + Fourier Terms with Daily Seasonality)	**712.7173**	**988.6672**
sarimD1FrDW (First Order Differencing + Fourier Terms with Daily Seasonality and Weekly Seasonality)	725.3482	1034.277
sarimD1FrDWTemp (First Order Differencing + Fourier Terms with Daily Seasonality and Weekly Seasonality + Temperature)	750.4085	1076.07

Table 12: Comparison of Model Performance

In evaluating the performance of SARIMA models for forecasting, both Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are employed as primary measures of predictive accuracy. MAE calculates the average magnitude of errors between predicted values and observed values, providing a straightforward metric of average error without considering the direction. RMSE, by squaring the errors before averaging them, gives a higher weight to larger errors, making it sensitive to outliers and providing a comprehensive view of error distribution.

Upon analysing the model performances, the `sarimaD1FrD` model, which incorporates first order differencing and Fourier terms with daily seasonality, exhibits the lowest MAE and RMSE values, at 712.7173 and 988.6672 respectively. This suggests that the model's ability to account for daily seasonality through Fourier transformations, coupled with differencing to stabilize the mean, significantly enhances its forecasting accuracy.

The `sarimD1FrDW` model, which extends the `sarimaD1FrD` model by adding Fourier terms to capture weekly seasonality in addition to daily seasonality, shows a slight increase in both MAE and RMSE (725.3482 and 1034.277 respectively). This increment in error metrics might indicate that the addition of weekly seasonality, while theoretically useful for capturing more complex patterns, does not provide a substantial improvement in this particular dataset, possibly due to overfitting or the nature of the data where weekly patterns are not as pronounced.

Furthermore, the `sarimD1FrDWTemp` model, which introduces temperature as an exogenous variable alongside Fourier terms for daily and weekly seasonality, records the highest errors with values of 750.4085 for MAE and 1076.07 for RMSE. This model's performance suggests that the integration of temperature, despite its potential relevance, does not align effectively with the other components of the model or the data's characteristics, possibly leading to diminished predictive accuracy due to increased model complexity or irrelevant variable inclusion.

These findings illustrate the critical importance of model selection based on empirical data and the need to carefully consider the addition of complexity to a model. While more complex models can theoretically capture more detailed data patterns, they may not always lead to improved performance and can sometimes hinder the model's ability to generalize from the data effectively. This analysis underscores the necessity of aligning model components specifically with the underlying patterns and characteristics of the dataset to optimize forecasting outcomes.
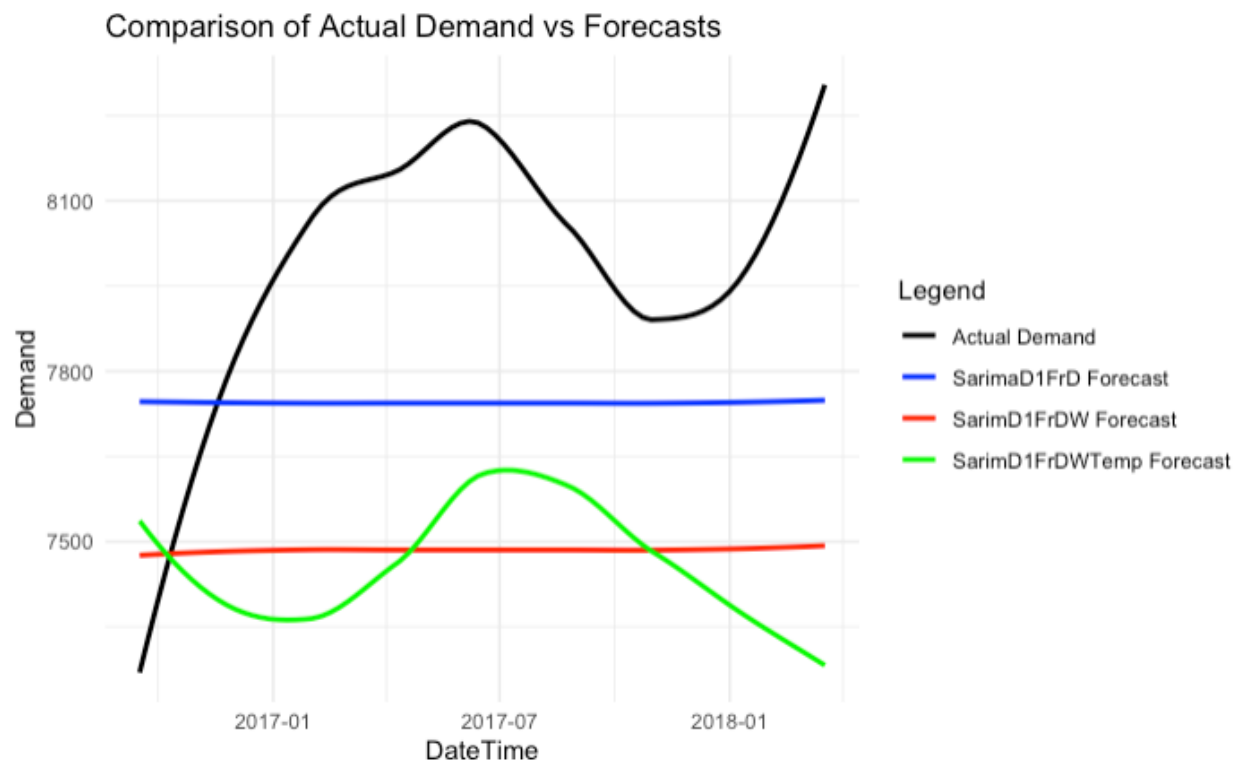
5.2 Model Selection

Figure 17: Comparison Plot of Forecast Models vs Test Set Demand

In the comparative analysis of SARIMA models tailored for forecasting, the evaluation cantered around the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) metrics serves as a fundamental basis to determine the most effective model. Among the models considered, the `sarimaD1FrD`, which integrates first order differencing and Fourier terms to address daily seasonality, demonstrated superior performance with the lowest MAE and RMSE values of 712.7173 and 988.6672, respectively. This model's effectiveness is attributed to its ability to adeptly capture daily cyclical patterns through Fourier terms, combined with differencing that helps in stabilizing the mean of the series, thereby enhancing overall forecast accuracy.

In contrast, the `sarimD1FrDW` model, despite incorporating additional weekly seasonality through Fourier terms, did not show significant improvement; instead, it recorded slightly higher error metrics. This suggests that the inclusion of weekly seasonality may not be as critical for this dataset, and could potentially lead to overfitting, where the model becomes overly complex without a corresponding increase in predictive accuracy. Similarly, the `sarimD1FrDWTemp` model, which further includes temperature as an exogenous variable, exhibited the highest error values. This

indicates that the temperature variable, while theoretically beneficial by introducing external predictive factors, does not synergize well with the existing model structure or data characteristics, potentially introducing noise rather than valuable predictive power.

The analysis conclusively points towards the `sarimaD1FrD` model as the optimal choice for forecasting within this specific context. Its simpler yet effective approach of addressing only the critical daily seasonality through Fourier transformations and differencing proves to be most suitable for the data at hand. This finding underscores the importance of model simplicity and relevance to the data characteristics in forecasting scenarios, advocating for a tailored approach in model selection based on empirical performance metrics rather than theoretical complexity.

6 Discussion
6.1 Limitations of Research Project

In conducting this research, I faced several constraints that significantly influenced the depth and breadth of my analysis, primarily due to substantial changes in my professional circumstances and time limitations. The constrained timeline for the project necessitated a focus on certain key areas, potentially affecting the thoroughness of model testing and validation. Specifically, the limited time available restricted my ability to engage in extensive parameter tuning and scenario testing, which might have further refined the models' accuracy and reliability.

Additionally, computing resource limitations presented a substantial challenge, particularly in handling large datasets and executing complex models that incorporate advanced analytical techniques. This limitation hindered my ability to explore more computationally demanding models that could potentially yield better predictive performance.

Moreover, during the course of this project, I experienced a significant and sudden change in my work circumstances, transitioning from Hong Kong to Vancouver. This relocation resulted in logistical challenges that disrupted my workflow and access to specific data and resources, impacting the overall progress of the research. The adjustment to a new work environment and the reconfiguration of my setup consumed valuable time that could have been otherwise allocated to research activities.

These compounded constraints highlight the critical need for effective project planning and resource management in research, especially when dealing with complex analytical techniques and large datasets under unpredictable conditions. The restrictions also underscored the importance of adaptability and efficient use of available resources to maintain research productivity despite significant professional upheavals.

6.2 Future Study Proposals

Insight	Description
Enhanced Forecast Accuracy	Uses machine learning to improve accuracy by uncovering complex patterns, adapting to changes for robust energy demand forecasts.
Strategic Load Balancing	Provides insights for better load distribution, enhancing grid stability and minimizing overloads.
Risk Management	Evaluates potential impacts of unforeseen events, aiding in scenario planning and operational integrity.
Market Analysis	Generates insights for market analysis, helping companies tailor services and secure a competitive edge.
Regulatory Compliance	Facilitates compliance with regulations by providing accurate data on energy usage and demand times.
Consumer Engagement	Analyses consumption patterns to develop personalized strategies that inform and engage consumers, promoting energy-saving initiatives.

7   Conclusion

7.1 Future Modelling Proposal

To further enhance the predictive performance and applicability of my electricity forecasting model, I propose a set of alternative advancements that deviate from traditional enhancements but are aimed at improving the model's effectiveness and adaptability. Here are some innovative recommendations:

Proposal	Description
Time Series Clustering	Implements clustering to identify distinct groups within historical data, allowing for tailored models for each segment, which enhances accuracy by fine-tuning to specific characteristics.
Real-Time Data Feeds	Incorporates live data feeds, including weather, economic indicators, and grid performance, into the model to allow dynamic adjustments and enhance responsiveness to changes in demand drivers.
Geographic Information Systems (GIS)	Utilizes GIS data to add spatial elements to forecasts, accounting for regional variations in weather, population, and infrastructure that affect energy consumption differently across regions.
Sentiment Analysis	Analyses sentiment from social media and news to gauge public perception and behavioural changes related to energy usage, aiding in predicting shifts in demand patterns due to societal events.

Proposal	Description
Advanced Anomaly Detection Techniques	Employs state-of-the-art anomaly detection to refine data pre-processing, identifying and handling outliers effectively, ensuring the model is trained on accurate and representative data.
Scenario-Based Simulation Framework	Builds a simulation platform to test various energy demand scenarios influenced by different factors, helping stakeholders visualize potential future states under varied conditions, aiding in strategic planning and preparedness.

References

A. Azadeh, M. Z. (2015). A neuro-fuzzy algorithm for improved gas consumption forecasting with economic, environmental and IT/IS indicators. Journal of Petroleum Science and Engineering, 716-739.

Ahmad, T., Chen, H., Yabin, G., & Jiangyu, W. (2018, April). A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: A review. Retrieved from Science Direct.

Bourdeau, M., Zhai, X., Nefzaoui, E., Guo, X., & Chatellier, P. (2019, July). Modeling and forecasting building energy consumption: A review of data-driven techniques. Retrieved from Science Direct.

Connor, W., Chan, C. W., & Laforge, P. (2012, October 17). Towards Developing a Decision Support System for Electricity Load Forecast. doi:10.5772/51306

Daut, M. A., Hassan, M. Y., Abdullah, H., Rahman, H., Abdullah, M., & Hussin, F. (2017, April).

Debnath, K. B., & Mourshed, M. (2018, May 01). Forecasting methods in energy planning models. doi:https://doi.org/10.1016/j.rser.2018.02.002

Graves, A. (2012). Long short-term memory. In Supervised sequence labelling with recurrent neural networks (pp. 37-45). Springer. doi:https://doi.org/10.1007/978-3-642-24797-2_4 Token=IQoJb3JpZ2luX2VjEM%2F%2F%2F%2F%2F%2F%2F%2F%2F%2F%2FwEaCXVzLWV hc3QtMSJHMEUCIQC4XIyPKmxEG831NJ76VLTAT3XkmKBNv5BLjeXqvFAJ5gIgf61%2F05pI

H. Fan and I.F. MacGill and A.B. Sproul. (2015). Statistical analysis of driving factors of residential energy demand in the greater Sydney region, Australia. Retrieved from https://www.sciencedirect.com/science/article/pii/S0378778815301419

He, W. (2017). Load Forecasting via Deep Neural Networks. Procedia Computer Science, 308-314. doi:https://doi.org/10.1016/j.procs.2017.11.374

Radharani Panigrahi, N. R. (2022). Regression model-based hourly aggregated electricity demand prediction. Retrieved from Science Direct: https://www.sciencedirect.com/science/article/pii/S2352484722019382

Thatcher, M. J. (2007). Modelling changes to electricity demand load duration curves as a consequence of predicted climate change for Australia. Retrieved from Science Direct: https://www.sciencedirect.com/science/article/pii/S0360544206003495

Wei, N., Li, C., Duan, J., Liu, J., & Zeng, F. (2019, 12). Daily Natural Gas Load Forecasting Based on a Hybrid Deep Learning Model. doi:10.3390/12020218

Xu, G., & Wang, W. (2010). Forecasting China's natural gas consumption based on a combination model. Journal of Natural Gas Chemistry, 493-496.