CAPSTONE PROJECT BY TEAM GROUP 3

A DATA SCIENCE APPROACH TO FORECAST ELECTRICITY CONSUMPTION IN NSW

John Student (z123456), Jim Student2 (zID), Rishantha Rajakaruna (z5441528).

School of Mathematics and Statistics
UNSW Sydney

September 2024

Submitted in partial fulfilment of the requirements of the capstone course ZZSC9020

Plagiarism statement

I declare that this thesis is my own work, except where acknowledged, and has not been submitted for academic credit elsewhere.

I acknowledge that the assessor of this thesis may, for the purpose of assessing it:

- Reproduce it and provide a copy to another member of the University; and/or,
- Communicate a copy of it to a plagiarism checking service (which may then retain a copy of it on its database for the purpose of future plagiarism checking).

I certify that I have read and understood the University Rules in respect of Student Academic Misconduct, and am aware of any potential plagiarism penalties which may apply.

By signing this declaration I am agreeing to the statements and conditions above.

Signed: _____ Date: _____

Signed: _____ Date: _____

Signed: _____ Date: _____

Acknowledgements

TBW...

25/07/2020.

Abstract

TBW ...

Contents

Chapter 1

Introduction

Accurate Electricity demand forecasting is a fundamental component of decision making for governments, regulatory bodies and businesses. To forecast electricity demand we need to understand the drivers of demand fluctuation and corresponding relationships. Weather changes, long-term climate change and macroeconomic influences such as, population growth, economic growth are key factors that influence demand [a20(2024)]. Additionally, generation of electricity through renewable sources such as Solar and Wind which doubled over the last decade [Energy.gov.au Department of Climate Change and Water(2024)] has added more complexity to forecasting due to its dependency on weather.

The demand forecasting varies from short term to medium and long term. The parameters that influence each period type varies. Short term forecasting focus on small time intervals. i.e. few mins upto few days. It can be heavily influenced by daily weather, time of the day and holidays. In case of medium to long term forecast, the timeframe could vary from few months to few years into the future and determined by long term factors such as climate change, population growth, government policy to name a few.

This research paper focus on short term forecasting. Therefore, primarily use half hourly temperature, historical half hourly actual demand, calendar information (holidays, weekends) and time of the day as key data sources. We analyse in detail the relationship between demand and the aforementioned features and use two models, **XGBoost** and **Prophet by Facebook** to forecast daily demand.

By using multiple matrices, we compare the performance of each model. Additionally, we also compare the results of the two models against the short-term demand forecast published by Australian Energy Market Operator.

We hope the findings will assist market participants to make informed decisions with regards to short term electricity demand forecasting and contribute to a more sustainable and efficient energy landscape.

Chapter 2

Literature Review

Here are a few references that can be useful: [Xie et al.(2018)Xie, Allaire and Grolemund] and [Lafaye de Micheaux et al.(2013)Lafaye de Micheaux, Drouilhet and Liquet]. See also https://bookdown.org/yihui/rmarkdown-cookbook/

In order to incorporate your own references in this report, we strongly advise you use BibTeX. Your references then needs to be recorded in the file `references.bib`.

Chapter 3

Material and Methods

3.1 Software

Primary software used for analysis and model build is Python. We have used extensive set of python libraries. List of libraries used are,

- holidays - Derive public holidays in NSW
- pandas - Data selection and manipulation
- numpy - Calculations and data manipulation
- statsmodels - Used for statistical tests and statistical data exploration
- matplotlib, seaborn - To create plots and visualize statistical analysis
- scipy - statistical analysis
- sklearn - Model preparation and analysis
- xgboost - Used for XGBoost model build
- prophet - Used for Facebook Prophet model build

Jupyter Notebook and RStudio was used as integrated development environments. Github repository was used for code management. Teams and One Drive were used for collaboration and communication.

3.2 Description of the Data

Following Data sets are used for this project.

1. Total Electricity Demand NSW
2. Air Temperature NSW
3. NSW Calendar
4. Total Forecast Demand NSW

3.2.1 Total Electricity Demand NSW

Total Electricity Demand in 30 min increments. This data is sourced from the Market Management System database, which is published by the market operator from the National Electricity Market (NEM) system.

Row Count	File Size (Approx)	File Format	File Name
196513	5.6 MB	CSV	totaldemand_nsw.csv

3.2.2 Air Temperature NSW

Air temperature in NSW (as measured from the Bankstown Airport weather station). This data is sourced from the Australian Data Archive for Meteorology. Note: Unlike the total demand and forecast demand, the time interval between each observation may not be constant (i.e. half-hourly data). As noted in the literature

Attributes	Description	Attribute Characteristics
DATETIME	Date and time interval of each observation. Format (dd/mm/yyyy hh:mm)	Timestamp
TOTALDEMAND	Total demand in MW	Numeric
REGIONID	Region Identifier (i.e. NSW1)	String. Categorical

review, temperature is a key driver of demand. Therefore, this dataset is critically important for the research question

Row Count	File Size (Approx)	File Format	File Name
220326	6.7 MB	CSV	temperature_nsw.csv

Attributes	Description	Attribute Characteristics
DATETIME	Date and time interval of each observation. Format (dd/mm/yyyy hh:mm)	Timestamp
TEMPERATURE	Air temperature (°C)	Numeric
LOCATION	Location of a weather station (i.e., Bankstown weather station)	String Categorical

3.2.3 NSW Calendar

Use NSW Calendar to include holidays, seasons, weekend information for the model build

Attributes	Description	Attribute Characteristics
DATETIME	Date and time interval of each observation. Generated through Python library	Timestamp
HOLIDAY	Marked 1 if public holiday in NSW, otherwise 0. Generated using 'holidays' Python library Sourced from the python library https://pypi.org/project/holidays/	Numeric
SUMMER	Marked 1 if the month is in Summer season, else 0. Use one hot encoding	Numeric
AUTUMN	Marked 1 if the month is in Autumn season, else 0 Use one hot encoding	Numeric
WINTER	Marked 1 if the month is in Winter season, else 0 Use one hot encoding	Numeric
SPRING	Marked 1 if the month is in Spring season, else 0 Use one hot encoding	Numeric
WEEKDAY	Marked 1 if day of week is between Monday to Friday, else 0	Numeric
DAYOFWEEK	Marked 1 to 7 to indicate day of week	Numeric
MONTH	Month of the Year	Numeric

3.2.4 Total Forecast Demand NSW

Forecast demand in half-hourly increments for NSW. Data also sourced from the Market Management System database. This dataset would be valuable for us to validate the outcome of our model. Especially to understand the accuracy.

Row Count	File Size (Approx)	File Format	File Name
10906019	722 MB	CSV	forecastdemand_nsw.csv

Attributes	Description	Attribute Characteristics
DATETIME	Date and time interval of each observation. Format (dd/mm/yyyy hh:mm)	Timestamp
FORECASTDEMAND	Forecast demand in MW	Numeric
REGIONID	Region Identifier (i.e. NSW1)	String Categorical
PREDISTPATCHSEQNO	Unique identifier of predispatch run (YYYYMMDDPP). In energy generation, "dispatch" refers to process of sending out energy to the power grid to meet energy demand. "Predispatch" then is an estimated forecast of this amount.	String (Identifier)
PERIODID	Period count, starting from 1 for each predispatch run.	Numeric (Identifier)
LASTCHANGE	Date time interval of each update of the observation (dd/mm/yyyy hh:mm)	Timestamp

3.3 Pre-processing Steps

3.3.1 Merge and read zip files

All of the data files (excluding NSW Calendar) were stored in zip files. Instead of exacting the the content of the files, we read directly the zip file content for data cleaning. Only exception was the forecast demand dataset where it was split to two zip files. Therefore these files had to be merged as one zip prior to consumption.

3.4 Data Cleaning

Following activities were done as part of data cleaning.

- Verify the data types, no of rows/columns and measure of central tendency.
- Removal of unused columns.
  - Total Electricity Demand NSW - 'REGIONID' was removed as this attribute had only one value and was not useful.
  - Air Temperature NSW - 'LOCATION' was removed as this attribute had only one value and was not useful.
- Verify whether NULL values exist. None found in the files.
- Validate whether duplicate rows exist
  - Total Electricity Demand NSW - None found
  - Air Temperature NSW - 13 duplicate rows were removed
- Verify whether data is missing by validating that all the dates are available between the minimum and maximum dates in the file
  - Total Electricity Demand NSW - None found
  - Air Temperature NSW - Data for three dates were missing. 2016-07-16 till 2016-07-18. No action was taken as its a small percentage and also data is too far in the past and was not relevant for our research question.
- Validate whether demand is recorded consistently for each day.
  - Total Electricity Demand NSW - It was noted that for 2021, there were only 2 months of data. Also for month of March, there was only one row. Therefore this had to be removed to ensure consistency.

  - Air Temperature NSW - Temperature readings were not restricted to 30min intervals.Therefore we verified whether a temperature reading exist for every 30min. Where temperature readings were missing, we used fill forward method to add missing values. No of readings missing were 579 which is a small percentage.
- Generate Calendar Data set. Python library *holidays* was used to identify NSW holidays. Combining a date range and the holiday dates, we created a new calender dataset. Calender was limited to the date range of demand and temperature dataset.Additional attributes such as season (spring, summer, etc), day of week, weekday were derived.

Final dataset was prepared after completing the above activities. This dataset was stored as a separate csv file for further consumption. It should be noted that additional attributes 'HOUR' and 'PEAK' were also included. Here 'PEAK' is defined as '1' when the Time of the day is between 7:00 AM and 10:00 PM. The time period for peak was derived from https://www.canstarblue.com.au/electricity/peak-off-peak-electricity-times/.

3.5 Assumptions

The temperature data is limited to one location. Bankstown Airport. However the electricity demand is measured for the entire state. As we know, the temperature varies across different locations within the state. Therefore we make an assumption that temperature at a single point is sufficient for demand forecast for the entire state.

Popularity of rooftop solar has increased over the years whilst manufacturing landscape has changed (as noted in literature review). The impact of these factors are not part of the analysis due to lack of publicly available data.

3.6 Modelling Methods

As per our initial research combined with literature review and considering the time constraints, we will focus on two models for this project. An additive model **Prophet** by **Facebook** and another popular short term demand forecasting machine learning model **XGBoost**. These two models will be tested across various features such as temperature, hour of the day, holiday/non-holiday, weekdays/weekends and peak/off-peak period.

Since the research question relates to short term demand, we intend to use the latest available data instead of past data. Therefore the dataset would be restricted to 3 years. We would also use 80:20 split of data for training and testing. In relation to hyper-parameter tuning, based on the model either random or grid search would be used.

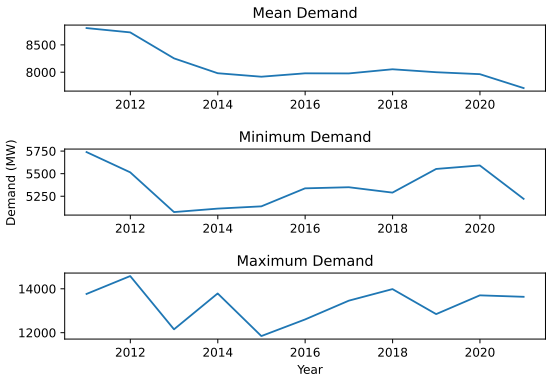
Chapter 4

Exploratory Data Analysis

Lets begin by analyzing the data to understand its characteristics.

4.1 Yearly Electricity Demand

Analysis of the electricity demand over the years would help identify historic trends and any seasonal effects. As a start, lets review average, Minimum and Maximum demand fluctuation over the years.



It is evident that the the average demand has reduced or flat lined over the years. We were expecting the demand to rise with the population growth over the years. Therefore, it would be prudent use latest available data for model build as we are focusing on short term demand.

The minimum and maximum demand seem to fluctuate within a range and does not indicate any significant trends.

Note: A potential reason for decrease in demand could be the increase use of roof top solar panels. Below graph from Australian PV Institute https://pv-map.apvi.org.au/analyses shows a clear increase in Solar panel installation. However further research is required to confirm any relationship.

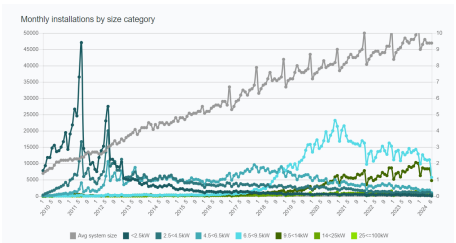
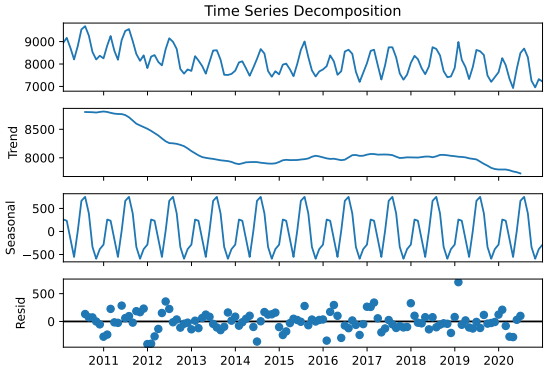


Figure 4.1: Yearly Solar PV Installation

4.2 Decompose Time series

Electricity demand data is of time series nature. According to Australian Energy market operator (AEMO), Time series models are more applicable in short-term forecasting [Australian Energy Market Operator (AEMO)(2023)] similar to our research question. Time series data can be decomposed to four components. [Brownlee(2017)]

- Level - The average value in the series.
- Trend - The increasing or decreasing value in the series.
- Seasonality - The repeating short-term cycle in the series.
- Noise - The random variation in the series.



The Level and Trend plots both show gradual decline in demand similar to what we observed earlier. The seasonal plot shows the peaks and troughs in a repetitive pattern. This maybe due to relative high usage of electricity during Winter (for heating) and Summer (for cooling) compared to Spring and Autumn.

Further, to verify that the data set used is suitable for time series analysis we perform a stationarity test using ADF (Augmented Dickey-Fuller).

The null(H0) and alternate hypothesis(H1) of ADF test are:

- H0: The series has a unit root (value of a=1),the series is non-stationary.
- H1: The series has no unit root, the series is stationary.

If we cannot reject the null hypothesis, we can say that the series is not stationary, and if we do, it is considered stationary.

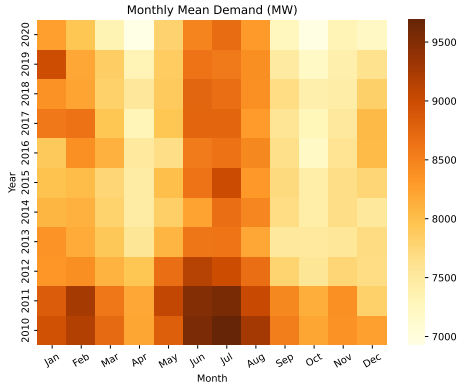
```
## Results of Dickey-Fuller Test:

## Test Statistic                  -5.586991
## p-value                          0.000001
## #Lags Used                      28.000000
## Number of Observations Used    3989.000000
## Critical Value (1%)              -3.431990
## Critical Value (5%)              -2.862265
## Critical Value (10%)             -2.567156
## dtype: float64
```

Based on the results we can observe that the test statistic is lower than the critical values. Therefore we can reject the null hypothesis and conclude that the time series is stationary.

4.3 Monthly Demand

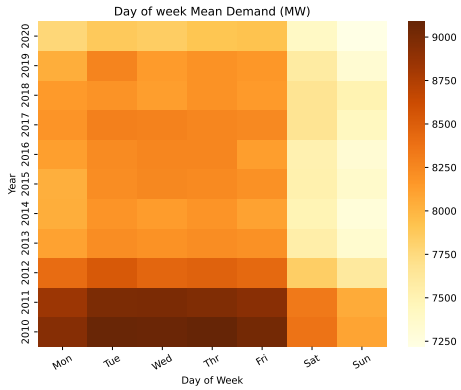
Lets analyse the impact of demand based on the month.



We tend to use heating during winter months and cooling during summer. The heatmap clearly indicates that June, July, August winter months and January, February Summer months have a higher average demand for electricity. Conversely Spring and Autumn has a lower average demand. Therefore the month/season should be considered for the model build.

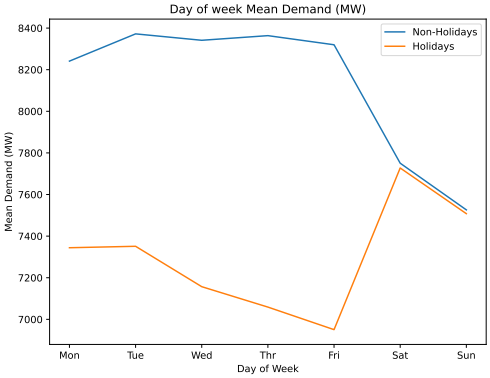
4.4 Day of the week Demand

Next we analyse whether the electricity demand fluctuate depending on the day of the week. As per the heatmap, weekends tend to have lower demand. This could be due to the fact that most offices, factories are closed during the weekend. Also, many people tend to spend weekends outside. Similar to month, day of the week seems to have an influence on the demand. Hence suitable to be included in the model build.



4.5 Demand on Holidays

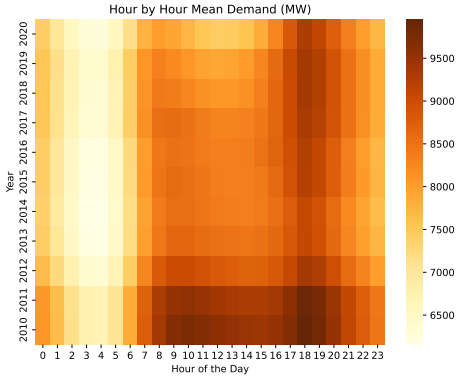
Extending on the day of the week demand, we would like to analyse the difference in average demand between holidays and non-holidays. For the purpose of the analysis, we consider Public holidays and weekends as *Holidays* and all the other days as non-holidays.



The graph clearly indicate a significant difference in mean demand between holidays and non-holidays. As noted previously for Saturday and Sunday, this may be due the fact that offices, factories not operating over holidays resulting in lower demand. Therefore we could conclude that holidays has a impact on the overall demand and therefore should be considered in the model.

4.6 Hour of the day demand

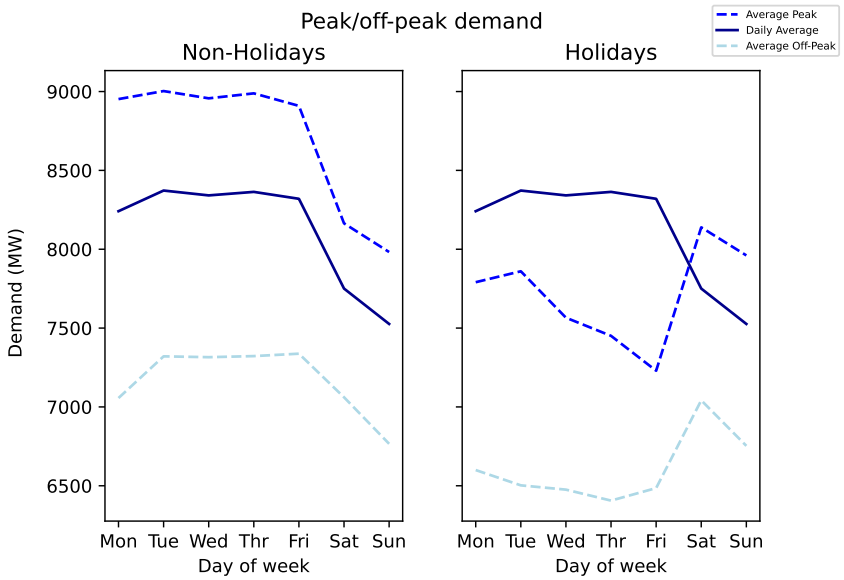
At the next granular level, we would like to observe how demand fluctuates within a day (hour by hour). Australian energy providers broadly segregates hourly demand to three groups [Wrigley(2019)]. Peak, off-peak and shoulder. There are variations of this by providers. For our analysis purposes, we would simplify to Peak and Off-peak only.



As evident from the graph, approx from 7:00 AM to 10:00 PM, the demand seems to be high. Therefore we would use that period as Peak demand and rest as off-peak demand for our model build.

4.7 Peak vs Off-Peak demand

Based on previous analysis, lets verify the variations in demand based on Peak hours vs off-peak. Additionally, we further drill-down on holidays vs non-holidays.



The graphs clearly indicate the difference in demand between peak and off-peak hours. This pattern is visible both during non-holidays and holidays. Therefore we should consider Peak/Off-Peak demand in the model build.

4.8 Autocorrelation & Lag

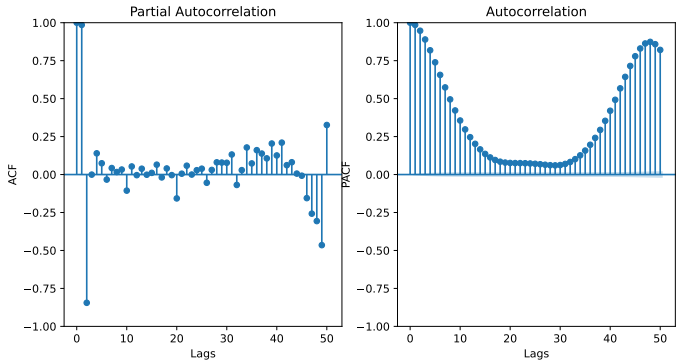
We would like to understand the influence of Lag in the chosen model. Prior to that lets review the concept.

Autocorrelation (serial correlation) is correlation between two values of the same variable at times t and t_k. When a value from a time series is regressed on previous values from that same time series, it is referred to as an autoregressive model. e.g y_t and y_{t-1}

$$y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t$$

The above is a first order autoregressive model. Meaning only one proceeding value is used as predictor variable and is written as AR(1). If we used two previous values as predictors, then it would be a second order autoregressive model i.e. AR(2). This can be generalised as AR(k), i.e k^{th} order autoregressive model [The Pennsylvania State University()].

The autocorrelation function (ACF) is given as, $Corr(y_t, y_{t-k}), k = 1, 2, ..n$ [NIST and Technology(2020)] where k is the time gap or the lag between values of the same variable. We are interested in Partial Autocorrelation, which measure the association between y_t and y_{t-k} directly and filter out the linear influence of the random variables that lie in between. PACF is useful to identify the order of autocorrelation.

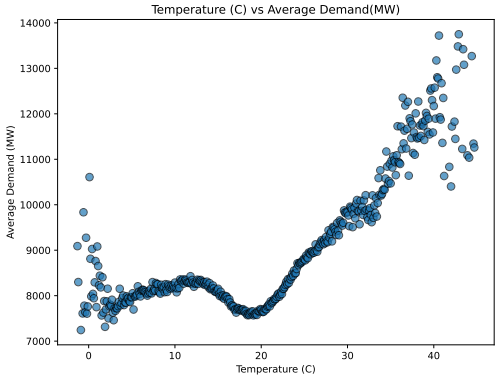


The ACF plot seems to indicate cyclical pattern and this could be due to seasonality effect on the demand data. It also seem to highlight high correlation between adjacent data points.

The PACF graph shows significant spike in lag 1,2 and 3 but seems to decay as it moves along. Hence it maybe useful to limit to 3 lags for the model.

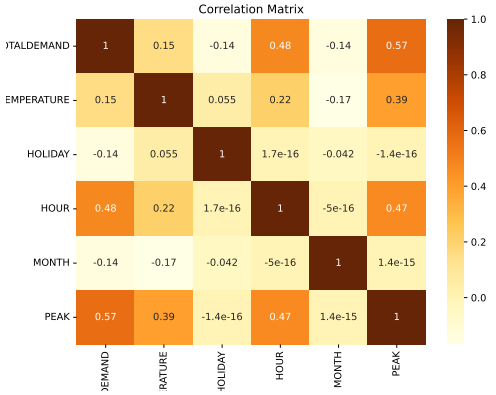
4.9 Temperature and demand relationship

The relationship between temperature and electricity demand is well known. In below grpah, as the temperature increase, it is clearly evident that demand increases. However it is interesting to note that when temperature decreases, especially below 10 degrees, we see a limited spike in demand. Potential reason could be that in NSW temperature falls mostly during the night / early morning and therefore consumers do not necessarily need heating. However the high temperatures are mostly during day time and as a result people use electricity for cooling driving up the demand.



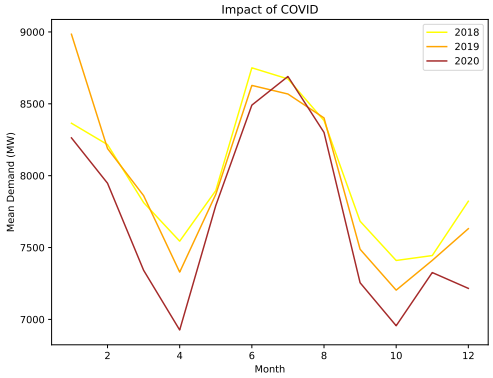
4.10 Correlation matrix

Finally we look at the correlation between the attributes. The graph below shows very low correlation between Total Demand and Temperature this due the the fact that the relationship is non linear. Similarly, Holiday and Month seems to have little or no correlation with respect to demand. Hour and Peak/off-peak in turn seems to have higher correlation.



4.11 Covid impact on Demand

Since the dataset used overlaps with the Covid period, it is important to understand if there is any impact on overall demand.



The plot does not indicate any significant deviations in the demand. One could conclude that 2020 covid period had minimal impact on overall demand based on the graph. However, further research is warranted to confirm the finding.

Chapter 5

Analysis and Results

5.1 A First Model

Having a very simple model is always good so that you can benchmark any result you would obtain with a more elaborate model.

For example, one can use the linear regression model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots \beta_p x_{pi} + \epsilon_i, \qquad i = 1, \ldots, n.$$

where it is assumed that the ϵ_i's are i.i.d. $N(0, 1)$.

Chapter 6

Discussion

Put the results you got in the previous chapter in perspective with respect to the problem studied.

Chapter 7

Conclusion and Further Issues

What are the main conclusions? What are your recommendations for the "client"? What further analysis could be done in the future?

A figure:

Figure 7.1: A caption

In the text, see Figure 7.1.

References

[a20(2024)] , 2024. Nem forecasting and planning guidelines. URL: https://aemo.com.au/en/energy-systems/electricity/national-electricity-market-nem/nem-forecasting-and-planning/forecasting-approach/forecasting-and-planning-guidelines.

[Australian Energy Market Operator (AEMO)(2023)] Australian Energy Market Operator (AEMO), 2023. Forecasting Approach: Electricity Demand Forecasting Methodology. Technical Report. Australian Energy Market Operator. URL: https://aemo.com.au/-/media/files/electricity/nem/planning_and_forecasting/nem_esoo/2023/forecasting-approach_electricity-demand-forecasting-methodology_final.pdf. accessed: 2024-09-22.

[Brownlee(2017)] Brownlee, J., 2017. How to decompose time series data into trend and seasonality. URL: https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/.

[Energy.gov.au Department of Climate Change and Water(2024)] Energy.gov.au Department of Climate Change, Energy, t.E., Water, 2024. Renewables — energy.gov.au. URL: https://www.energy.gov.au/energy-data/australian-energy-statistics/renewables#:~:text=In%202023%2C%2035%25%20of%20Australia.

[Lafaye de Micheaux et al.(2013)Lafaye de Micheaux, Drouilhet and Liquet] Lafaye de Micheaux, P., Drouilhet, R., Liquet, B., 2013. The R Software: Fundamentals of Programming and Statistical Analysis. Statistics and Computing, Springer New York. URL: https://books.google.fr/books?id=Ji-8BAAAQBAJ.

[NIST and Technology(2020)] NIST, N.I.o.S., Technology, 2020. 1.3.5.12. autocorrelation. URL: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35c.htm.

[The Pennsylvania State University()] The Pennsylvania State University, P.E.C.o.S., . 10.2 - autocorrelation and time series methods - stat 462. URL: https://online.stat.psu.edu/stat462/node/188/.

[Wrigley(2019)] Wrigley, K., 2019. Peak and off-peak electricity times - tariffs and rates. URL: https://www.canstarblue.com.au/electricity/peak-off-peak-electricity-times/.

[Xie et al.(2018)Xie, Allaire and Grolemund] Xie, Y., Allaire, J., Grolemund, G., 2018. R Markdown, The Definitive Guide. Chapman and Hall/CRC. URL: https://bookdown.org/yihui/rmarkdown/.

Appendix

Codes

Add you codes here.

Tables

If you have tables, you can add them here.

Use https://www.tablesgenerator.com/markdown_tables to crete very simple markdown tables, otherwise use LaTeX.

Tables	Are	Cool
col 1 is	left-aligned	$1600
col 2 is	centered	$12
col 3 is	right-aligned	$1