Comp20008 Assignment 2 Report

Due 19th May 2022
Word Count (excluding titles, tables and in-text citations): 2191

Table of Contents

Research Question and Target Audience

How can the trends in energy demand over the past year be explained by weather conditions?

The purpose of this project is to investigate how patterns in weather can help predict energy consumption or demand in different Australian states**.** Given the multiple ways weather can be measured, a variety of independent variables have been tested to best discover which weather metric is most highly correlated with the energy-demand data. Importantly, the report differentiates energy-demand and energy-consumption. Energy demand is concerned with how much energy is used at a particular point in time (AMEO, 2021). Energy consumption is concerned with a period over which electricity or gas is demanded (AMEO, 2021). Thus, consumption is the product of demand and time. Some measurements in the dataset are more suited for energy consumption whilst others are more suite for demand, justifying the separation of such data.

The results from the experiment could be utilised by economists to develop and improve predictive models that are used by policy makers. The improved forecasting via this experiment can assist with subsidisation initiatives that promote ‘green energy’. Meteorologists could use the weather data in tandem with their natural-disaster predictions to ascertain how floods or droughts impact a business or household’s energy consumption pattens.

Dataset Description and Contextualisation

1. 4 CSV files containing weather metrics for the cities: Brisbane, Melbourne, Sydney, and Adelaide

The weather metrics used as explanatory variables were

- Minimum and Maximum temperature – *useful during winter and summer where these two extremities are more distinguished*
- Rainfall and Evaporation – u*seful during natural-disaster prone seasons*
- Sunshine hours: *useful for checking whether solar-powered electricity would be suitable for a household or business*
- 9am and 3pm Wind Metrics (Direction, Max Speed, Time of Maximum Gust): *useful for understanding where potential wind-powered turbines should be positioned*
- 9am and 3pm relative humidity – *useful for improving the accuracy of the ‘feels like’ section on the weather apps*
- 9am and 3pm cloud cover

- 9am and 3pm mean sea level (MSL) pressure – *considered to be the least relevant out of all the weather metrics given that its connection with electricity-demand is unclear and indirect*

An important limitation from this dataset was that some weather metrics were unavailable for some cities (Adelaide had no data for Evaporation and Sunshine).

2. CSV file containing energy demand between February 2021 and March 2022 for the states: Queensland (QLD), New South Wales (NSW), South Australia (SA) and Victoria (VIC)

   Total demand was recorded based on region (state) as opposed to city and was measured in megawatts (MW). A price surge column existed in the dataset, but the feature is not prevalent for this investigation.

   Demand was recorded at inconsistent intervals (both 30 minute and 5-minute intervals throughout a 24-hour period), requiring further pre-processing.

Pre-Processing and Wrangling

1. **Weather CSV Files per City:**

These files contained multiple cells with missing values or had data inconsistencies; namely, multiple data types in the same column and outliers. Using statistical measurement imputation, these missing values and outliers were replaced by the mean to retain all data records. This was considered an appropriate data-quality adjustment given that the information was not fatal. Moreover, missing data would hinder the calculation of Euclidean distance required for the KNN regression model.

Additionally, some features were categorical such as the wind direction which made such cells difficult to semantically interpret. Therefore, these categorical variables were converted into a numerical form.

2. **Pre-processing of the Demand CSV file**

This file had a column called *SettlementDate* which included the Date and Time as one entity. This was inconsistent with the way the date was recorded for the weather CSV files. For consistency, two regular expressions were created that took parts of the raw string in the *SettlementDate* column and separated the date and time from each other.

Given that the research question is concerned with demand and consumption, the Dataframe was expanded to include an additional column that contained the daily consumption of energy based on demand. The *'price surge'* column was deleted given that all the Boolean values were ‘false’, and thus stakeholders who choose to use this program should assume that prices remain relatively stable throughout the duration of the dataset provided. Therefore, this feature should not affect the results and is thus unnecessary for this model.

3. Merging two Dataframes together based on Date and Region

Both ‘Date’ columns of the two files were converted into the same datetime format using an in-built Pandas module. It was observed that the weather files were recorded on a city basis whereas the demand file were recorded by state. As most state populations live within capital cities (Melbourne population is around 76.2% of VIC and the Adelaide population is around 73.7% of SA), the weather conditions of the cities were renamed as their state to allow the merging of the datasets (ABS, 2020).

4. Creation of two distinct data frames after merging (for each state)

The first Dataframe contained weather measurements measured over a period and its dependent variable column was the energy consumption per day.

The second Dataframe contained weather measurements measured at a point in time, with the dependent variable column being the energy demand at 9am and 3pm.

Understanding relationships with Normalised Mutual Information (NMI)

NMI was considered an appropriate strategy to analyse the correlation between one feature variable and the output variable (consumption or demand). Multiple weather features when plotted against consumption illustrated a non-linear shape, eliminating Pearson Correlation (PC) as an effective method.

During the calculation of the NMI score, data was discretised, with the number of bins set to 20. This was considered an appropriate number due to the relatively large number of rows to analyse. However, different bin sizes could change the estimations of MI, making this a limitation for this investigation.

Table 1.1- Normalised Mutual Information Scores for New South Wales

	Feature	Correlation
Top 3 Consumption Metrics	Minimum Temperature	0.225
	Maximum Temperature	0.200
	Sunshine	0.165
Top 9am Metric	9 am Temperature	0.240
Top 3pm Metric	3 pm Temperature	0.195

Table 1.2 - Normalised Mutual Information Scores for Queensland

	Feature	Correlation
Top 3 Consumption Metrics	Minimum Temperature	0.252
	Maximum Temperature	0.196
	Sunshine	0.176
Top 9am Metric	9 am Temperature	0.159
Top 3pm Metric	3 pm Relative Humidity	0.204

Table 1.3 - Normalised Mutual Information Scores for South Australia

	Feature	Correlation
Top 3 Consumption Metrics	Minimum Temperature	0.195
	Maximum Temperature	0.209
	Time of maximum wind gust	0.151
Top 9am Metric	9 am Temperature	0.207
Top 3pm Metric	3 pm Temperature	0.196

Table 1.4 - Normalised Mutual Information Scores for Victoria

	Feature	Correlation
Top 3 Consumption Metrics	Minimum Temperature	0.178
	Maximum Temperature	0.215
	Evaporation	0.171
Top 9am Metric	9 am Temperature	0.218
Top 3pm Metric	3 pm Temperature	0.195

Figure 1.1 Queensland Correlation Scatter Plot

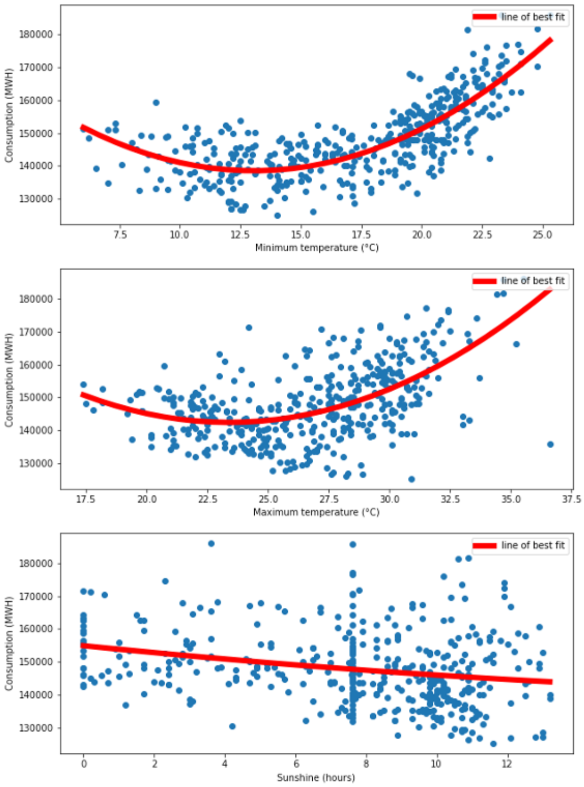


Figure 1.2 Victoria Correlation Scatter Plot

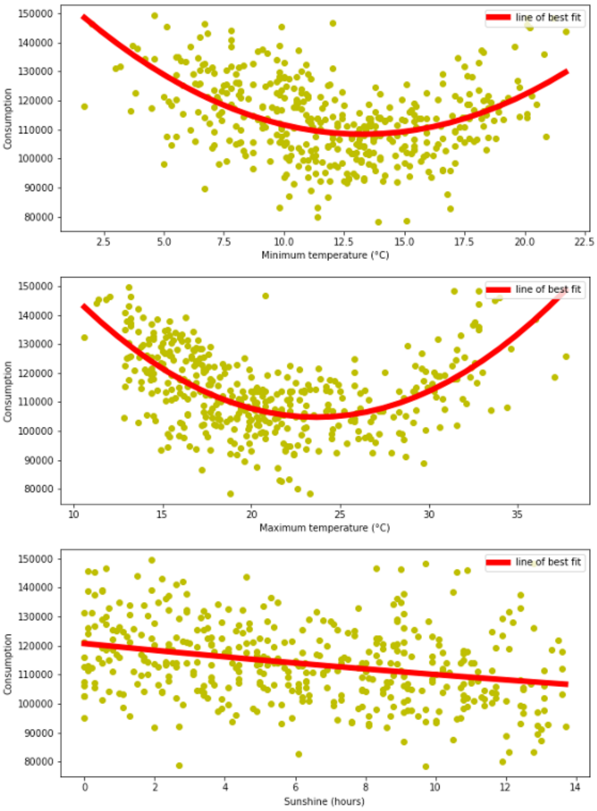


Figure 1.3 – New South Wales Correlation Scatter Plot

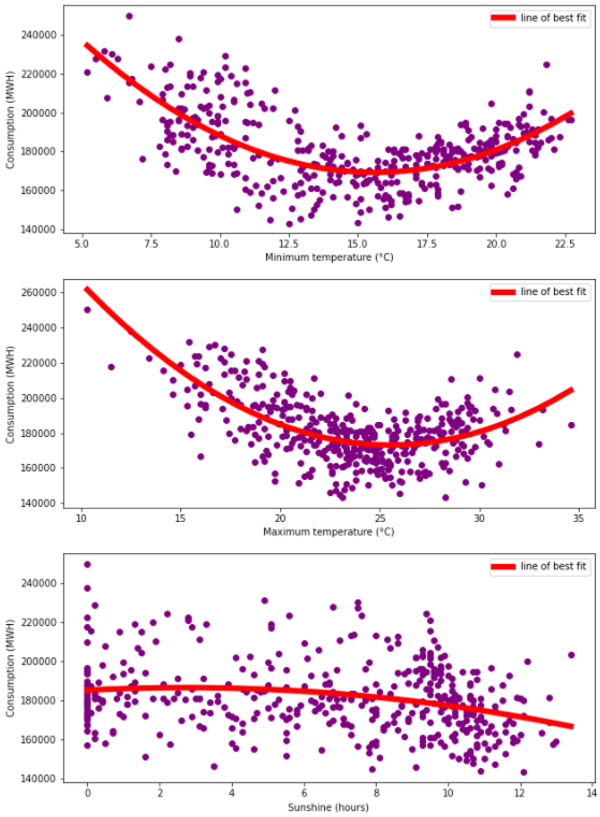
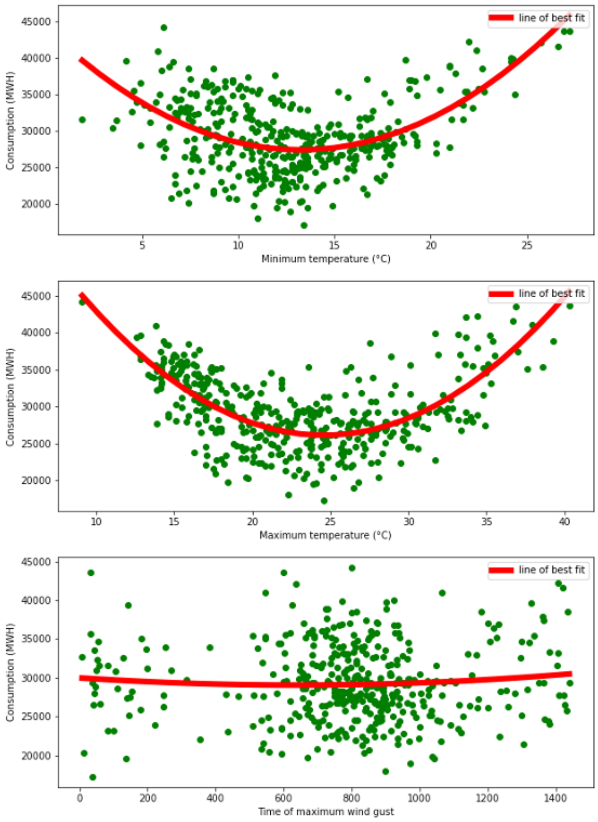


Figure 1.4 – South Australia Correlation Scatter Plot



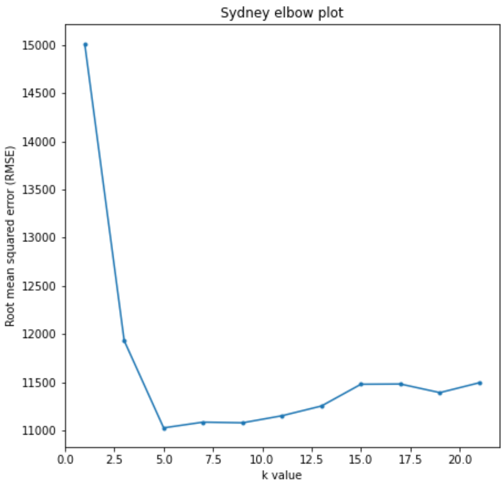
Analysis methods:

K Nearest Neighbours (KNN) Regression Algorithm

This regression algorithm is used for the prediction of a continuous output variable. Upon initial visualisation, it was evident that the relationship between weather parameters and corresponding energy consumption and demand was non-linear. Therefore, a linear regression analysis was unable to be applied. Instead, KNN regression was chosen to deal with this non-linear relationship.

The algorithm involves calculating the average distance from the numerical target to its k-neighbours (Sci-Kit Learn, 2022). The hyperparameter (k) was chosen based on the 'elbow plot' method which involves plotting a range of k values against the RMSE and finding the minimum. This is the optimal number of neighbours where error will be most reduced (Figure 2).

Figure 2: Elbow Plot



Univariate Feature Selection Process:

Chi Squared Test and Mutual Information Scoring (MI) were considered because of the non-linearity of the dataset. For Chi-Squared Test, the lower the Chi Squared value, the higher the correlation and dependence between a feature and its class variable. After performing the null hypothesis testing, it was discovered that not enough features were rejected, meaning this technique was not considered viable.

For MI, a high score indicates that the weather feature selected is well correlated with the class, strongly predicting the class label. To perform the MI calculation, the continuous data was discretised by applying a 3-bin equal width frequency. The top three highest scoring features were then selected to be inputs for the machine learning (ML) model. Since MI scores could be ranked, this method was chosen.

Training and Testing

The data was split into a 70% training and 30% testing ratio. This ratio ensures the model does not overfit consistently, decreasing variance error and keeping bias error relatively low (predicts well on training data). Whilst in an ideal situation, both a well-fitting model (best learning result) and highly accurate performance (best validation) would be preferred, performing a training and testing split comes with this inherent trade-off.

Within the training dataset, 10-fold cross validation was performed. This ensures that all 10 sub-samples have an opportunity to be a validating sub-sample as part of the training model, removing the limitations associated with the hold-out method (all data is used for training as well as testing), improving accuracy. The model results are then averaged across all the folds.

Considering the small sample size, 10 folds seemed appropriate for this dataset.

The process was repeated for the 'time-specific' datasets (9am and 3pm weather measurements), with energy-demanded used as the y output variable to be predicted.

Figure 3: Good Fit Observation from Queensland Dataset after using KNN Regressor Algorithm

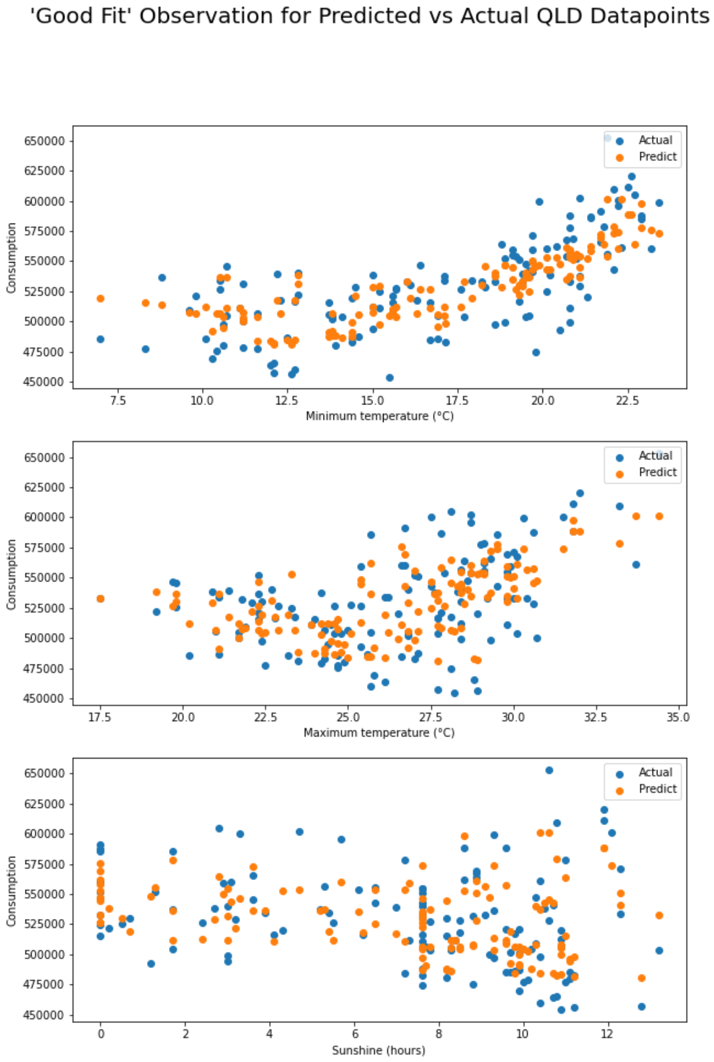


Figure 4: Linear Relationship Between Actual and Predicted Consumption (QLD)

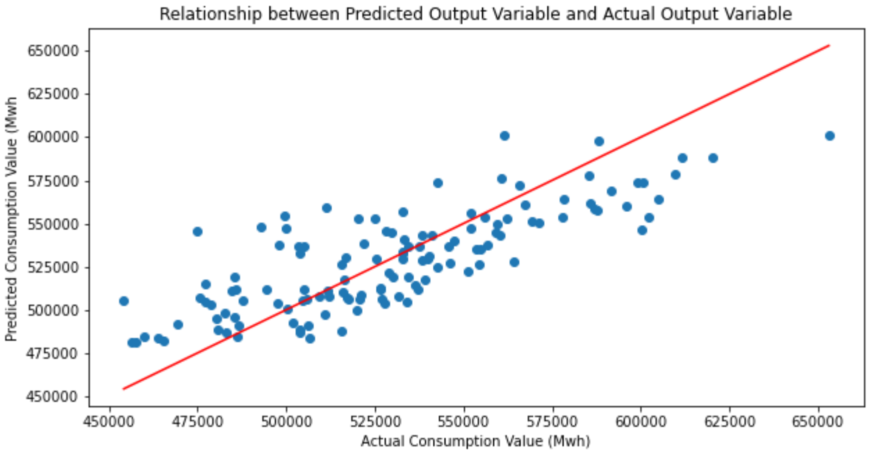


Table 2: Means Absolute Error (MAE)

Machine Learning Experiment	*QLD*	*NSW*	*VIC*	*SA*
Consumption	4% – 5399.82	6% -- 10130.96	7% -- 7541.68	13% -- 3831.91
9 am Metrics	8% – 446.16	9% -- 698.44	13% -- 638.3	21% -- 251.5
3pm Metrics	7% – 423.15	10% -- 725.02	13% -- 590.06	26% -- 234.47

Table 3: Root Means Squared Error (RMSE)

Machine Learning Experiment	*QLD*	*NSW*	*VIC*	*SA*
Consumption	5% -- 6645.27	7% -- 12550.88	8% -- 9376.33	16% - 4734.97
9 am Metrics	9% -- 537.08	12% -- 888.41	16% -- 801.4	25% -- 305.63
3pm Metrics	9% -- 545.03	13% -- 948.25	15% -- 710.6	34% -- 298.44

Table 4: R-squared value

Machine Learning Experiment	*QLD*	*NSW*	*VIC*	*SA*
Consumption	0.63	0.402	0.524	0.07
9 am Metrics	0.39	0.384	0.26	0.003
3pm Metrics	0.554	0.283	0.32	0.320

Discussion of Key Results and Significance

This experimental result answers the research question from two angles:

1. It relaxes the assumption that weather and energy-consumption/demand are linearly related
2. It confirms that temperature is a core indicator of how consumers and businesses will utilise electricity and gas

Analysis of Correlation and Scatter Plots

Across all four states, the minimum and maximum temperature data appeared to be the most correlated with energy-demand and energy-consumption. Consumption tends to increase at extreme temperatures, hence creating a parabolic shape (Figures 1.1-1.4). This relationship between consumption and extreme temperatures seems feasible given that households and businesses are more likely to use heating or cooling devices when temperatures fall or rise drastically.

Interestingly, the time of maximum wind gust was considered a highly correlated variable against energy-consumption for SA. This may be attributable to the Roaring Forties influence on SA, a strong wind pattern that occurs most noticeably in the Southern Hemisphere (NOAA, 2022). Additionally, over 70% of SA's population live in the capital city Adelaide, located along the coast where wind is most prevalent (ABS, 2020). This may impact consumption of energy through the empirical correlation recorded between a climates wind and its associated humidity. High wind gusts could lead to a change in humidity, affecting the perceived temperature for the residents, and increasing energy consumption.

The relationship between Sunshine and Consumption is most linear compared to other metrics, but this is not reflected in Tables 1.1-1.4 given the choice to use NMI over PC. These lines of best fits also exhibit a slightly negative slope. This indicates that as the amount of sunshine per day increases, consumption of energy decreases. A reduction in the use of artificial lighting on days with more sunlight could explain this. Policy makers could use this knowledge to encourage the purchase of solar panels during periods where there is low energy consumption and high amounts of sunshine.

Significance of results:

The result of this experiment demonstrates that the KNN regression algorithm is a somewhat effective predictor of non-linear continuous output variables. If weather is estimated for tomorrow, consumption per day can be predicted with an error of 10% on average.
The diversity of features from the dataset ensures that the ML model will be able to deal with foreign data and still reasonably predict future energy consumption/demand.

Analysis of the Machine Learning Performance:

The Mean Absolute Error (MAE) was chosen because its value is measured to the same scale as the residuals and is the most robust to outliers because residuals contribute to the total error proportionally.

The consumption model's MAE range (4-13%) was relatively low, indicating that the predicted output value for consumption/demand were close to the training values (Table 2). Thus, economists can be more confident that the predicted outcomes from this model are relatively accurate. SA recorded relatively higher MAE terms for all three models (13- 26%). SA was missing more data than other states, possibly hindering the training component. Therefore, meteorologists' analysis of SA should rely upon other factors contributing to trends in energy demand / energy consumption because of these high error terms.

Since RMSE is the most used ML performance analysis metric, it was included as a performance measure, but it is more sensitive to outliers because the MSE aspect amplifies residuals. The inflexible nature of the model does not consider the delayed demand in response to weather patterns at these two specific times. It instead assumes that the demand changes simultaneously with changes to weather (assuming a relationship exists) which is an unrealistic assumption, potentially explaining the higher RMSE for the 9am and 3pm models.

For all four states, the consumption ML model had the highest recorded R-Squared values. The R-squared value may not be valid goodness of fit given that the R-Squared metric assumes a linear relationship (Minitab, 2014). However, it can be used to measure the quality of regression; a higher r squared value indicates more variation of consumption or demand is explained by the variation in a weather metric (Table 4).

Limitations and Improvements:

Limitations:

The period over which data was recorded was not entirely over 2021-2022. There were no values for January 2021, nor were there values for the rest of 2022. This impacted the ability to develop a research question that explained a trend throughout an entire year.

The demand file lacked a level of specificity that hindered the ability to answer the question in a more nuanced fashion. For example, there was no specification as to whether energy-demand was referring to energy in the form of electricity or gas. Furthermore, demand in suburbs instead of states could help policy makers better target their future investment strategies.

Recommendations:

1. Consumption was calculated by assuming that demand was constant over a period. This calculation of consumption is not always applicable to real situations. A better measurement would be the integral of demand with respect to time.
2. Collate data on a weekly basis by choosing features based on a weekend, weekday, and holiday period. It would be expected that on a weekend as people and businesses are not working, energy consumption may be higher
3. Employ a heuristic-based approach for feature selection along with a numerical metric to improve the validity of results, enabling more targeted research for professionals
4. Select multiple regression models such as Decision Tree Regression and Random Forest Regression to ascertain for the best fitting model for the data
5. Perform One-Hot Encoding when dealing with categorical data (e.g Wind Direction). PCA extraction would be required as an additional step.

References

AMEO. (2022). *Operational Consumption Definition*. Https://Www.Aemo.Com.Au/-/Media/Files/Electricity/NEM/Planning_and_Forecasting/Demand-Forecasts/Operational-Consumption-Definition.Pdf; AMEO.

Australia's Energy Markets and Systems. (2022). Retrieved 18 May 2022, from https://aemo.com.au/learn/energy-markets-and-systems

Headquarters, M., LLC Global. (2014). *Regression Analysis*. Retrieved 18 May 2022, from https://blog.minitab.com/en/tag/regression-analysis-3

Regional Population, 2020-21 Financial Year | Australian Bureau of Statistics. (2022, April 29). https://www.abs.gov.au/statistics/people/population/regional-population/latest-release

Sklearn. Neighbors. Kneighborsregressor. (2022). Scikit-Learn. Retrieved 18 May 2022, from https://scikit-learn/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html

US Department of Commerce, N. O. and A. A. (2022). *What are the Roaring Forties?* Retrieved 18 May 2022, from https://oceanservice.noaa.gov/facts/roaring-forties.html