**Melbourne’s Daily Energy Use and Pricing Based on Weather Data**

This report, codes, and supporting files can be accessed via this GitHub repository:

<https://github.com/YunaWuu/Data_Analytics_Assignment2_Group2/tree/main>

1. **Executive Summary**

This report was produced to investigate the relationship between weather and energy demand/pricing.

**Key Insights**

* There is an underlying causal relationship between energy demand and temperature with demand peaking when the temperature is at its coldest but also when it is at its hottest.
* There also appears to be a causal relationship between energy prices and weather. Increases in energy prices are especially evident during colder months (when the average daily temperature is less than 15°C).
* Given the limitations of the datasets and the techniques used in preparing the models, further work needs to be undertaken to improve the predictive capabilities of the models.

**Key Recommendations**

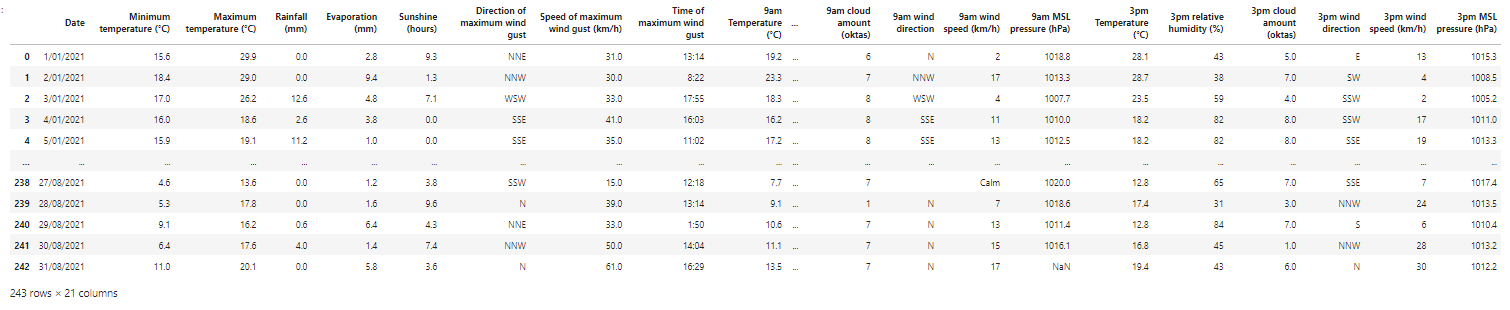
* It is strongly recommended that more advanced data analysis techniques such as the use of multiple regression, dimensionality reduction, or Principal component analysis be employed to create more sophisticated predictive models.
* Additional data may also be incorporated to reduce uncertainties and improve the predictive capabilities of the models.

1. **Data Preparation**
   1. **Datasets**

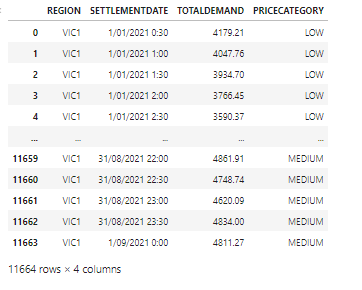
The following datasets were provided for this exercise.

*Table A – Overview of datasets*

|  |  |  |
| --- | --- | --- |
| **Dataset name** | Weather data | Price Demand Data |
| **Description** | Key weather indicators for the City of Melbourne for each day between January to August 2021 inclusive from the Bureau of Meteorology. | Energy price and demand figures for the State of Victoria for each half-hour period between January and August 2021 from the Australian Energy Market Operator. |
| **Format** | csv | csv |
| **Number of columns** | 21 | 4 |
| **Number of rows** | 243 | 11,663 |



*Figure 1 – Snapshot of weather data (raw)*



*Figure 2 – Snapshot of price demand data (raw)*

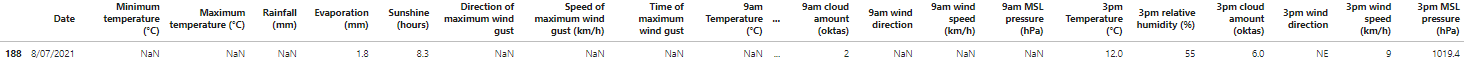
* 1. **Data Wrangling**

Analysis of both datasets indicates that there are no unexpected values (e.g. negative values where only positive values are expected) or outliers (e.g. unusually high temperature beyond 39°C).

* + 1. *Missing data*

However, analysis of the weather dataset indicates that there are missing observations from multiple variables across the weather data. The missing observations impact different variables on different dates. For example, the *9 am wind direction* is missing on 23/01/2021 and the *3 pm wind direction* is missing on 21/03/2021. This suggests that the missing values are missing completely at random (MCAR).

Crucially, 11 values are missing on 08/07/2021 including temperature data for the day. For Model 1, the median maximum temperature in °C across the weather dataset has been inputted to replace the missing value. This substitution should not have a material impact as it is a small subset of the data population (1 out of 243 records).



*Figure 3 – Weather data on 08/07/2021 before the substitution of missing information*

For Model 2, the following substitutions were made for missing values:

* For string datatype - substituted with the mode of the corresponding variable.
* For integer or float datatype - substituted with the mean of the corresponding variable.
  + 1. *Inconsistent formatting*

The sole common denominator identified for both datasets is the date values. However, the date values are inconsistent (“*dd/mm/yyyy*” (see column *Date* - Figure 1) vs “*dd/mm/yyyy hh:mm*” (see column *Settlement Date* - Figure 2). The date formats were converted so they are consistent.

* + 1. *Additional variable – Variance between maximum temperature to average maximum temperature*

For Model 1, the independent variable selected for the regression model is the variance between the maximum temperature of the day against the average maximum temperature across the provided dataset. A new variable for this was added to the dataset. Absolute value is used because energy consumption is expected to peak when the weather is at its coldest or hottest.

* + 1. *Additional variables – average temperature, average date temperature and average humidity*

For Model 2, the below variables were added as they scored highly in the correlation matrix with the maximum price category:

* Average temperature – average between maximum and minimum temperature of the day.
* Average day temperature – average between 9 am and 3 pm temperature.
* Average humidity – average between 9 am and 3 pm relative humidity.
  1. **Data Aggregation**
     1. *Aggregating daily energy demand*

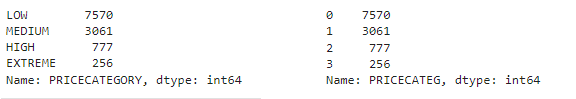
For Model 1, the Total Demand values were aggregated to obtain the daily total. This value will serve as the dependent variable (target) for Model 1.

As a result of the date conversion method employed, there is a half-hour data for 01/09/2021 as the data is incomplete and the removal will not be material to the dataset.

The total demand data is then joined with the weather dataset based on the date value.

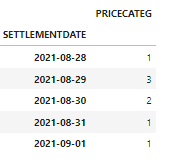
* + 1. *Assigning maximum daily price category*

To obtain the maximum daily price category for Model 2, a numerical value of 0-3 is first assigned to each price category:



*Figure 4 – Maximum daily price category data and assignment of numerical value.*

Then, the relevant dataset for Model 2 (date and price category) is grouped by the maximum value of the price category for each date.



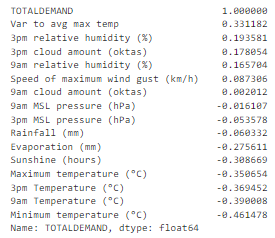
*Figure 5 – Maximum value of price category for each date*

For Model 2, the price category is then joined up with the weather dataset based on the date value.

1. **Model Building**
   1. **Model 1 – Prediction of Daily Total Energy Usage Based on Provided Weather Data**

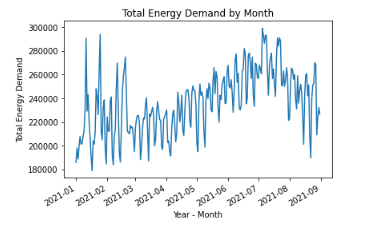
For Model 1, a simple linear regression model is employed to predict total energy use based on *var to avg max temp*. This method is selected as both variables contain numerical data and regression is a useful statistical tool to quantify the relationship between two variables.

This variable has a Pearson correlation coefficient of 0.3316 to total energy demand, which suggests a moderate correlation. There are other variables with stronger Pearson correlation coefficients (including negative correlations) such as Minimum temperature. However, visual inspection of line charts of these variables suggests that the correlation may be inconsistent. For instance, it is expected that the minimum temperature during winter months would be lower than during the warmer months but energy demand peaks during both the hottest and coldest days (see Figure 8).

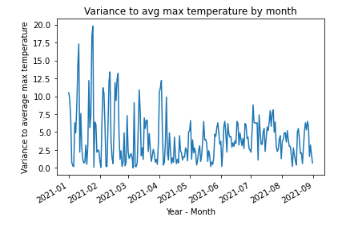


*Figure 6 – Person correlation coefficient of variables in weather data against the total demand*

A visual inspection of line charts created on the dependent (Figure 7) and independent variables (Figure 8) also suggests that there is a strong correlation. In particular, the correlation appears to be stronger during the warmer months (January 2021 to April 2021).

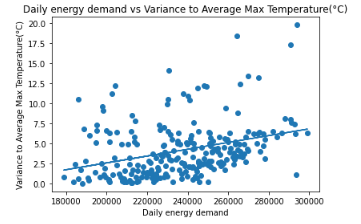


*Figure 7 – Line chart on daily total energy demand*



*Figure 8 – Line chart on daily variance to average maximum temperature*

The below scatter plot also provides a visual hint that there are potential causal relationships.



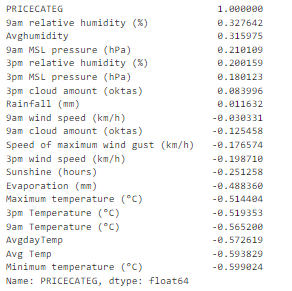
*Figure 9 – Scatter plot of daily energy demand vs variance to average max temperature*

For Model 1, 80% of the dataset is assigned as the training dataset with the remaining 20% assigned as the test dataset to avoid overfitting. Based on its learning from the training dataset, Model 1 predicts total energy demand based on the variance between the maximum temperature °C for the day to the average maximum temperature.

* 1. **Model 2 – Prediction of Maximum Daily Price Category Based on Provided Weather Data**

For Model 2, two types of models are employed – A K-Nearest Neighbour classifier and a Decision Tree. The classification models are selected because the target is finite (i.e. there are 4 price categories - low, medium, high and extreme).

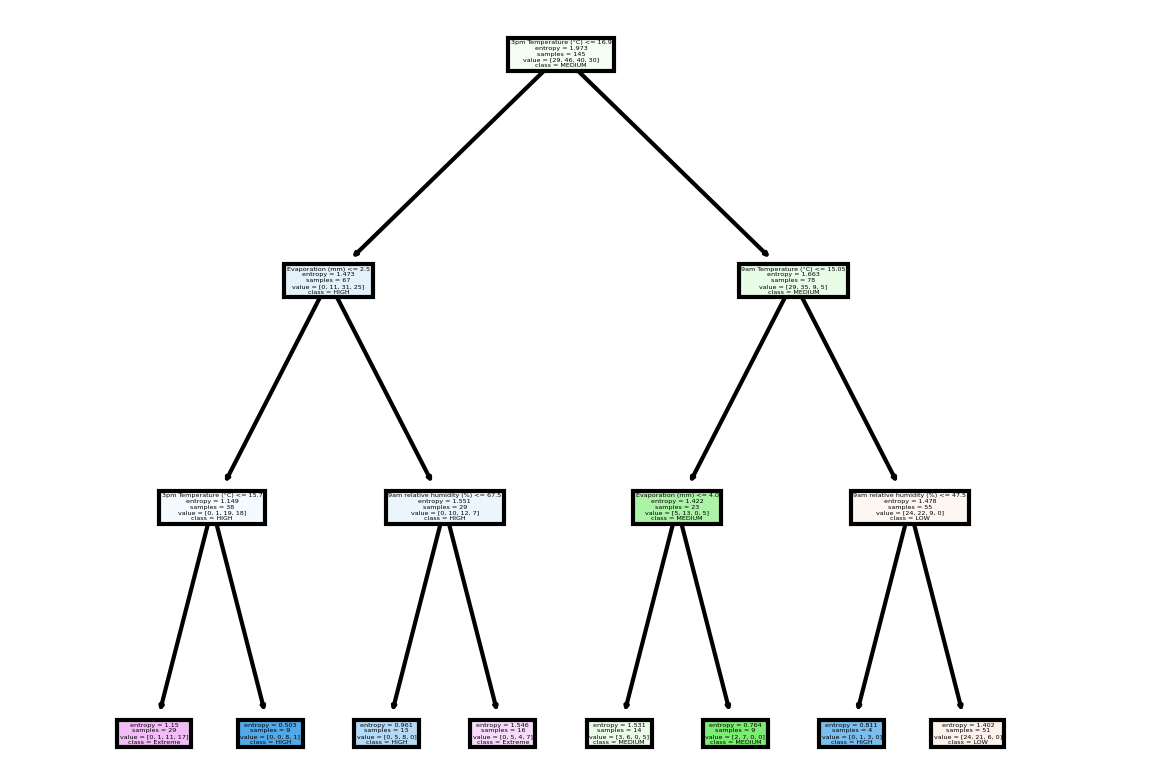
The features in the aggregated dataset are first reviewed using the correlation matrix. Features with correlation scores of +/- 0.2 and/or above are selected.



*Figure 10 – Correlation matrix of the features against price category*

For Model 2, 60% of the dataset is assigned as the training dataset with the remaining 40% assigned as the test dataset to avoid overfitting. Based on its learning from the training dataset, Model 2 predicts the classification of the maximum price category for the day based on the feature dataset.

For the Decision Tree, below is an outline of the model’s operations:



*Figure 11 – Decision tree model operations*

1. **Results**
   1. **Model Effectiveness**
      1. *Model 1 – Effectiveness*

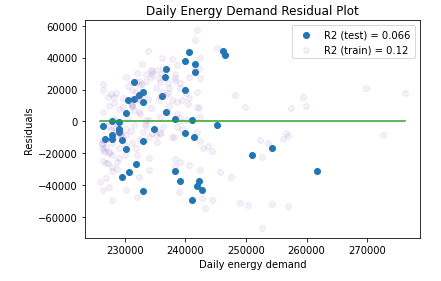
To test the effectiveness of Model 1, the following tests were performed:

* The coefficient of determination (r2) values were calculated against both the training and test dataset:

|  |  |
| --- | --- |
| r2 -Training dataset | 0.1149 |
| r2 -Test dataset | 0.0661 |

The r2 scores suggest that the model is not effective in predicting the total energy demand as only 6.61% - 11.49% of the variation between maximum temperature against average maximum temperature can be used to predict daily energy demand.

* The final test performed to assess the effectiveness of the model is a residual analysis.



*Figure 11 – Residual Analysis*

The residual analysis (refer to Figure 11) suggests that the residuals are independent. It is, however, observed that the residuals are not linear and the variances are not constant. This suggests that there are violations of the regression assumptions and that the linear regression model is not necessarily appropriate for two variables.

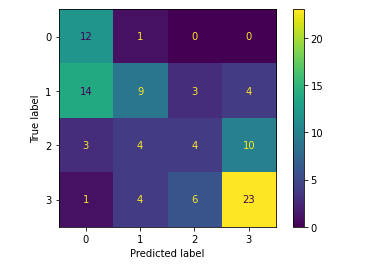
* + 1. *Model 2 – Effectiveness*

To test the effectiveness of the K-NN, the following tests were performed:

* The K-NN model has an accuracy score of 0.4591 which means that the model makes a correct prediction approximately 45.91% of the time.
* To further test the accuracy of the model, the K-fold cross-validation is employed. First, the dataset is split into 10 partitions. The first fold will be used as test data and the remaining fold will be used as training data. In the next iteration, the second fold will be the test data, and so forth until the last fold. Using this method, the accuracy score achieved for the K-NN model ranges from 0.4198 to 0.4692.

To test the effectiveness of the decision tree, the following tests were performed:

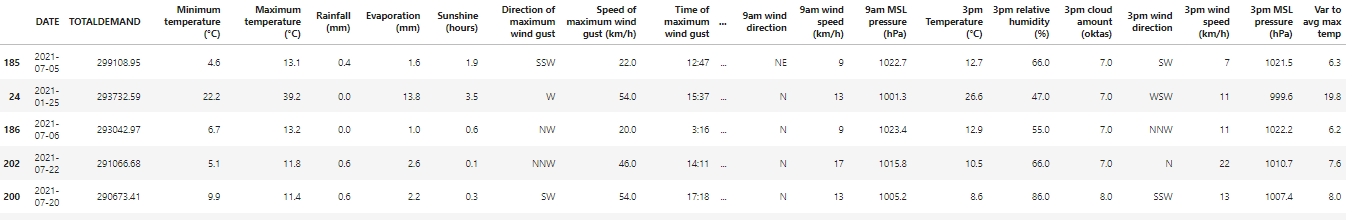
* The decision tree has an accuracy score of 0.4898 which means that the model makes a correct prediction approximately 48.98% of the time. The decision tree performs slightly better than the K-NN model.
* The accuracy score achieved by the model can also be visualized in a confusion matrix (see Figure 12 below).



*Figure 12 – Confusion Matrix*

* 1. **Insights from Data Analysis**

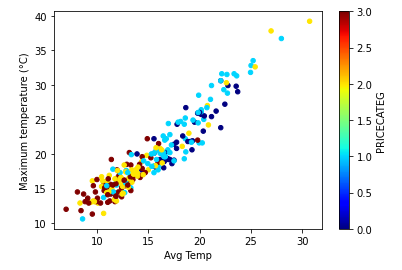
Across the dataset, total energy demand peaks on either the hottest days or the coldest days (see Figure 9 below). This is consistent with the initial conclusion that there is at least a moderate correlation between the variance between the maximum temperature to the average maximum temperature across the dataset.



*Figure 13 – Merged dataset ranked by total energy demand*

However, the use of this variable in Model 1 does not lead to highly accurate predictions of total energy demand. The variability in daily energy demand differs between colder and hotter days.

An observation of the distribution of the maximum daily price category (see Figure 14 below) suggests that Extreme price categories are more likely to occur when the temperature is colder. As observed below, the red dots representing the Extreme price category are more prominent where the average temperature is less than 15°C.



*Figure 14 – Distribution of maximum daily price category*

Model 2 performs considerably better than Model 1 in predicting the maximum price category.

Observation suggests that energy prices seem to be at their highest during colder weather when the energy demand is high. Interestingly, high energy demand during hotter weather does not seem to reach the same level of energy prices.

An improved model that can provide more accurate predictions of energy demand/usage based on weather data can be used by energy providers to plan for peak demand periods and prevent outages. From a profitability perspective, energy companies may also wish to increase their pricing when energy demand peaks during warmer months.

* 1. **Limitations & Improvements**
     1. *Model 1 – Limitations & Improvements*

For Model 1, a linear regression model was utilized and therefore only a single independent variable was considered. The model did not consider additional variables which may improve the accuracy of predictions such as other weather-type variables (e.g. rainfall) or the impact of the different types of days weekdays vs weekends, school holidays, public holidays) to energy usage.

The following improvements may improve the model’s ability to predict energy demand:

* This report has not explored the use of a multiple regression model. From the analysis above, we know that the energy demand moves differently during the colder and hotter months and the inclusion of additional variables may improve the quality of prediction.
* Principal component analysis (PCA) may also be employed to find a new set of features that better capture the variability in energy demand.
  + 1. *Model 2 – Limitations & Improvements*

As noted above, Model 2 achieves a better predictive outcome in comparison to Model 1 but the accuracy scores achieved are below 0.5:

The following improvements may improve the model’s ability to predict the maximum price category:

* Although features with lower correlation have been filtered out, it is possible that the number of features can be narrowed down further using dimensionality reduction techniques to improve performance.
* PCA can be utilized to find a new set of features that improves classification accuracy.