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

**Executive Summary**

**Key Insights**

<TBA>

**Key Recommendations**

<TBA>

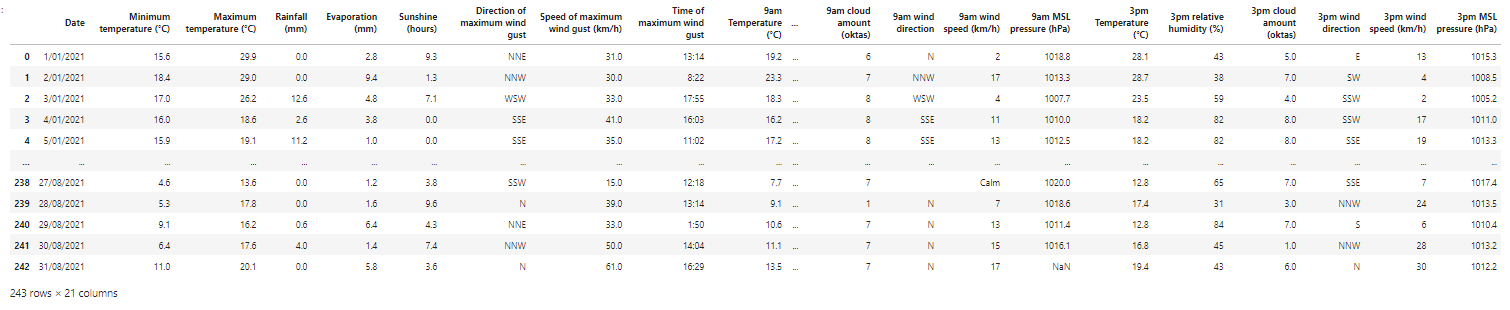
**Data Preparation**

**Datasets**

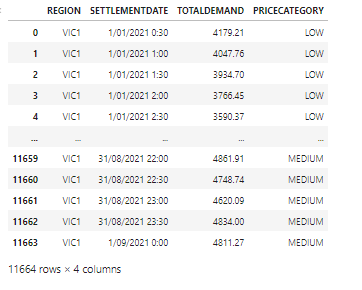
The following datasets were provided for the purpose of this exercise.

*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)*

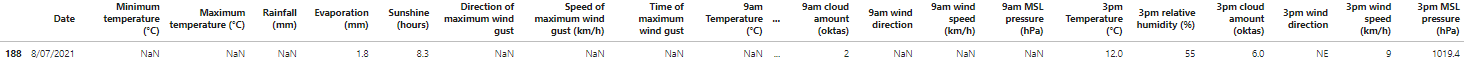
**Data Wrangling**

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).

*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 value for the *9 am wind direction* is missing on 23/01/2021 whereas the value for 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 the purpose of preparing Model 1, the median maximum temperature in °C across the weather dataset has been inputted to replace the missing value. Given that this missing data represent a small subset of the total data population (1 of 243 rows), it is unlikely that this substitution will have a material impact.



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

<TBA – insert information on data transformation on missing data for model 2, explain what method was undertaken>

*Inconsistent formatting*

The common denominator identified for both datasets is the date values. There is no other identified common denominator across the two datasets. However, the date format in the weather dataset is formatted as “dd/mm/yyyy” (see column *Date* - Figure 1) whereas the format in the price demand dataset is “dd/mm/yyyy hh:mm” (see column *Settlement Date* - Figure 2). To enable data aggregation, the date format in both datasets was converted so they are consistent.

*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. Therefore, a new variable containing the absolute value (i.e. negative values are converted into positive values) of this is added to the weather data. The reason why absolute value is used is that it is not expected that lower temperature will necessarily result in reduced energy consumption as energy consumption is expected to increase during colder weather.

**Data Aggregation**

*Aggregating daily energy demand*

For model 1, the Total Demand values for every half-hour are summed up so a daily total demand is obtained for each settlement date. This is to enable the model to predict total daily energy demand/usage (the dependent variable) based on the selected independent variable which is *the var to avg max temp*.

As a result of the date conversion method employed, there is a half-hour data for 01/09/2021 from 01/09/2021 00:00. This was removed as the data is incomplete (i.e. data is only for a single half-hour period) and there is no corresponding weather data. Given that there is still a population of 243 records remaining once this is removed, the removal would not have a material impact.

**Model Building**

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

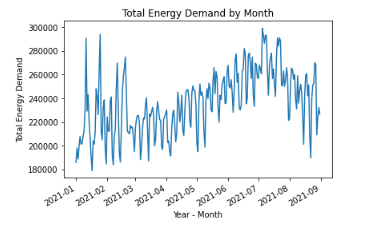
For Model 1, a simple linear regression model is employed to predict total energy use based on *var to avg max temp*. This method is selected as both variables contain numerical data and regression is a useful statistical tool to quantify the relationship between two variables.

This variable is selected because it produced the highest Pearson correlation coefficient of 0.3316 in comparison to existing variables within the existing weather dataset. This suggests that there is a moderate correlation between this variable and total energy demand.

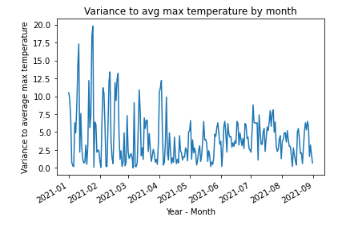
<TBA – insert Pearson coefficient figures from Zach>

*Figure 4 – Person correlation coefficient of variables in weather data against the total demand*

A visual inspection of line charts created on the dependent (Figure 5) and independent variables (Figure 6) 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).

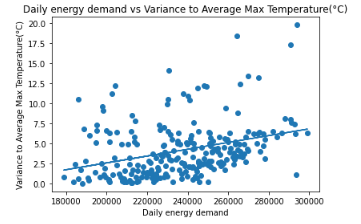


*Figure 5 – Line chart on daily total energy demand*



*Figure 6 – Line chart on daily variance to average maximum temperature*

A scatter plot created also provides a visual hint that there are potential causal relationships between the two variables.



*Figure 7 – 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.

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

<TBA – input information relating to classification model>

**Results**

**Assessment of Model Effectiveness**

*Model 1 – Effectiveness*

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

* The first test is to use the model to predict total energy demand on 16/01/2022. The results were as follows:

|  |  |
| --- | --- |
| Model prediction – total energy demand | 230,808.77 |
| Actual – total energy demand | 229,073.49 |

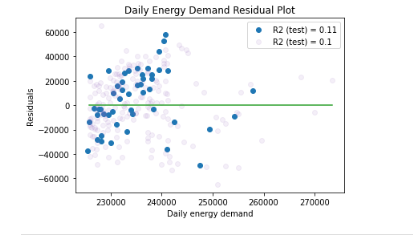
Although the result produced by the model appears to be pretty close to the actual, it is not efficient to manually test a few different dates using this method.

* The next test performed was to calculate the coefficient of determination (r2) values. The r2 score is the portion of the total variation in the dependent variable that is explained by variation in the independent variable. The r2 score was calculated on both the training and test dataset and below were the results:

|  |  |
| --- | --- |
| r2 -Training dataset | 0.1023 |
| r2 -Test dataset | 0.1144 |

The r2 scores suggest that the model is not effective in predicting the total energy demand as only 10.23% - 11.14% of the variation in these is explained by the variation of variance between maximum temperature against average maximum temperature.

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



*Figure 8 – Residual Analysis*

The residual analysis (refer to Figure 8) suggests that the residuals are independent. However, it can also be 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.

**Insights from Data Analysis**

**Limitations & Improvements**

Notes for others (will be deleted before report submission)

Unimelb technical report writing guide

<https://students.unimelb.edu.au/academic-skills/explore-our-resources/report-writing/technical-report-writing>

Example of unimelb technical report

<https://rest.neptune-prod.its.unimelb.edu.au/server/api/core/bitstreams/bc7abe33-87f1-5e12-8ef1-9b200912af3b/content>