**GROUP ONE**

Text

Description automatically generated with medium confidenceA picture containing train

Description automatically generatedA screen shot of a computer

Description automatically generated with low confidenceA picture containing indoor, automaton

Description automatically generated

[IMAGERY SOURCES]:

1 Photo by ARTHUR YAO on Unsplash

2 Photo by Artturi Jalli on Unsplash

3 Photo by Owen Beard on Unsplash

4 Photo by Stephen Dawson on Unsplash

**DATA ANALYTICS**

**{with Python} ==** *[DATA0006]*

DATA0006: Data Analytics with Python

Assignment 2

*07/08/2022*

**CO-AUTHORS:**

Liam Duffy  
Anqi Liao  
Ingrid Mayboehm

Dan Kenney

**GitHub ReadMe File:**

*https://github.com/Ann-chi/Assignment-2-COMP20008-Data-Science-Project/blob/main/README.md*

**GitHub Group One Assignment Two:**

[*https://github.com/Ann-chi/Assignment-2-COMP20008-Data-Science-Project*](https://github.com/Ann-chi/Assignment-2-COMP20008-Data-Science-Project)

Contents

[1. Introduction 5](#_Toc110797194)

[2. Data Exploration 6](#_Toc110797195)

[Scatter Plot assessment of results 7](#_Toc110797196)

[Pearson’s Correlation (Appendix) 7](#_Toc110797197)

[Data exploration analysis 7](#_Toc110797198)

[3. Predictive Modelling 9](#_Toc110797199)

[Modelling the change in energy consumption and price 9](#_Toc110797200)

[Methodology 9](#_Toc110797201)

[Model Results 10](#_Toc110797202)

[Model Equation 1 10](#_Toc110797203)

[Model Equation 2 11](#_Toc110797204)

[Model Equation 3 12](#_Toc110797205)

[Modelling the maximum daily price category 13](#_Toc110797206)

[Classification Tree 13](#_Toc110797207)

[Methodology 13](#_Toc110797208)

[Model Results (Appendix) 13](#_Toc110797209)

[4. K-Nearest Neighbours (KNN) 14](#_Toc110797213)

[Methodology 14](#_Toc110797214)

[Model Results 14](#_Toc110797215)

[5. Model Comparison and Selection 16](#_Toc110797223)

[6. Conclusion 17](#_Toc110797229)

[7. Limitations 18](#_Toc110797240)

[REFERENCES 19](#_Toc110797241)

[APPENDIX: 20](#_Toc110797249)

[CORRELATION SCORES 20](#_Toc110797250)

[CLASSIFICATION TREE 21](#_Toc110797251)

[KNN RESULTS 21](#_Toc110797252)

[HISTOGRAMS 22](#_Toc110797253)

[PAIR PLOTS 23](#_Toc110797254)

# Introduction

This data science project is based on the observations of energy consumption and the weather. Our team was tasked with exploring this issue and developing models to predict likely future energy demands based on weather forecasts. Two datasets were provided.

**Weather Data**

Supplied by the Bureau of Meteorology. This dataset contained key weather indicators for the City of Melbourne, covering a range of eight months from mid-summer to the end of winter.

**Price Demand Data**

Supplied by the Australian Energy Market Operator. This dataset contained energy prices and demand figures for the State of Victoria.

Both data sets were cleaned and merged by their date as index. This data set featured MAR data points which were filled in with the mean of column observations.

Our objective was to predict maximum daily energy usage and maximum daily price category, in accordance with the supplied weather data. This report outlines the steps used by our team to develop statistical models and concludes with an evaluation of their effectiveness.

# Data Exploration

A picture containing scatter chart

Description automatically generated

## Scatter Plot assessment of results

To search for any intuitive patterns and detect outliers, scatterplots were used to check the total demand for energy under different time periods of the day. We also checked the features concerning demand for energy and overall temperature. Finally, we used histograms to check the central tendency of the data.

## Pearson’s Correlation (Appendix)

To further quantify our analysis, a confusion matrix was constructed and represented by colour to measure the size and strength of all coefficients to identify potential linear relationships. The results of this table were expected to be helpful for use in our feature selection process.

## Data exploration analysis

This working dataset has a relatively even distribution of data points that centers around a mean. Observing the medium price category has the highest frequency count has provided insights into which price category is the most popular. Visually assessing the scatter plots, we have noticed linear patterns with a considerable number of data points outside the trend. These may form outliers in this model or in fact be representative of an underlying nonlinear relationship. According to the correlation matrix, the strongest coefficients all feature a negative relationship as energy demand increases, validating the scatterplot visual assessment. With further analysis, we sought to uncover a cumulative effect.

A picture containing text, scoreboard

Description automatically generated

# Predictive Modelling

## Modelling the change in energy consumption and price

To answer our first research question, we analysed the data types that would be used. From our initial assessment, there are mostly non-linear relationships however this may change when modelled in combination with other variables.

Our choice of model would be relative to the type of data we are using. Daily usage demand is a continuous variable, and this required an appropriate model. First, we tried linear regression where we constructed two models and compared the r2 values to decide which could provide a better goodness of fit for the daily demand of power usage in Victoria.

## Methodology

The data set was divided into two splits, training set and testing set. The training set is used to construct a predictive model by learning the central tendency according to 80% of the randomly assigned data. Consequently, the efficacy and predictive power of the former model, allows it to be tested against 20% of the data which has not been seen and results are evaluated for model selection.

We opted to used 80% of the data so the model had more opportunities to learn and reduce the error in prediction. Acknowledging that this introduces a lower degree of variance in favour of bias, we lacked a sufficient amount of data to rely firmly on the testing results. The degree of randomness included was also listed as 1 to improve the goodness of fit.

Feature selection was performed using two methods. Initially we used hypothesis testing with a confidence interval of 95%. Which means that below a P value of 0.05 we accept there is a dependent relationship. Our test for independence showed no significance in relationship to daily power usage and weather conditions, so we were unable to reject the null hypothesis and conclude that these variables are independent. We then conducted a separate test on price categories and found there is no significant relationship between each price category alone and consumer demand for which energy was explained in 95% of the data.

Our feature selection method did not hold all independent relationships constant, so we opted to try the **wrapper** method to build a model that was both explainable and could comparatively explain as much variance as possible. Over three iterations R2 and MSE were evaluated.

## Model Results

**Our first model** equation excluded wind variables in exchange for time of maximum gust. Using the wrapper method, we constructed a model that achieved an R2 of 0.34. However, the Y intercept produced an average demand of 15243 which was deemed spurious because it exceeded the maximum daily usage identified in the data.

*Chart, scatter chart

Description automatically generated*

### Model Equation 1

*Demand = 15243.1 + -122.46\*MinTemp + 154.85\*MaxTemp -4.42\*Rainfall + 19.36\*Evaporation -63.67\*Sunshine -0.47\*max\_gust -67.87\*am\_cloud + 58.75\*msl\_pres -133.21\*pm\_temp + 2.53\*pm\_relative\_hum + -3.85\*pm\_cloud -66.19\*pm\_msl\_pres + e*

**Our second model** was a less dimensional version of the first, which managed to achieve a slightly less R2 value at 0.29, using the same tuning parameters and featuring less variables. With an average demand of 7153.99 at base 0, it was far more reasonable. A visual assessment of the model observations shows it is more likely to under predict the correct daily energy consumption.

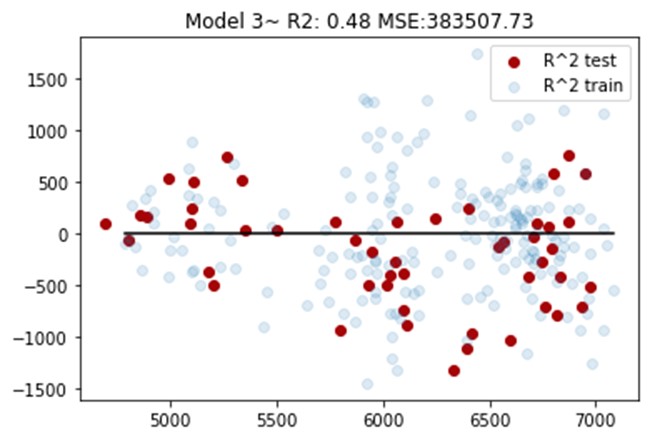
Chart, scatter chart

Description automatically generated

### Model Equation 2

*Demand = 7153.99 + -204.85\*Min\_Temp\_2 + 106.65\*Temp\_9am\_2 -4.10\*Temp\_3pm\_2 + 21.23\*Evap\_2 + 29.86\*Sun\_2 + e*

**Our third model** is built on our second model. Here we see an increase in R2 value to 0.48 and similar MSE. Showing that these price categories offer greater predictive power when forecasting demand. A visual analysis shows that the plots appear to follow the training dataset at the minimum and maximum ends of the regression line better than the previous two models.



### Model Equation 3

*Demand = 5220.713 + -9.81\*Min\_Temp\_dummy + 5.66\*Temp\_9am\_dummy + 1.96\*Temp\_3pm\_dummy + 6.07\*Evap\_dummy -2.78\*Sun\_dummy, + 8.05\*price\_2\_dummy + 1.41\*price\_3\_dummy + 1.63\*price\_4\_dummy*

## Modelling the maximum daily price category

### Classification Tree

Taking the results from our regression into consideration, a classification method was required to predict the most likely daily price category and see which features would have the highest level of information gain. It is expected that the highest ranked features will present at the base of the tree with the least amount of bins and reduce, as leaf nodes are formed.

## Methodology

Prior to running the algorithm, binning was performed to assign class labels to continuous variables in a multi-way split. This is a technique required for the classification tree to work effectively. Using a program, features were ranked according to highest information gain. This provided a baseline understanding of what we could reasonably expect. We then opted to use all features excluding ‘Date’ and let the algorithm tell us what is important.

We were looking for the highest accuracy score with the lowest number of bins, to help determine important features. Noting that accuracy score is not always the best method of judging a model’s effectiveness, the negative impact of models incorrectly guessing is relatively low and should not be relied upon when limited data is provided to build a model.

## Model Results (Appendix)

This model showed that Maximum temperature had the largest impact on price category with the strongest level of information gain, followed by minimum temperature, 3pm temperature, 9am wind speed and evaporation. This tree was able to correctly guess the outcome of the daily price category with an accuracy score of 0.63.



# K-Nearest Neighbours (KNN)

As a comparison, we decided to compare with another model using cross validation. For this we chose a KNN supervised learning algorithm that clusters features together according to the lowest dissimilarity calculated by the Euclidean distance. As this is a supervised machine learning algorithm it required assistance in forming medoids, for which we selected four.

## Methodology

For this model we opted to use K fold cross validation to allow the model to train over multiple iterations. The data set was stratified into five partitions with the 5th being the test set and each fold would be used to allow the algorithm to build a predictive model with high accuracy. Completing this three times produced the best overall aggregate result.

The data was first scaled to stop large clusters overpowering the results, any missing variables were imputed with the column mean and outliers were removed prior to this.

## Model Results

Cleaning the results from the KNN produced a high accuracy score of 0.67 with an expected average accuracy of just 0.59. The maximum result is an improvement over the classification tree while the average expected result is below. The risk appetite will ultimately decide which we consider useful.

**Text, whiteboard

Description automatically generated**

*Accuracy score: TP+TN / TP + FN + TN + FP*

According to the confusion matrix, this model was better at classifying medium and extreme price categories.

**Graphical user interface

Description automatically generated**



# Model Comparison and Selection

Our initial linear regression was not able to produce a valuable model on weather data alone without the inclusion of price category variables, leading us to believe that perhaps the greatest indication of energy demand is in relation to price category.

To model question one, we selected model three to provide the greatest predictive power and comparative error rating:

*Demand = 5220.712 + -9.81\*Min\_Temp\_dummy + 5.66\*Temp\_9am\_dummy + 1.96\*Temp\_3pm\_dummy + 6.07\*Evap\_dummy -2.78\*Sun\_dummy + 8.05\*price\_2\_dummy + 1.41\*price\_3\_dummy + 1.63\*price\_4\_dummy +E 355581*

On average we expect power demand to be 5220 at a low-price category. Holding other variables constant, the medium price category was overall weighted the highest in energy demand and would increase 8.05 times per 1 point of energy demand, making it the most likely category to influence price demand. Comparing to our correlation analysis, Evaporation and Sunshine did not have strong correlations in our data exploration but became useful in building our model. In comparison to other features, evaporation featured a relatively high coefficient. Perhaps it is worth considering that there are significant outliers present in this data set and removing them would produce a better model.

Intuitively min temperature was relatively high in this model and negatively correlated, as minimum temperature declines, we can reasonably assume that maximum temperature increases.

Comparing this to our classification results, to answer our second question, we compared two models using a decision tree and KNN. Using accuracy as our preferred metric we believe the KNN model is a better option which could predict medium and extreme price categories. Providing further validation to our linear regression that medium price category is the most important label and out feature selection was salient. The selected weather factors can be used as a leading indicator, but ultimately our results show the cost of energy will influence demand.



# Conclusion



To conclude, we selected our project modelling on the premise that it would address both questions with the lowest degree of error possible within the available dataset.

With careful consideration given to overfitting, the degree to which this model can be explained has been an essential focus so that real world value can be created.



# Limitations

We have observed there are noticeable limitations inherent within this modelling. These include the following observations:

* a limited amount of the data that does not cover a full year or insights from previous years
* demand and costs can be hard to measure without a benchmark figure to help understand whether a positive or negative outcome has been produced, thus influencing how well we can explain our analyses
* a sample size of data from previous years would allow us to construct a seasonality index and forecast the change in quarterly demand, thus enabling a more accurate model to be built

# REFERENCES

Data Science Tutorials. (2021, Nov 13). ‘*Installing Matplotlib learn in Anaconda Jupyter Notebook.’*  
<https://www.youtube.com/watch?v=Xmr-1L04mvQ>

Fessel, K. (2020, Nov 17). *‘How to change the axes limits of a matplotlib figure || Matplotlib Tips.’*   
<https://www.youtube.com/watch?v=tPUqnGjhwEg>

Fessel, K. (2021, Jan 12). *'Seaborn FacetGrid | How to make Small Multiples with Python Seaborn | Titles, Hue, Legend.’* <https://www.youtube.com/watch?v=YYeqJllXHxM>

Fessel, K. (2020, Dec 8). *‘Seaborn histplot | How to make a Seaborn histogram plot with Python code.’* <https://www.youtube.com/watch?v=Bjz00ygERxY>

Fessel, K. (2020, Aug 1). *‘Seaborn scatter plot | How to make and style a scatterplot in Python seaborn.’* <https://www.youtube.com/watch?v=4yz4cMXCkuw>

Pawandeep. (2021, 11 March). Tutorials Point Simply Easy Learning. *‘How to install matplotlib in Python?’* <https://www.tutorialspoint.com/how-to-install-matplotlib-in-python>

Python Tutorial. *‘Read CSV with Pandas.’* <https://pythonbasics.org/read-csv-with-pandas/>

Stack Overflow.(2013, May 29). *‘Is it possible to print using different colors in ipython's Notebook?’* <https://stackoverflow.com/questions/16816013/is-it-possible-to-print-using-different-colors-in-ipythons-notebook>

Waskom, M. Seaborn. *'Seaborn.color\_palette.'* <https://seaborn.pydata.org/tutorial/color_palettes.html>



# APPENDIX:

## CORRELATION SCORES

Between various dataset features

TOTALDEMAND 1.000000

PRICECATEGORY 0.685448

9am wind speed (km/h) 0.117655

9am relative humidity (%) 0.103267

Speed of maximum wind gust (km/h) 0.081024

3pm cloud amount (oktas) 0.070940

3pm relative humidity (%) 0.064301

9am MSL pressure (hPa) 0.051994

3pm MSL pressure (hPa) -0.005067

3pm wind speed (km/h) -0.040821

Rainfall (mm) -0.072715

Direction of maximum wind gust -0.076616

Sunshine (hours) -0.139581

9am cloud amount (oktas) -0.167373

3pm wind direction -0.168290

Time of maximum wind gust -0.178466

Evaporation (mm) -0.264008

Maximum temperature (°C) -0.290004

3pm Temperature (°C) -0.325252

9am wind direction -0.337905

9am Temperature (°C) -0.390843

Minimum temperature (°C) -0.488244

## CLASSIFICATION TREE

Graphical user interface, application, Teams

Description automatically generated

## KNN RESULTS

Best Score: 0.6714035087719298

Parameters: {'n\_neighbors': 26}

Average score: 0.598650540492646

Average std: 0.09487809603307505

Metric Used: euclidean

## HISTOGRAMS

Plot different variables and their count value

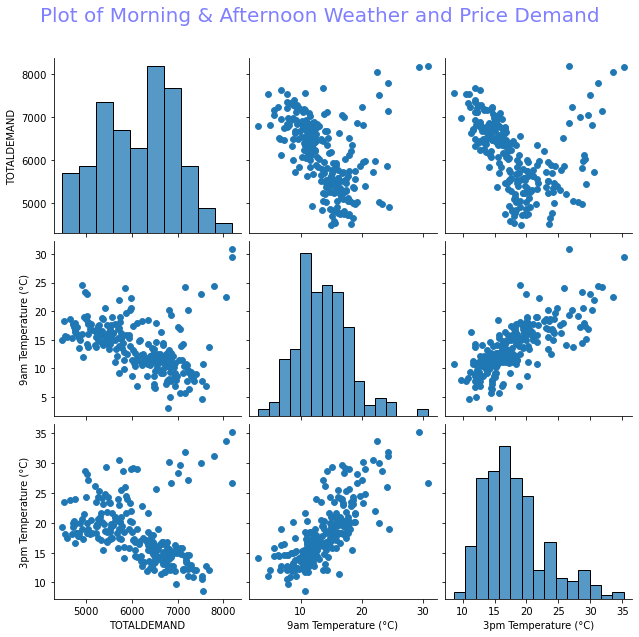
**Chart, histogram

Description automatically generated**

## 

## PAIR PLOTS

Plot different variables on x- and y- axes for comparative analysis

****

Chart, histogram

Description automatically generated