A picture containing sky, grass, outdoor, field

Description automatically generated

Daily energy usage and price forecasting based on the weather data

Teams member : Foad Momeni , Wenjie Yuan, Nasim Nemati Stephanie Xie

[A2-Project-6](https://github.com/foadmomeni/A2-Project-6) | [0DATA0006\_2022\_SEP\_PAR\_1](https://canvas.lms.unimelb.edu.au/courses/167755) | Dec 2022

# Exploration and Cleaning

In those two datasets provided, we have two .csv files: 'price\_demand\_data.csv' and 'weather\_data.csv', which contain key weather indicators in Melbourne and energy price and demand figures for Victoria between January and August 2021. We noted there is geographic difference between those two datasets as the energy demand and price information is at the state level while the weather information only relates to Melbourne. However, Melbourne contributes to most of the energy usage, since Melbourne is the largest city in Victoria and in fact comprise around 75% of the total population of Victoria.

## Price and demand dataset

Our first step is to remove the columns which will not be used in our analysis. Each record in the ‘price\_demand\_data’ dataset is for half hour period. Therefore, we need to summarize data as we don’t need to know the demands for different hours in a day. So, we drop all the hours from the SETTLEMENT column. We have also checked values in the REGION column are all ‘VIC1’, so we dropped this column.

To use the Date column as index in both datasets to merge them together later, we have renamed SETTLEMENTDATE to Date. Besides, to find maximum daily price category, we replace 'LOW' to 1 , 'MEDIUM' to 2, 'HIGH' to 3 and 'EXTREME' to 4 to find maximum price category, then we will replace to original. We then grouped the dataset by date and aggregate by max demand and max price category. The purpose for our first model is to predict the total daily energy usage but we think that for total demand we should consider maximum demand not adding all the demand during every day, so we have taken only maximum demand each day.

## Weather features dataset

Although we noted there are Minimum temperature and Maximum temperature in the dataset, we also calculated the Average temperature and Range per day and add to the dataset which might be help us in our model. We also put Date as index in both DataFrame so we can merge them together later. Checking inconsistency of data and correct them as well as changing data type have been done in our DataFrame.

## Merging the two datasets

After we merged two datasets, Price dataset has 244 rows and weather dataset has 243 when we merge inner, we delete one row from price which is data for 1/9/20.

We have also done further analysis to drop all the columns that are constituted by NaNs, as they are unusable. The result shows that there are 23 NaNs in our DataFrame located in 6 rows.

we could replace NaNs with mean or find similar rows and find the best match but 6 of 243 looks acceptable to discard.

After the above steps, the datasets have been cleaned and merged successfully and are ready for further use as input into our model.

All these wrangling and cleaning data is based on our needs for our model, we did back and forth many times to do wrangling and cleaning. As our model is predicting maximum demand based on daily weather data, dropping hours was the best way and replace maximum demand for each day instead of demand for every half an is what we need for first model. By replacing different price category to numbers temporary, we could find maximum price category per day.

# Forecasting Total Daily Energy Usage Based on Weather Data

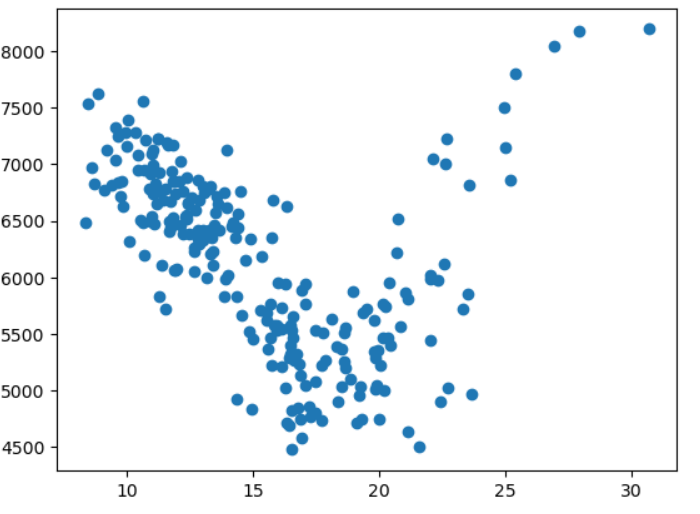
Before start to write any code we need to know what do we want to do? Demand and Usage are two different things, Demand is like the speed of a car during driving and usage is distance traveled, adding all demands is not right way, it is like adding all different speed, same as car which need to be able to drive maximum speed, power plant need to be able to produce maximum demand. Thus, we decided use maximum demand.

Next step we decided to predict which variables should have correlation with maximum demand, asking two questions:

* Which variables should have correlation with maximum demand? The answers were temperature and sunshine
* what do you think our scatter plot should be based on previous question? For temperature should increase as temperature going to minimum or maximum and less sunshine should lead to increase demand.

The first model is to predict the total daily energy usage based on the provided weather data.

For all columns with continues data (float and integer) we use scatter plot and Pearson correlation to identify the protentional independent variables that could be used in our model, including Temperature, Rainfall, Evaporation, Sunshine (hours) and etc. As we investigate the scatter graphs, we find out there is relation between temperature and demand. However, as we are cautious about the fluctuation of the temperature during a day, we decided to use Average temperature for our model.



As seen in the scatter graph between Average temperature (°C) and Max\_Demand above, the curve is a kind of parabola. We can define a temperature where is about minimum demand and have two model. When we look at scatter it looks 19 degree is a proper point, so we will have two model for temperature in range (8,19]) and (19,31). Quadratic regression should be a better solution and we surf through internet, it looks there are some functions achieve this, but we need to get familiar with it.

We firstly look at data with Average temperature (°C)' less than 19. The Pearson r is -0.83 which is strongly linear, as shown in the graph below.

Chart, scatter chart

Description automatically generated

Before knowing about cross validation, we realized using just one random state cannot certain us about our model. To make our model reliable we used 100 different random state by using a for loop and then make an average for gradient and intercept for our model as well as evaluate our model based on average of r squared for training and test sets. The results from running the code are detailed below:

* r square for training set = 0.686803560149947
* r square for testing set = 0.6749955637014509
* our model is Max\_Demand = [-226.05981583] x Average temperature (°C) + 9277.050466450404

It looks we got a good r square. Our model will be y = - 226.06 x 'Average temperature ' +9277.05 but this model is for x in range (8,19). When we have just one independent variable probably doesn’t need to scale our variable.

Next step we did Residual study, again because we finish this part before knowing the code, we make a DataFrame for predictions and calculate difference and this is the result:

Blue colors are predictions diffrence with actual, orange is other data. It looks residual for prediction are the same.

Chart, scatter chart

Description automatically generated

We have then build the model for temperature in range [19,31). Similarly, we have a Pearson r of 0.79 which is strongly linear.

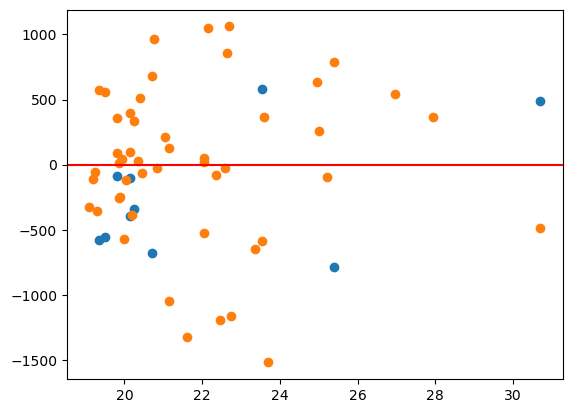
Chart, scatter chart

Description automatically generated

Next, same as before we use a for loop to check our model performance and finding more accurate model. The results are detailed below:

* r square for training set = 0.6271911058637015
* r square for testing set = 0.43857115005458447
* our model is Max\_Demand = [314.42228657] x Average temperature (°C) -967.9138771227513

this is residual for second model, it is clear that we don’t have enough data for second model.

* 

Our second model looks ok but not as good as model for temperature in range (8,19). One potential reason is we don't have enough data for temperature in range [19,31). Our model will be y = 314.42 x'Average temperature' – 967.91 but this model is for x in range [19,31).

In summary, we have split the model according to the average temperate within two ranges:

* When the average temperature is in range (8,19), the model is:

Max\_Demand = -226.06 x Average temperature (°C) + 9277.05

* When the average temperature is range [19,31), the model is:

Max\_Demand = 314.42 x Average temperature (°C) - 967.91

We cannot use these models for any other temperature.

There are a few limitations of our model and ways to improve it:

1. As mentioned at the beginning of the report, there is an issue in our data which is important: the weather dataset is for Melbourne, but demand is based on Victoria. Both datasets should cover the same geographic region otherwise there is potential detriment to the accuracy of our model.
2. The time frame of the dataset is only 9 months between January and August. The model can be improved if we have at least a full year of data.
3. To improve our model, it is better to consider holidays and weekends as lots of business not working in these dates.

# Forecasting Maximum Daily Price Category Based on Weather Data

The second model is to predict the maximum daily price category based on the provided weather data.