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

GitHub: https://github.com/foadmomeni/A2-Project-6.git

[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

We have two .csv files 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.

* Data wrangling and aggregation methods

## Price and demand dataset

Our first step is to remove the useless data. Each record in the ‘price\_demand’ file 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. Furthermore, we need daily basis data so we can merge ‘price\_demand ‘and ‘weather’ later. We start by dropping hours from the SETTLEMENT column. The value of REGION column is all ‘VIC1’ so we dropped it.

To use the Date column as index to merge datasets later, we 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, then we replace to original afterwards. We 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 of each day.

## Weather features dataset

Although we noted there are Minimum and Maximum temperature in the dataset, we also calculated the Average temperature and Range per day and add to the dataset which might be helpful in our model. We also set Date as index in both DataFrame to 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 merged two datasets, We have also done further analysis to drop unusable columns that are constituted by zeroes and NaNs. The result shows that there are 23 NaNs in our DataFrame located in 6 rows has NaN we could replace NaN with mean or find similar rows and find the best match but 6 of 243 looks acceptable to discard.

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 hour is what we need for our models. By replacing different price category to numbers temporary, we could find maximum price category per day.

Now the datasets have been cleaned and merged for further analysis.

# Forecasting Total Daily Energy Usage Based on Weather Data

* Model building

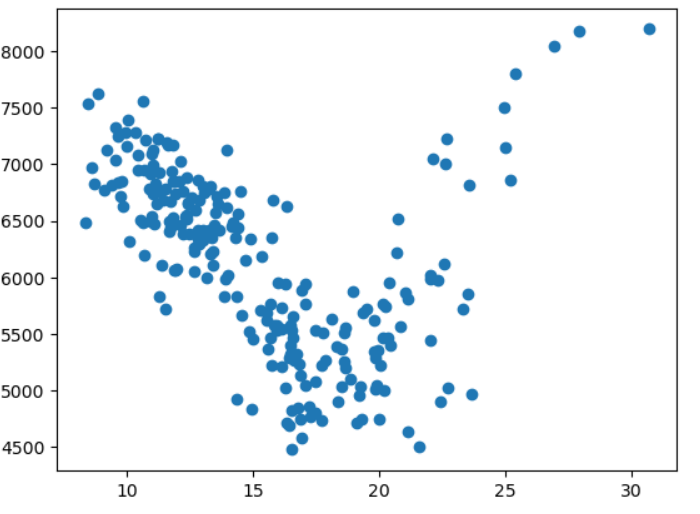
Before start to write any code we need to know what do we want to do? Demand and Usage are two different things, Demand is like the speed of a car during driving and usage is distance traveled, adding all demands is not right way, it is like adding all different speed, same as car which need to be able 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 without writing any code or even checking our data just our feeling, asking two questions:

* Which variables should have correlation with maximum demand? our 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) 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. The red line is our model:

Chart, scatter chart

Description automatically generated

* Model evaluation

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 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

Chart, scatter chart

Description automatically generated

We have then built 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.

Chart, scatter chart

Description automatically generated

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 in 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.

* Insights from analysis

From our model, we can tell the power demand have a linear relation within specific range of temperature, and we can use it for total daily demand prediction.

* Why the result significant and valuable

About 70 percent of errors is covered by our model but it is hard to answer this question as this is our first model. If we already had experience, we could say 70 percent is good or not, just based on our feeling, valuable of our model depend of purpose of study, if this study is for investment for power plant 30 percent error is too high, it means that government might invest in power plant more than what we need or built power plants which can’t cover the demand.

Our model is logic and easy to understand, demand is related to temperature.

* Limitation of result and future improvement

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.
4. It looks we got a good r square. However, if this model wants to be used for investing in power stations need to study more.
5. Our model is just for average temperature between (8,31) and for sure there are some days which average temperatures are below or above.

# 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.

* Data wrangling and aggregation methods

However, we did wrangling and cleaning at the start but still need more for second model, First we split our DataFrame to numerical variable and categorical variable, the reason is that most of variable are numeric and in our analysis we used just numeric variables, then because we want to study price category based on weather so demand shouldn’t be as a variable, it is logic as well because demand and price categories are simultaneous and price category should have relation with demand but wholesaler can’t wait to see demand for today and define price category for past.

* Model building, evaluation and Insights from analysis

We group by Price categories and check the boxplot for each variable.

Chart, box and whisker chart

Description automatically generated

Chart, scatter chart

Description automatically generated

In first boxplot we see there is relation between price category and average temperature and in second we see rainfall looks same for different price categories.

As we wrote our code before learning chi2 and feature selection, we used ANOVA, to find best features for our model. We decided to use 5 top features based on p values.

In ANOVA same as other statistic hypotheses, we have two hypotheses:

H0: there is no difference in our variable (mean) between the price categories

H1: there is difference in our variable (mean) between the price categories

Using a for loop and find 5 best features with lowest p value by considering choosing variables which doesn’t have correlation.

|  |  |
| --- | --- |
| features | P values |
| Average temperature (°C) | 3.6013854502644483e-22 |
| Evaporation (mm) | 7.302455004359034e-14 |
| 9am relative humidity (%) | 3.958740219051292e-06 |
| Sunshine (hours) | 7.636002610266395e-06 |
| Speed of maximum wind gust(km/h) | 5.086061547900228e-05 |

we think that for ANOVA we should consider some assumptions, this is just our thinking and honestly need more time to deeply understand behind the scenes of ANOVA but if we are checking the mean of two variable are different or not, these two variables should have probably same distribution.

We used k\_ nearest neighbor algorithm and to make our model reliable we used 100 different random state by using a for loop and then make an average for accuracy, overall accuracy is about 48 percent, we think this is not good enough, if we didn't have any model just by probability, we can predict the price category 25 percent correct and by our model 25 percent increased to 48 percent, so sad!

We used best feature selection method based on chi squared and surprisingly features were exactly same as ANOVA so output of our model won’t change, still about 48 percent.

Finally, we used principal component analysis with cross validation and k\_nearest neighbor and Decision tree and accuracy of our knn model increased just by 2 percent and accuracy of decision tree drop to 42 percent.

* Why the result significant and valuable and limitations of results, improvement for future

We tried different algorithms to achieve a better result, but 50 percent accuracy doesn’t look good enough, it looks there is another variable which we don’t have it in our model, we checked categorical variable and doesn’t mean wind direction has effect on price categories.

Price categories should be based on demand and that’s why some wholesalers offer different price for day and night, we didn’t consider price categories and time per day.

In our DataFrame there are some high demands with low price category and vice versa.

The question is if price categories can be predicted just by weather?

* Conclusion:

In our first model we realized there is a relationship between temperature and demand which is logical, when temperature goes down heaters turn on, in for example 20 degree rarely people using heater or air conditioning system. Our model can explain about 70 percent of errors, our model is logic and understandable.

In second model about 50 percent of predictions are correct. If we didn’t make our model 25 percent of prediction should be correct (one out of four), which means our model unfortunately is not applicable. In future we should check our data if there is any out of range, however it doesn’t show any out of range in boxplot. Some data are important for our model for example, price of oil or day and night hours.