Assignment 2: Data Science Project

[Data Analytics with Python (0DATA0006\_2022\_MAY\_PAR\_1)](https://canvas.lms.unimelb.edu.au/courses/159930)

GitHub: https://github.com/smbayat11/Assignment2\_Group9

Group members:

* Sidhra Jani (sidhrajani@yahoo.com)
* Somayyeh Bayat (smbayat@gmail.com)
* Ashraful Alam ([alam.ashraful31@gmail.com](mailto:alam.ashraful31@gmail.com))
* Ming Him Tong ([leo121713@hotmail.com](mailto:leo121713@hotmail.com))

Introduction

Energy is an essential need for everyone. The sources of energy production have emerged in various types. Although the primary energy consumption tends to depend on temperature, other factors affect energy production and consumption. For example, solar energy can be produced when the sunshine increases and the sunshine may unevenly heat the earth's atmosphere by solar radiation, increasing the temperature. In the meantime, the cloud amount will significantly impact sunshine reaching the earth's surface. Wind can produce energy, but this has affected by topography.

This analysis is based on how the daily maximum energy demand is affected by temperature and the maximum daily price category according to the weather data for the given dates. We built the model to predict and forecast the future energy demand and price category.

Assessment

1. What wrangling and aggregation methods have you applied? Why have you chosen

these methods over other alternatives?

All the missing values were dropped to clean the data set. OpenRefine tool was used to identify the missing values. Twenty-five rows of data were dropped from the 'weather\_data' table. The cells that contain missing data were not uniformly distributed in the table. When we checked raw-wise, there were many empty cells, so we removed the values rather than imputing them.

The columns' Direction of maximum wind gust', '9 am wind direction', and '3 pm wind direction' column data were replaced. The eight directions were annotated with one or two letters to make them uniform for analysis(aggregation).

One of the important columns in our analysis is the Settlement date column in the 'price\_demand\_data' in object format. We changed it into the DateTime format to ease the study by splitting the date and times into two separate columns.

2. How have you gone about building your models, and how do your models work?

2-1- A model which predicts the maximum daily energy usage based on the provided weather data:

Based on the data provided, we built a linear regression model. We merged both CSV files to create one table. Applied the Groupby method on the date column, the maximum demand for the day (MAXDEMAND) was considered to build the model.

We utilized Correlation Map and Scatter Plot to explain the strength of the relationship for each factor. They revealed that the most affecting factor to maximum demand is temperature. As shown in figure 1, the relationship between maximum Direction and minimum temperature, maximum temperature, and average temperature are -0.47, -0.26, and -0.37, respectively.

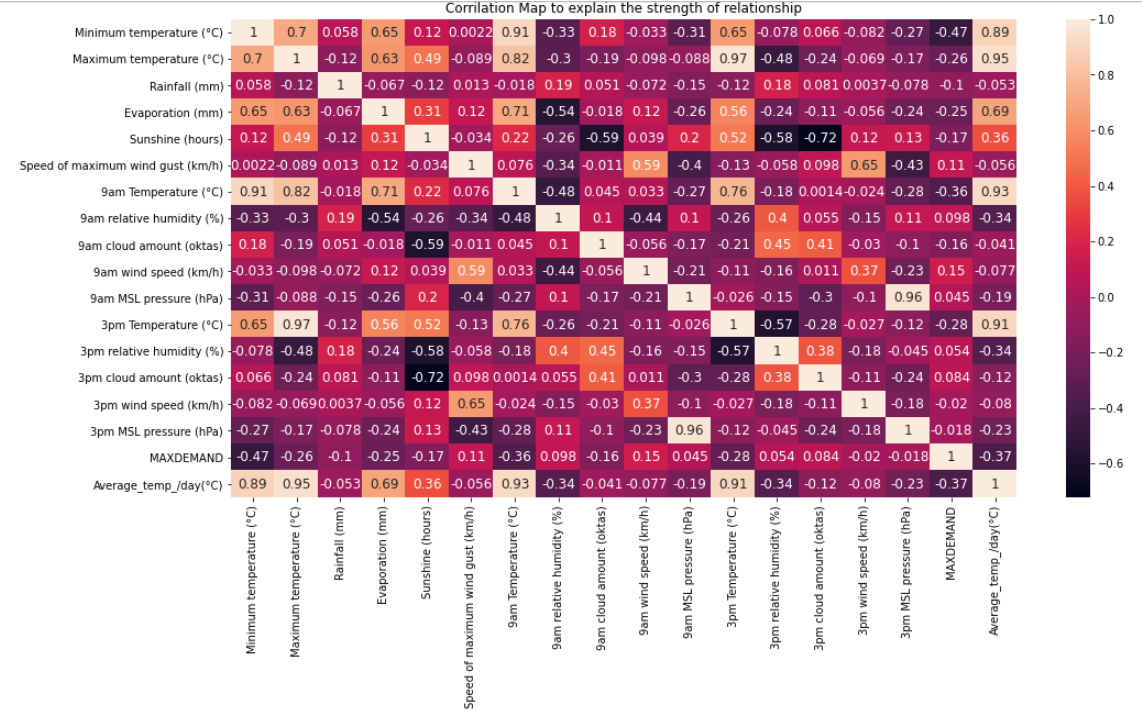


Figure 1: the correlation values of all features

Having the Correlation Map, we assumed that minimum temperature must give us a robust regression model. Thus, we used Linear Regression Method to build the model.

It gave us the following coefficients, which are poor:

* Coefficient of determination (test): 0.10
* Coefficient of determination (training): 0.03

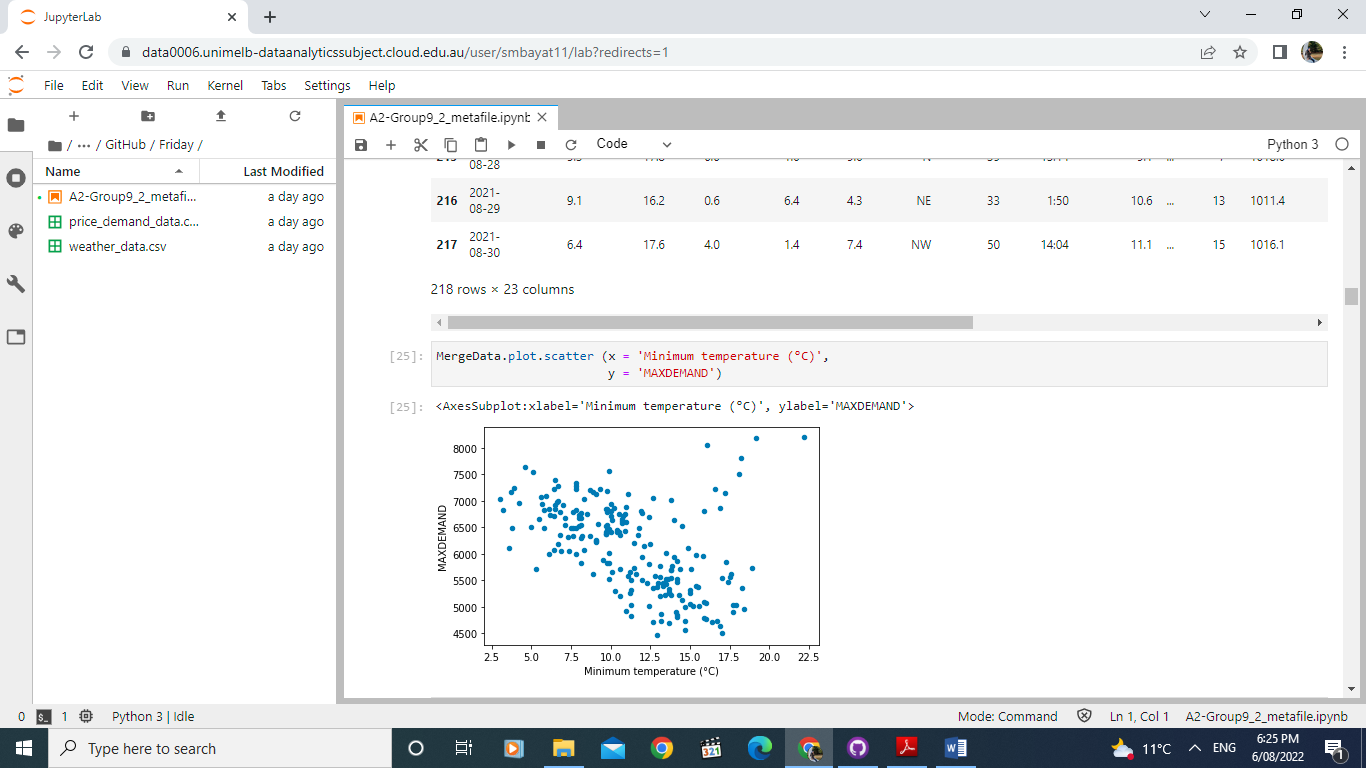


Figure 2: Maximum Demand versus minimum temperature

We also tried other temperature variables. We analyzed the maximum temperature, average temperature, and 3 pm temperature (as in summer, the temperature goes high around 3 pm). The pattern of the scatter plot is similar. However, the highest demand is the same as the lowest market slightly changes, but not significantly.

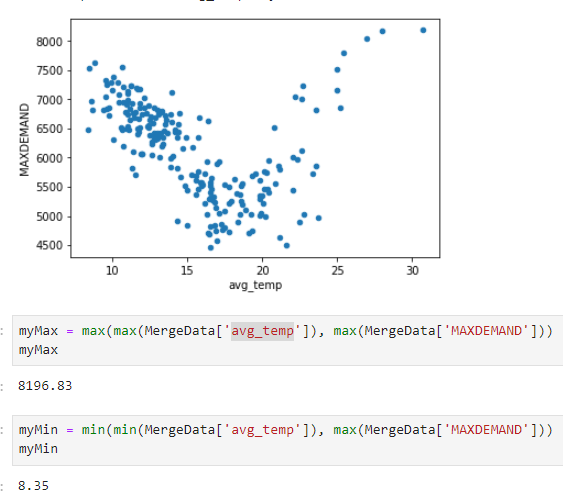
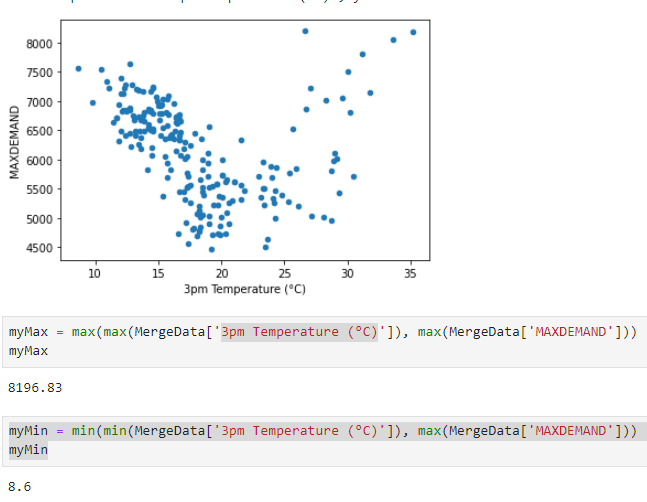
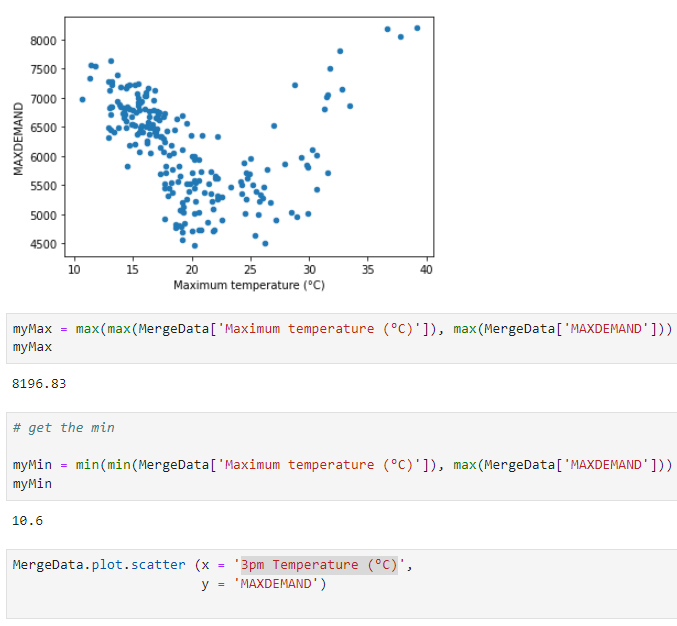


Figure 3: Maximum Demand versus maximum, average, and 3 pm temperature

Although Correlation Map proved that maximum demand has a strong relationship with temperature, the results above didn't lead to an appropriate regression model. The graphs were not linear, so we analyzed deep into them. All temperature-related patterns were U-shaped overall.

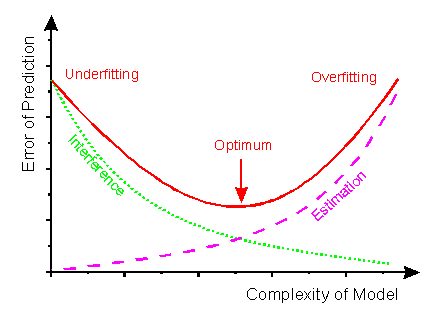


Figure 4: Model complexity (<http://www.frank-dieterle.de/phd/2_8_1.html>)

According to the model complexity diagram above, the optimum temperature was selected.

Based on the scatter plot, Maximum Demand is minimum when the Average daily temperature is 200C

It is in line with the fact that 200C is considered ideal room temperature.

Thus, we defined a new variable as below:

‘Diff\_temp\_/day(°C)’ = abs(20 - ‘ Average\_temp\_/day ’)

This variable gives us the absolute difference between the Average daily temperature and 200C.

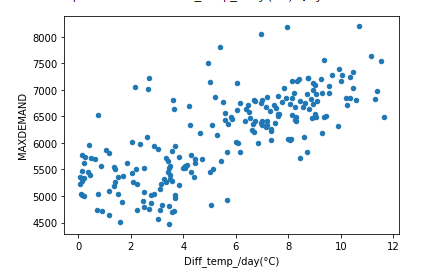


Figure 5: Maximum Demand versus Diff\_temp\_/day(°C)

Figure 5 indicates a linear pattern of Maximum Demand against Diff\_temp\_/day(°C). The correlation Map showed a strong relationship between Maximum Demand and Diff\_temp\_/day(°C), which is 0.74.

These results convinced us to use Diff\_temp\_/day(°C) for Linear Regression. The output of the Linear Regression Method proved that the model is reliable.

* Coefficient of determination (test): 0.56
* Coefficient of determination (training): 0.55

Then we were interested in analyzing how the other factors might affect the energy demand, such as sunshine, wind gust, evaporation, and cloud amount. The scatter plots were plotted against each given data, and correlation values were calculated.

Timeline, box and whisker chart

Description automatically generated with medium confidence

Figure 6: the correlation values of all features

The chart explains all the relationships, and the figure 6 correlation values define the relationship and their strength.

Sunshine:

The number of clouds will impact the sunshine. When the sunshine increases, the temperature increases. But when the sunshine increases, it produces solar power, affecting energy demand. So we cannot expect the same pattern on the sunshine-max demand plot as the temperature- max order scatters plot.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Figure 7: Impact of Sunshine on Temperature and Maximum demand

When the sunshine increases, we cannot observe a pattern on max demand as many other factors affect the temperature.

2-2- A model which predicts the maximum daily price category based on the provided weather data:

According to the price data, the energy price was categorized into four groups (Fig 8). First, we need to calculate the maximum daily price category, so we rated all category labels to the number (LOW=1, MEDIUM=2, HIGH=3, and EXTREME=4).

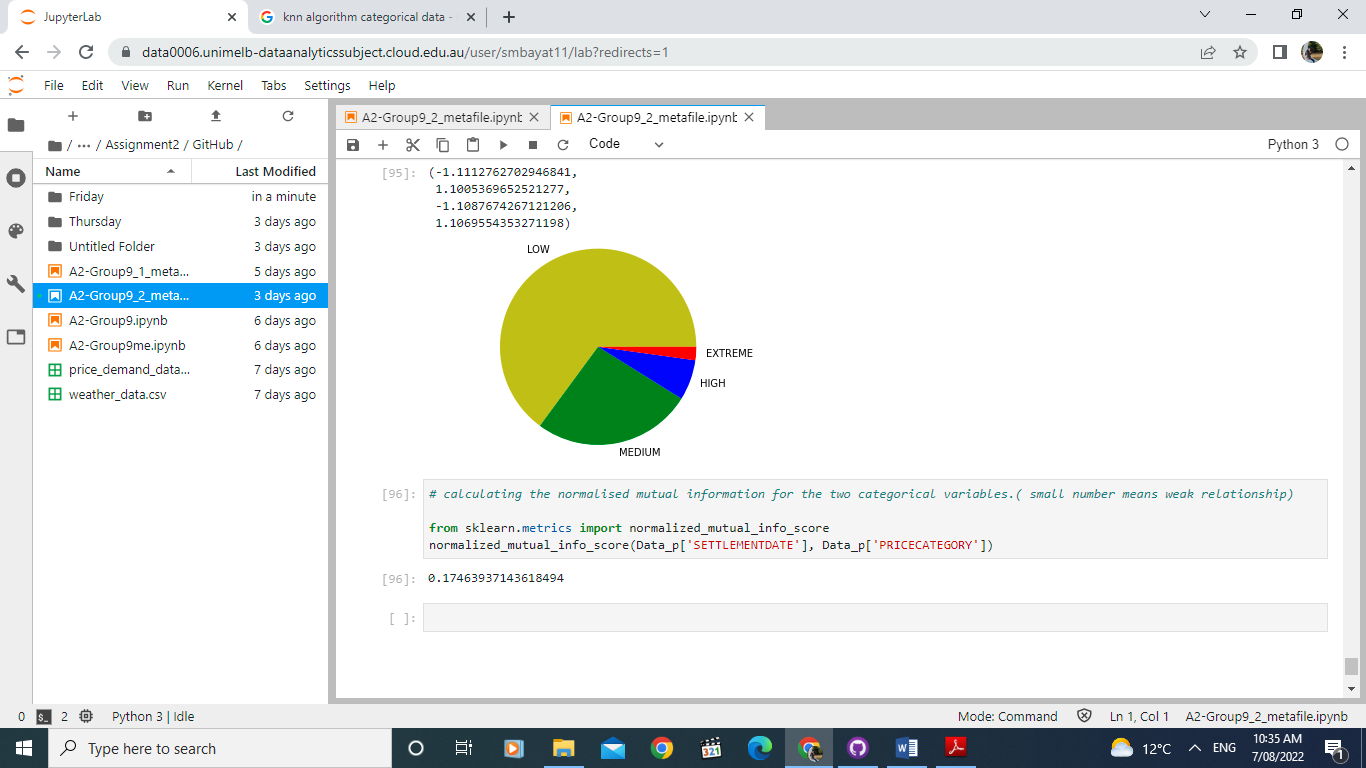


Figure 7: The price category's pie chart

Then after finding the maximum daily price category, we turned all rates to the label again. Finally we merged both weather and price csv files to create one table. Applied the Groupby method on the date column.

For this part, we used KNN algorithm to find the best model. We defined

'maximum\_daily\_price\_category' as a class label and weather data as a feature. We compared our prediction with the actual class label. The overall accuracy was 0.57 for train\_size=0.66 and test\_size=0.34.

We also applied Equal length bin technique for price category discretisation (Fig 8):

* [min, 4080.605), [4080.605, 5452.68), [5452.68, 6824.755) [6824.755, max]
* min = 2708.53
* max = 8196.83

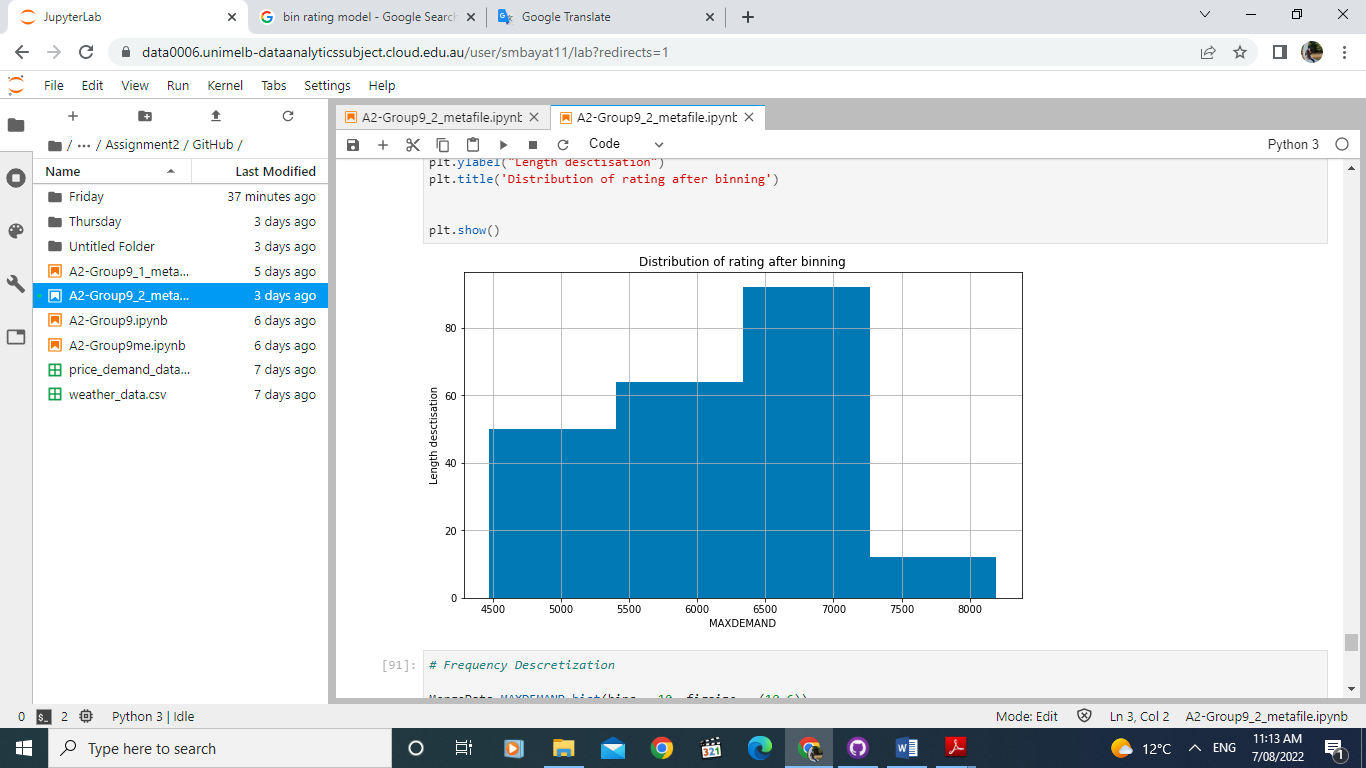


Figure 8: Distribution of MAXDEMAND rating after binning

3. How effective are your models? How have you evaluated this?

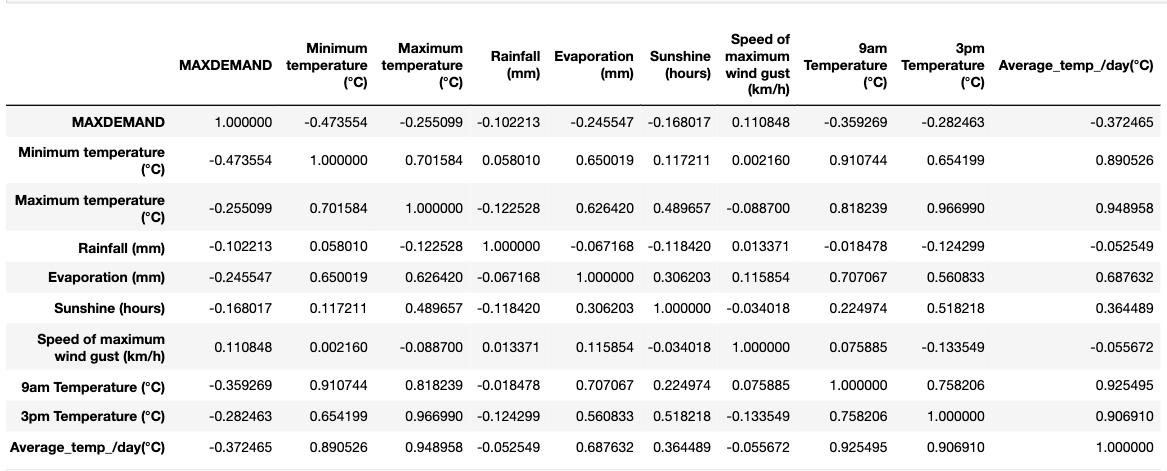
To avoid overfitting, we have applied machine learning by splitting the data table into two sections, test data, and training data. Calculated the R^2, intercept, and coefficient values.

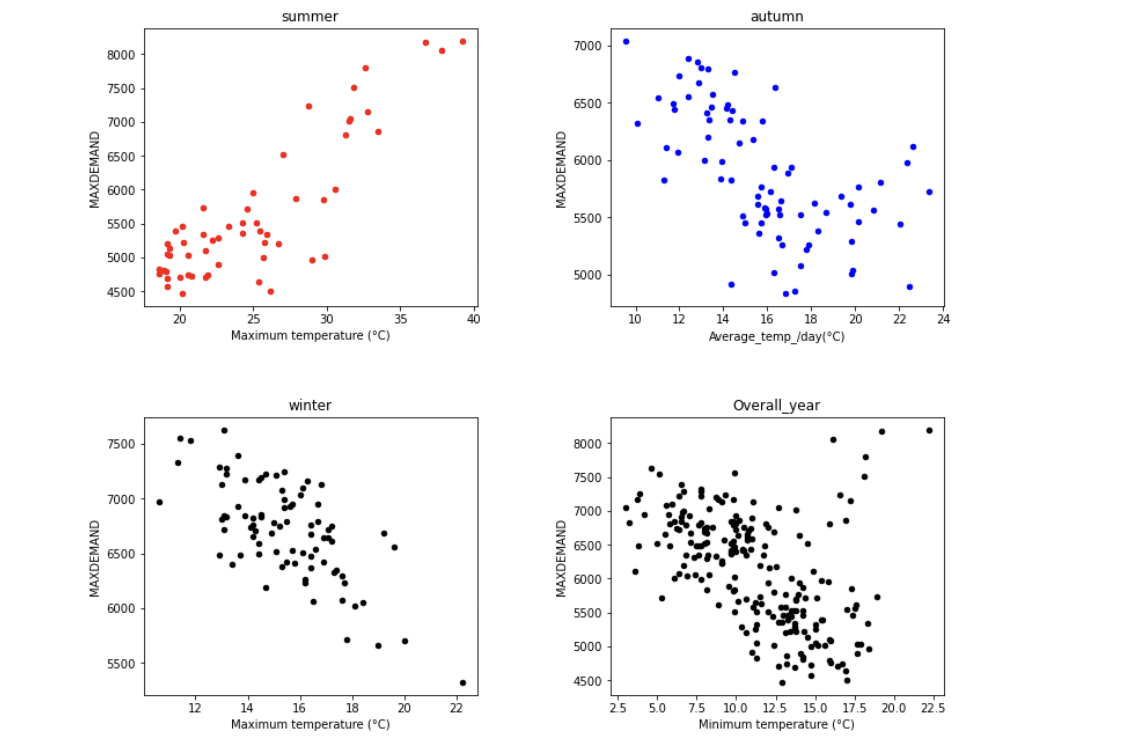
4. What insights can you draw from your analysis? For example, which input variables

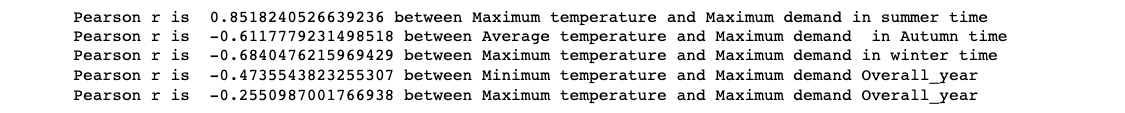
are most valuable for predicting energy usage/price?

### Model 1 - Maximum temperature vs. Maximum demand:

According to the analysis, the maximum temperature tends to affect the maximum daily energy usage by season. The hottest temperature is likely to consume the highest daily demand in summer, and lower temperatures tend to have higher daily demand in autumn and winter.

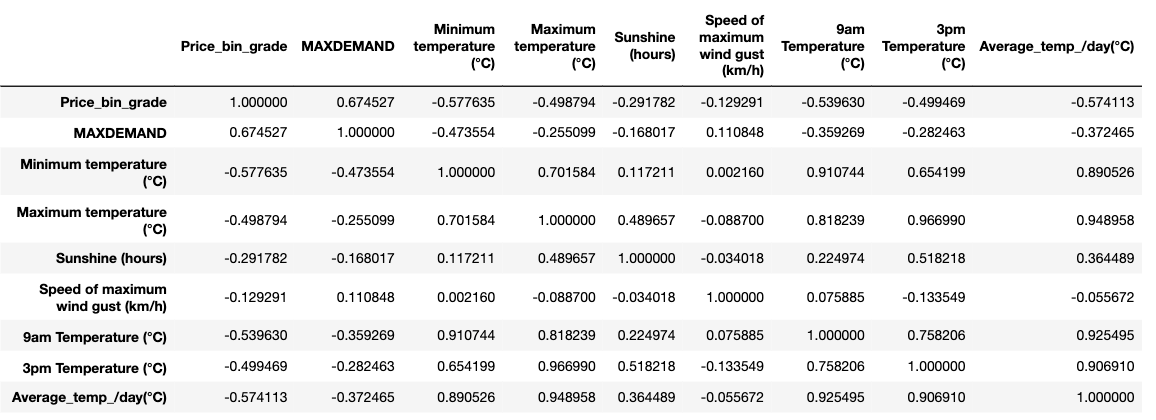


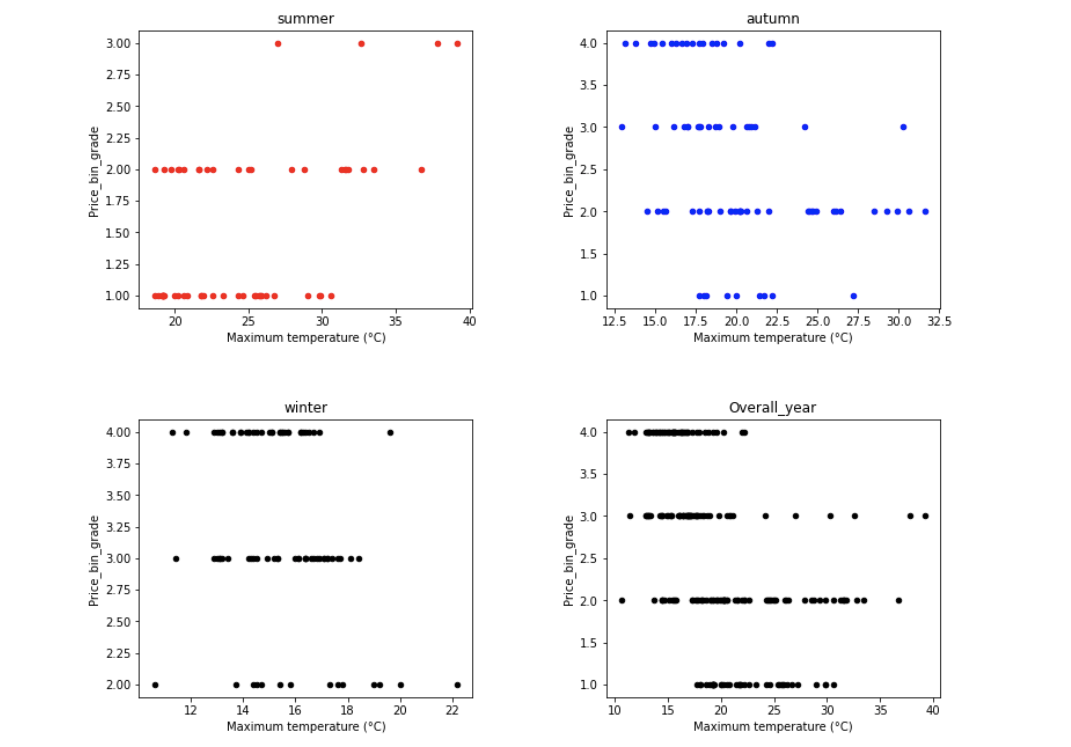




### Model 2 - Maximum temperature vs. Maximum daily price category:

With the maximum daily price category model, the price category is highly related to the maximum daily demand. However, compared to higher temperatures in summer, the lower temperatures in autumn and winter tend to have the highest price category. Solar energy production may reduce the price in hotter weather.





Text, letter

Description automatically generated

# feature selection

Although we selected the features using Intuition (possible to evaluate the "goodness" of each feature) for the second model, we analyzed the chi2 methods to showcase the feature selection.

#Interpret the model

5. Why are your results significant and valuable?

#Discussion

#conclusion

6. What are the limitations of your results, and how can the project be improved for the future?

There is not enough data. With only eight months of price category and weather data, we cannot build up the model to predict the demand, price category, and weather for a whole year. In addition, with only 218 rows, it's hard to split the data set for training and testing the model. Without the actual price data value, it also limited the accuracy of the model prediction results.

The maximum demand can be impacted by many other factors such as sunshine, evaporation, and wind gust. Also, the sunshine can produce significant solar energy in summer, reducing the energy demand.

The wind gust can be used to produce energy using the kinetic energy created by air in motion, and it can reduce the total energy demand.

The data on solar and wind energy production is not given in the dataset, which we consider a limiting factor to analyze further. In the future, this analysis can be improved by adding (or machine learning) the solar energy production per hour and wind energy production by hour data, which will significantly reduce the energy demand. Then, there is a possibility of developing a multidimensional model, which can lead to a neural network if we feed more data to the model, and the accuracy will go further high.