# **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**

For building a future business model in any energy production or supplying company, a predictive model can be helpful by analyzing energy demand and prices in different weather conditions. In this analysis, we have developed two individual models based on energy demand and prices in a particular period. Here we represent our model and detailed analysis, presenting how these parameters influence the overall model performance based on weather conditions.

**Assessment**

**1. What wrangling and aggregation methods have you applied? Why have you chosen these methods over other alternatives?**

In this analysis, all the missing values were dropped to clean the data set. We have used the OpenRefine tool to identify the missing values. Twenty-five rows of data were dropped from the 'weather\_data' table because they had some NaN values and missing variables. The cells that contained missing data were not uniform. 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?**

Based on the data, we have built a linear regression model. We merged both CSV files to create one table. Applied the Group by 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. The correlation values of all features revealed that the maximum demand had been affected by temperature. As shown in figure 1, the relationship between the maximum Direction and minimum temperature, maximum temperature, and average temperature is -0.47, -0.26, and -0.37, respectively.

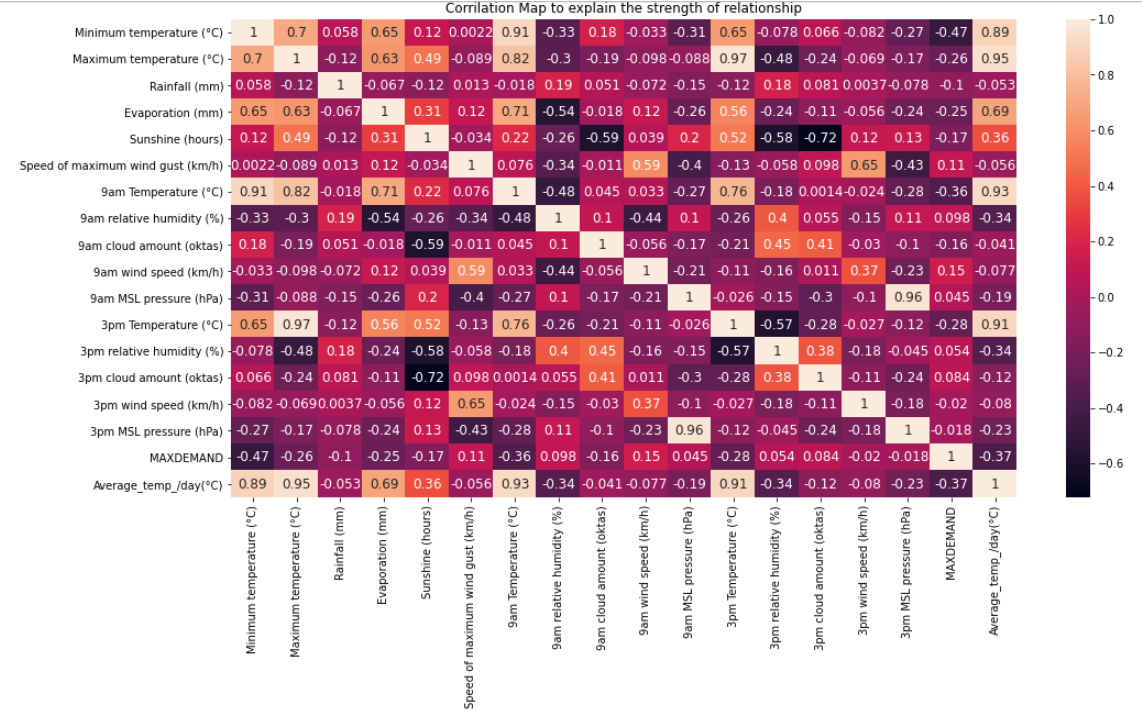


Figure 1: the correlation values of all features

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 were poor:

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

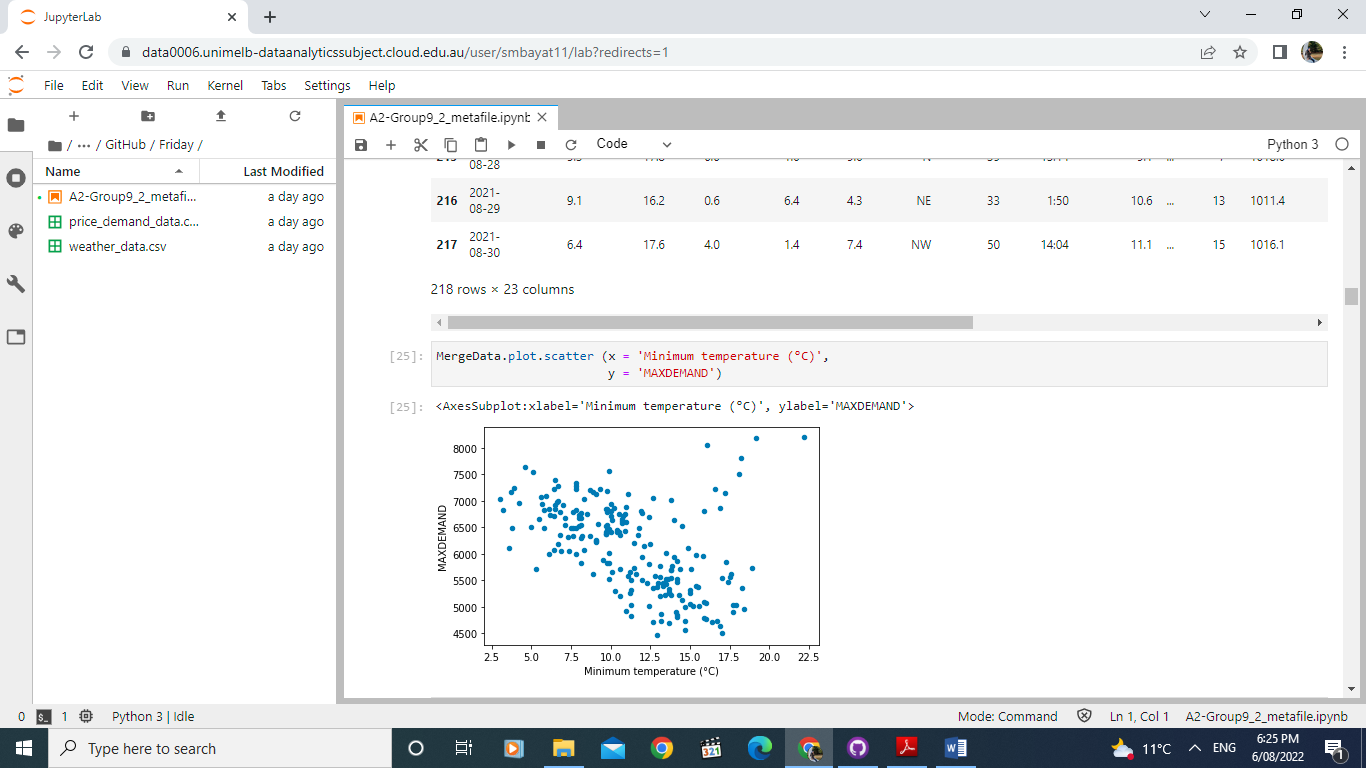


Figure 2: Maximum Demand versus minimum temperature

We also conducted a time-specific temperature variables analysis. We have analyzed the variation of the maximum demand at 3 pm and the daily average temperature. The pattern of the scatter plot is similar. However, the highest order is the same as the lowest market slightly changes, but not significantly.

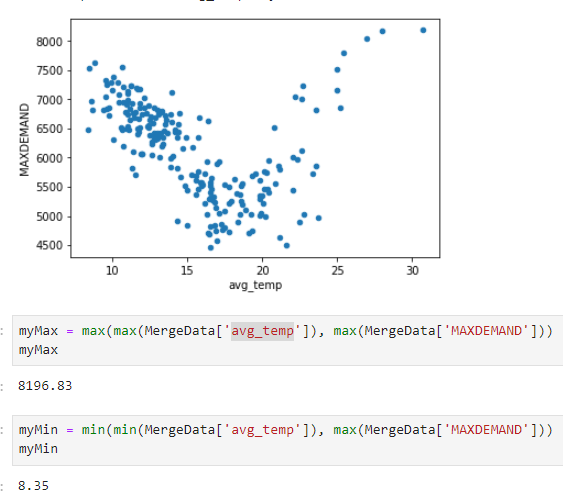
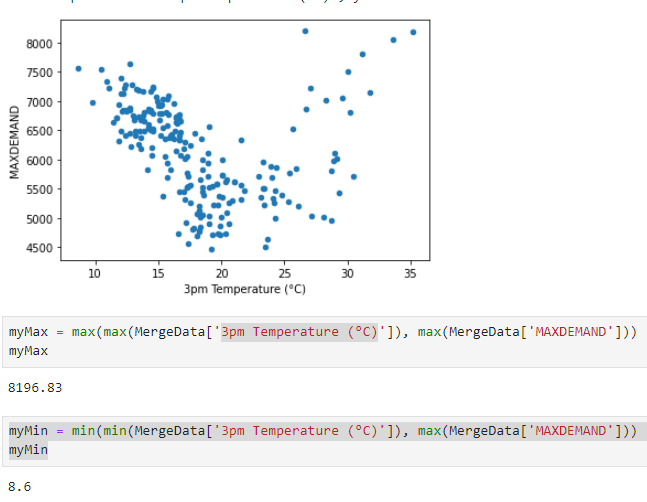
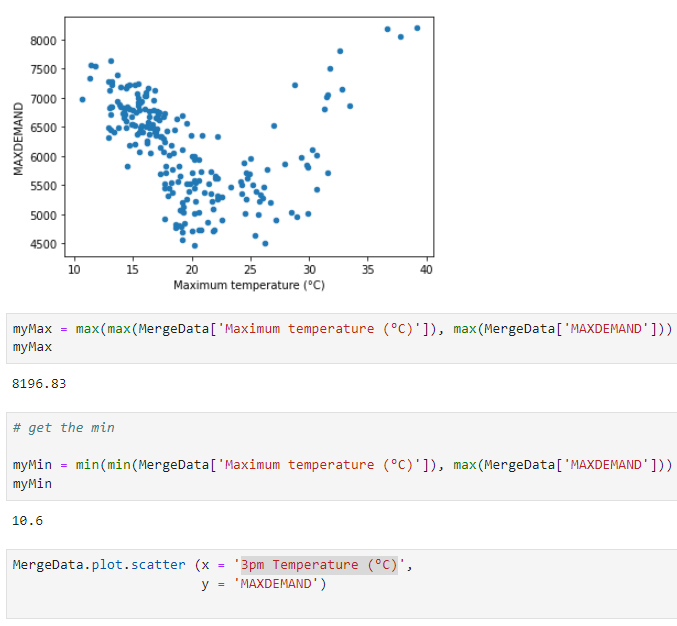


Figure 3: Maximum Demand versus maximum, average, and 3 pm temperature

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 conducted a further analysis. Our analysis shows that all temperature-related patterns were U-shaped overall.

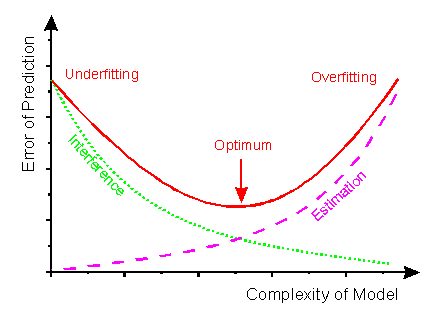


Figure 4: Model complexity (<http://www.frank-dieterle.de/phd/2_8_1.html>)

The optimum temperature was selected according to the model complexity, as shown above.

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.

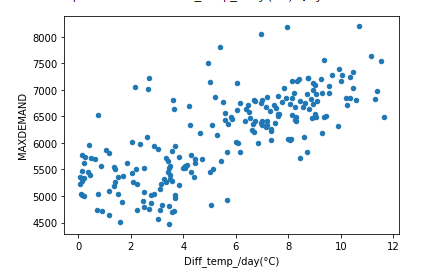


Figure 5: Maximum Demand versus Diff\_temp\_/day(°C)

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 other factors that might affect the energy demand, such as sunshine, wind gust, evaporation, and cloud amount. The scatter plots were plotted against each dataset's data, and correlation values were calculated. This plot demonstrated a linear relationship between the parameters and their dependence.

Timeline, box and whisker chart

Description automatically generated with medium confidence

Figure 6: the correlation values of all features

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. There might be little relation between sunshine hours and MAXDEMAD.

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

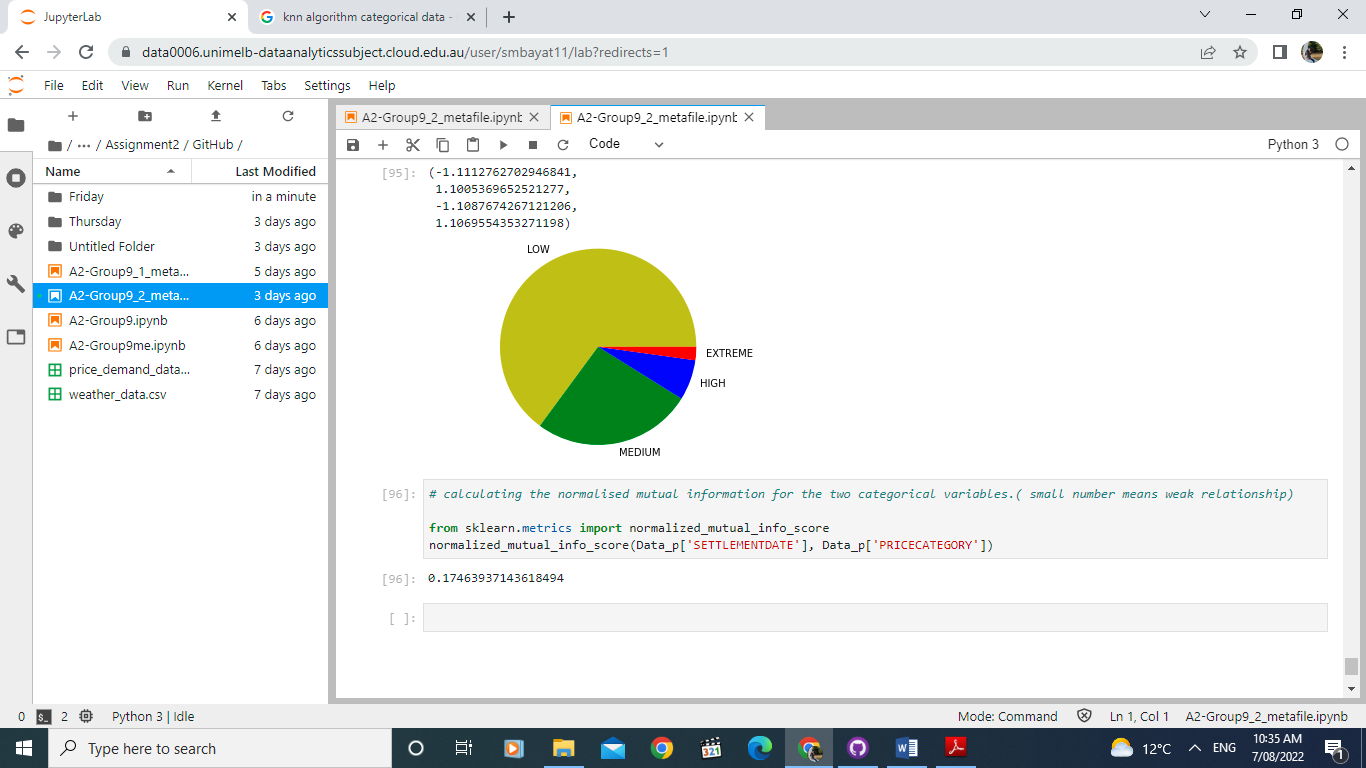


Figure 8: The price category's pie chart

Binning the maximum daily price category, we again turned all rates to the label. Finally, we merged weather and price CSV files to create one table. We have applied the Groupby method to the date column.

We used the KNN algorithm for this part 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 the Equal length bin technique for price category discretization (Fig 9):

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

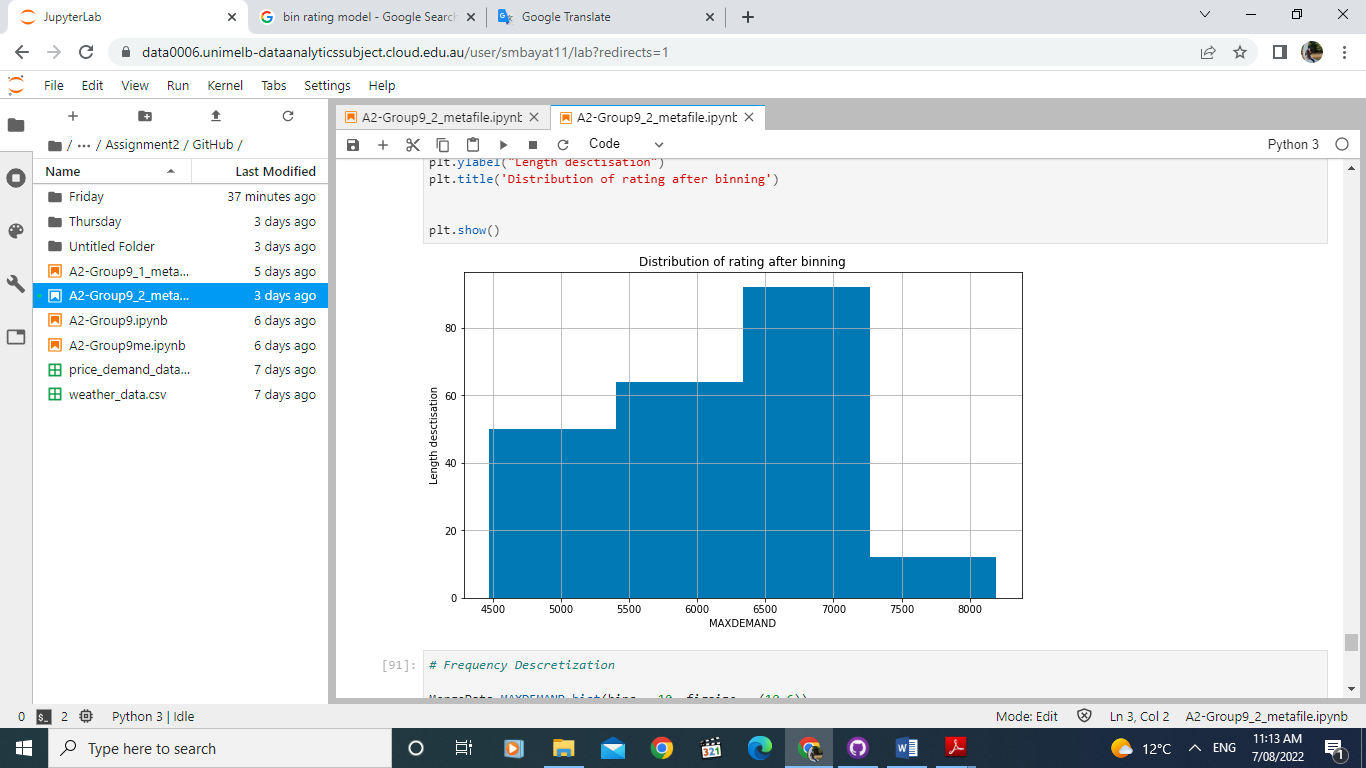


Figure 9: Distribution of MAXDEMAND rating after binning

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

**Determine Model accuracy scores by K-Nearest Neighbors (kNN) method:**

Effectively managing the maximum daily energy used, it is essential to analyze the energy uses in different sessions because the weather significantly affects total energy consumption. In this model, we have utilized K-Nearest Neighbors (kNN) classifier for maximum daily energy uses prediction based on classification. K-Nearest Neighbors is one of the most straightforward supervised classification techniques that provide exemplary performance results for the optimal value of K.

In this model evaluation, we have considered four stages: data processing, prediction, validation, and performance evaluation. The datasets containing weather conditions, total demand, and price categories particular period in a year that has been used in the experimentation. Two steps are involved in the prediction stage: training and testing. In the training stage, data values are given to the classifier to train it. The training data have labels associated with them that represent their class. During the testing phase, the KNN classifier is given unlabelled data points, and the algorithm generates a list of K nearest data values. In this model, we have divided it into different training and testing features for both models to find the performance and accuracy of the predictor.

For the first model, we have trained the dataset by considering 'Minimum temperature (°C)', 'Maximum temperature (°C)', 'Rainfall (mm)', 'Evaporation (mm)', 'Sunshine (hours)' as train predictors and binning maximum energy demand in a day as a class label. We have split the train and test size into 80% and 20% in 42 random states. The highest accuracy observed for our prediction is 72.7% for daily maximum energy use based on weather conditions.

For the second model, we have trained the dataset by considering 'Minimum temperature (°C)', 'Maximum temperature (°C)', 'Sunshine (hours)', 'Evaporation (mm)', 'Speed of maximum wind gust (km/h)', '9 am Temperature (°C)' as train predictors and binning maximum daily price category in a day as a class label. We have split the train and test size into 87% and 13% in 42 random states. The highest accuracy observed for our prediction is 55.1% for the maximum daily price category based on weather conditions.

**Determine Model accuracy scores by the K-Fold method:**

K-Fold cross-validation is a method where a given dataset is split into a K number of folds. Each fold is utilized as testing set in a particular selective point, and the remaining folds train the model. This process continues until each of the folds has acted as a testing fold individually and the rest of the fold acted as a training fold. A score is retained after completing each iteration, and the sum of total iteration scores is averaged to finalize the accuracy of the model performance. We have applied the K-Fold method in both models to understand the model performance in a particular training set. First, we split the data set into ten folds and shuffled each of the folds in every test in 42 random states. We have assigned the training and testing features and class sets for each iteration by indexing (train index, test index). We have trained the model in a specific iteration using the training index of each iteration of the K-Fold process and sum up the overall class test and prediction value.

For the first model, we have trained the feature dataset. The model accuracy observed for our prediction is 49.9% for daily maximum energy use based on weather conditions.

For the second model, we have trained the feature dataset. The model accuracy observed for our prediction is 49.9% for daily maximum energy use based on weather conditions.

**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, we have identified the daily temperature as are most valuable variable for predicting energy usage or price. Figure 10 demonstrates that the maximum temperature affects the maximum daily energy usage by season. In summer, we have observed the maximum demand increases with an increasing temperature, while winter shows the opposite scenario. Overall, we can see that the energy is remarkably higher in low and high temperatures. At comfortable temperatures, energy is not increasing as much as in winter and summer.

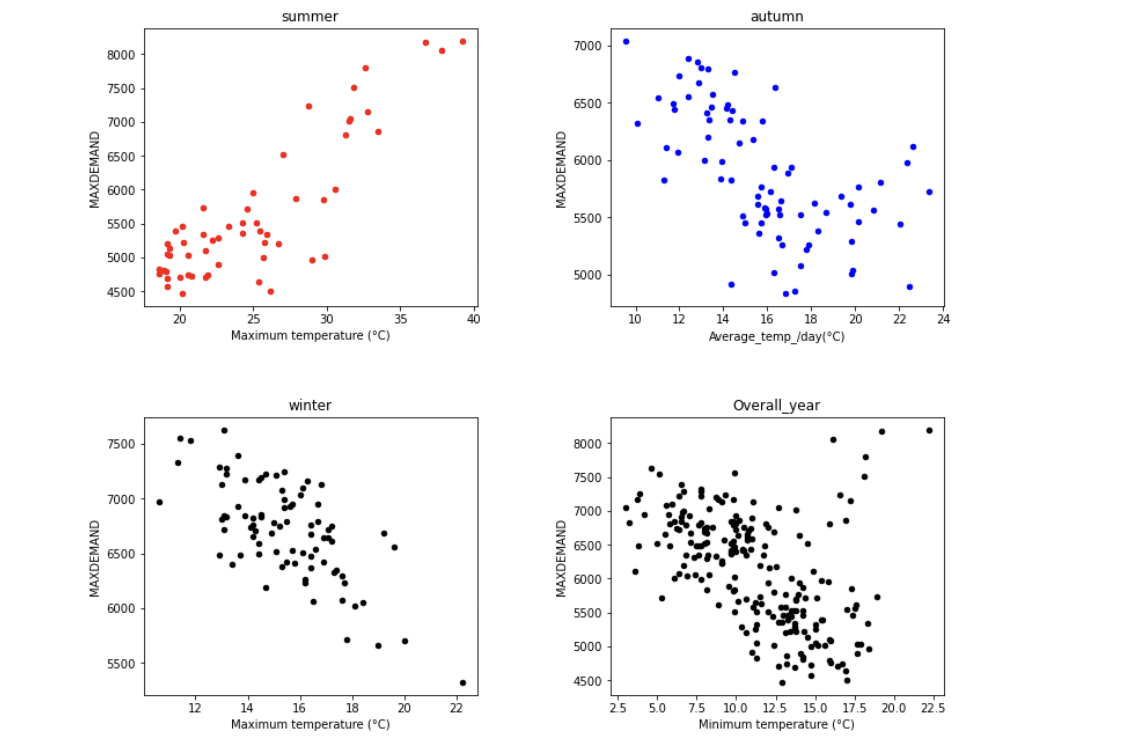
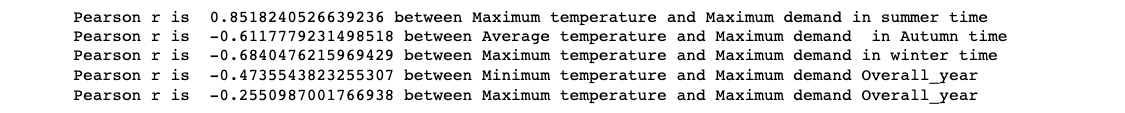


Figure 10: Maximum energy used vs. temperature curve in different sessions.

Pearson correlation between maximum energy used and temperature in a different session.



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

With the maximum daily price category model, the price category is highly correlated to the maximum daily energy demand. However, compared to higher temperatures in summer and the lower temperatures in Autumn and winter, we have observed the highest price category in the low temperature region compared with high temperature.

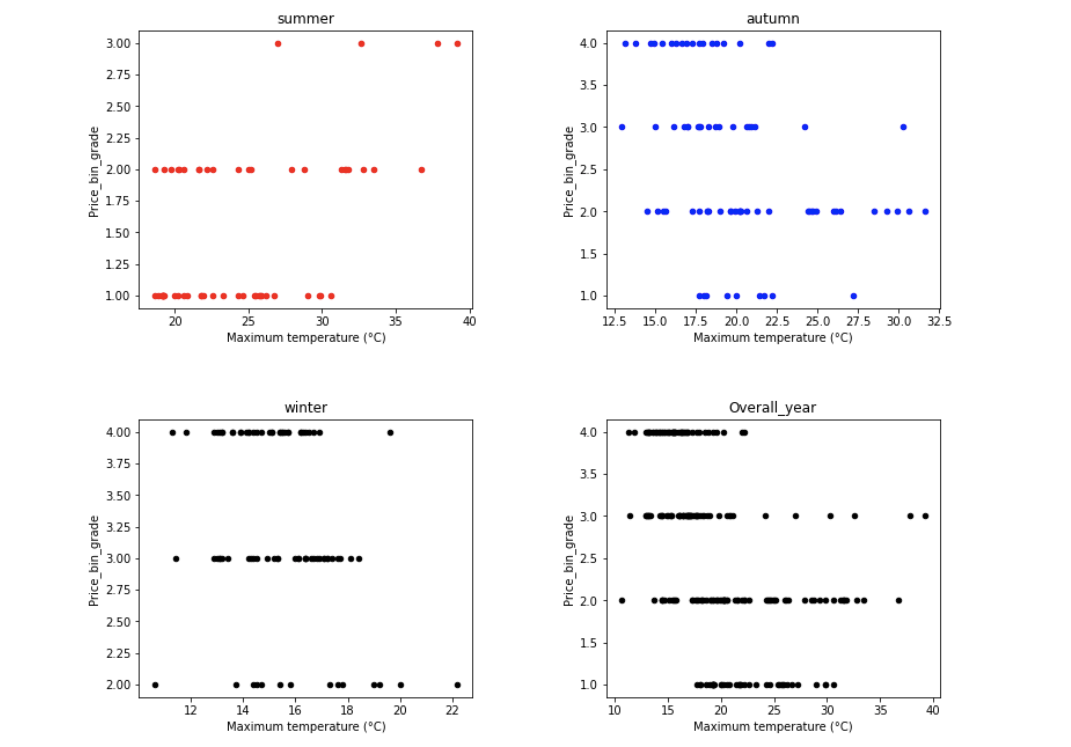


Figure 10: Energy price category vs. demand curve in a different session.

Pearson correlation between maximum energy used and temperature in a different session.

Text, letter

Description automatically generated

**5. Why are your results significant and valuable?**

K-Nearest Neighbors (kNN)model accuracy analysis for the first model shows that the highest accuracy for model prediction is 72.7% for daily maximum energy use based on weather conditions. For the second model, the highest accuracy observed for our forecast is 55.1% for the maximum daily price category based on weather conditions.

On the other hand, the Pearson correlation shows that our model predicted an 85% correlation between maximum energy uses and temperature in summer and 61% in Autumn, and 68% in winter. Similarly, on the price model, overall, we have observed our model predicted a 71% Pearson correlation between demand and energy price. Therefore, after analyzing k-NN and K-fold model accuracy and Pearson correlation, we conclude that our model results are more significant and valuable for model prediction.

**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 could not build the model to predict the demand, and price category 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.

**Discussion and conclusion**

The purpose of building these models is to predict the maximum daily energy used and energy pricing based on weather conditions that significantly help the energy companies to understand how weather conditions may influence the overall energy demands and the effect of several variables on energy prices that can help them a detailed understanding of their future business model based on future energy uses.

For both models, we have used supervised algorithms. The model between the maximum daily energy consumption and average temperature, due to MAXDEMAND, is continuous data. Regression analysis is suggested for modeling and finding the cause-and-effect relationship between variables. For the model which predicts the maximum daily price category based on the weather data, since the price category is labeled, KNN is an excellent option to build a model for discrete variables.