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 be at its highest on days with hotter temperatures, there are other factors affecting on energy production and consumption. For example, when the sunshine increases, the solar energy can be produced, and the sunshine may unevenly heat the earth’s atmosphere by solar radiation which will increases the temperature. Meantime the cloud amount will have a significant impact on sunshine getting all the way down to the earths surface. Wind can produce energy, but this has affected by topography.

This analysis is based on how the daily maximum energy needs affected by the maximum day temperature and the changes on price demand accordingly for the given set of dates. And we build the model to predict and forecast the future energy demand.

Assesment

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. 25 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 check raw wise, there were many empty cells, so we decided to remove the values rather than imputing.

The columns ‘Direction of maximum wind gust’, ‘9am wind direction’, ‘3pm wind direction’, column data were replaced. The 8 directions were annotated with one or two letters to make it uniform for analysis(aggregation).

One of the important columns in our analysis, is Settlement date column in the ‘price\_demand\_data’ was in object format. Changed it into the datetime format for ease the analysis by splitting the date and times into two separate columns.

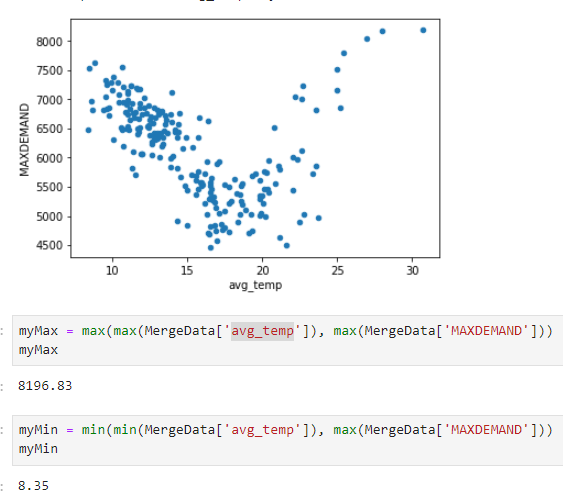
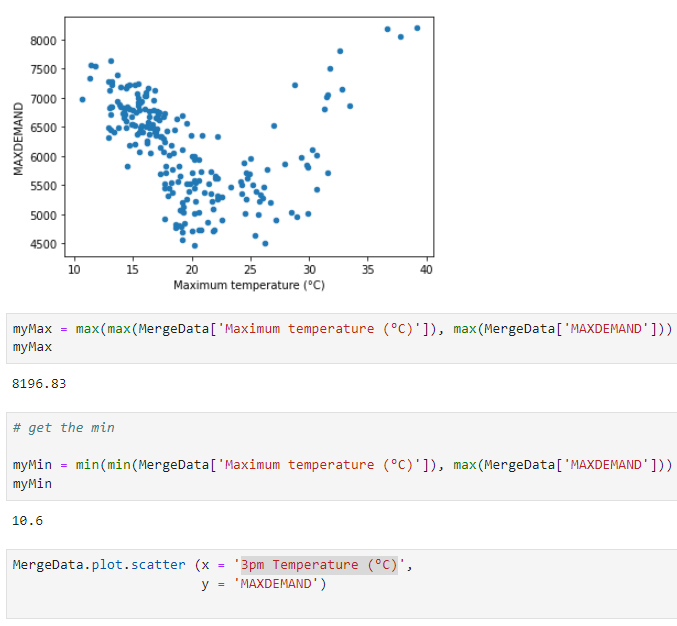
2. How have you gone about building your models and how do your models work?

Based on the data provided, we build a linear regression model.

Merged both Excel.csv files to create one table. Applied Groupby method on date column and the maximum demand for the day (Maxdemand) was considered to build the model.

We analysed the maximum temperature, average temperature and 3pm temperature(as in summer the temperature goes high around 3 pm). Although the highest demand is same the lowest demand slightly changes, but not significantly. The pattern of the scatter plot is similar.

Chart, scatter chart

Description automatically generated  

But the graphs were not linear. So we analysed deep into it.

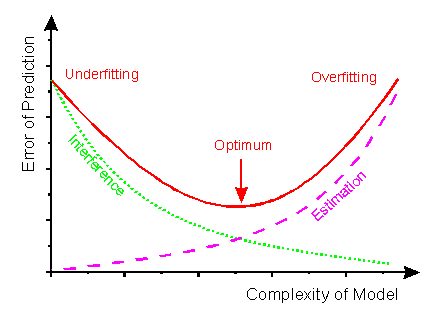


Figure: Model complexity (<http://www.frank-dieterle.de/phd/2_8_1.html>)

According to the model complexity diagram above, the optimum temperature was selected which is 200C. then we derived a linear graph as shown below.

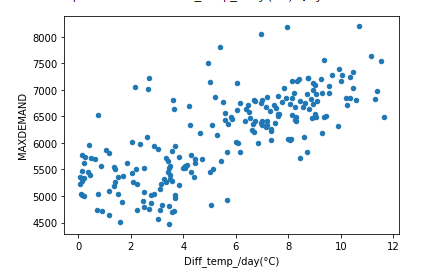
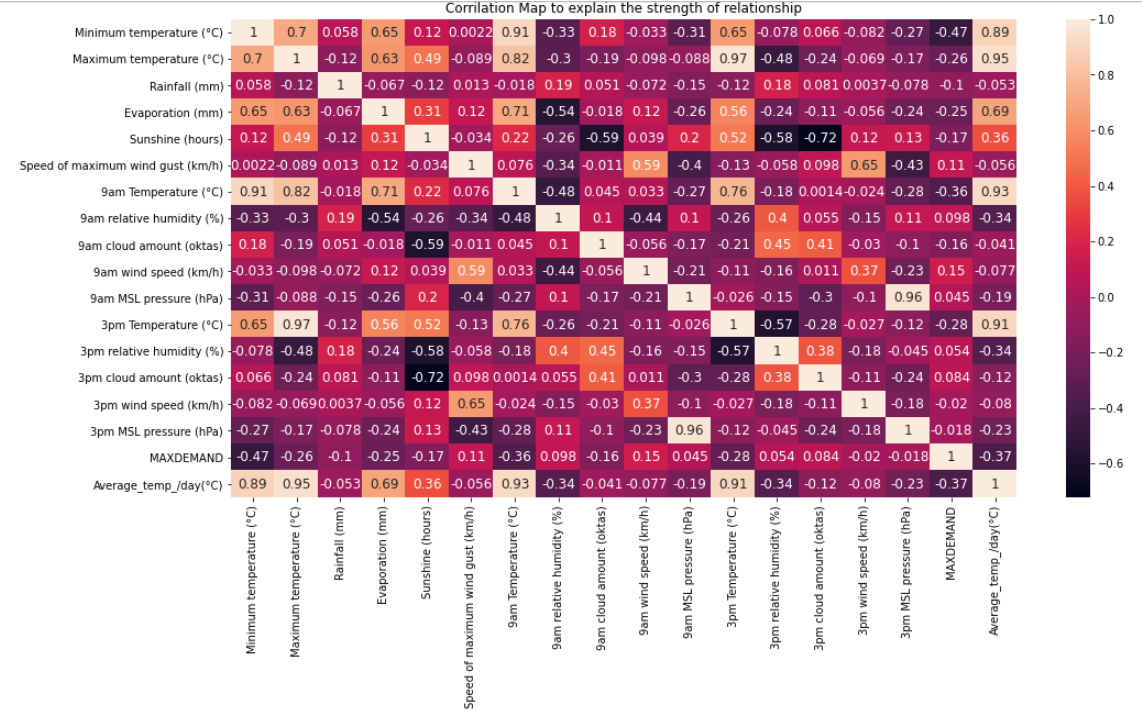


Figure: Starts with the optimum temperature, to give a linear model.

Then we were interested to analyse how are the other factors might affect on the energy demand, such as sunshine, wind gust, evaporation and the 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 1: the correlation values of all features. Figure 2: The scatter plots

The chart explains all the relationships and the figure 1 correlation values explains the relationship and their strength.

Sunshine:

When the sunshine increases, the temperature increases. The amount of cloud will impact on the sunshine. But when the sunshine increases, it produces the solar power, which will have an impact on energy demand. So we cannot expect the same pattern on sunshine-max demand plot as temperature- max demand scatter plot.

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Figure:1 Figure:2

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

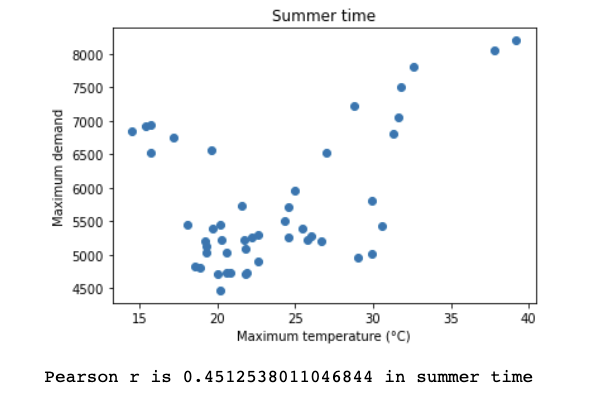
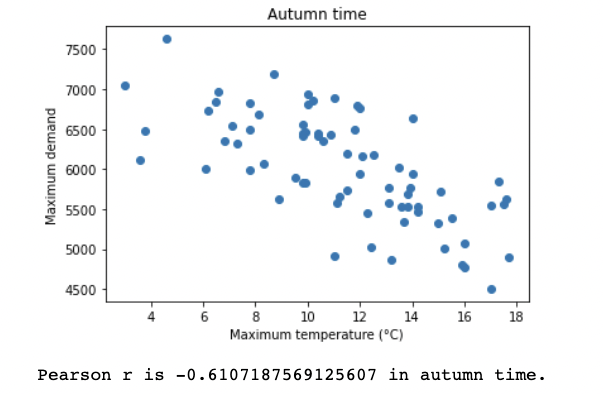
3. How effective are your models? How have you evaluated this?

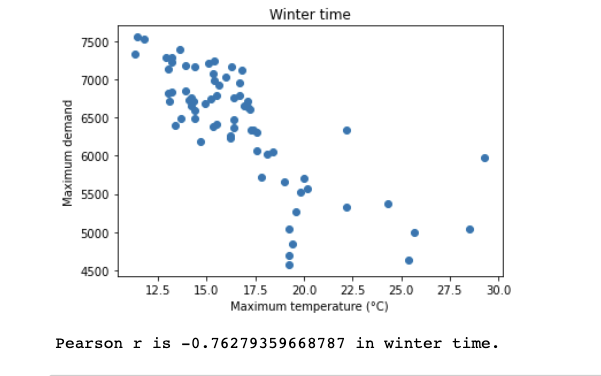
We have applied machine learning via splitting the data table into two sections, test data and train data to avoid overfitting. Calculated the R^2, intercept and coefficient values.

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

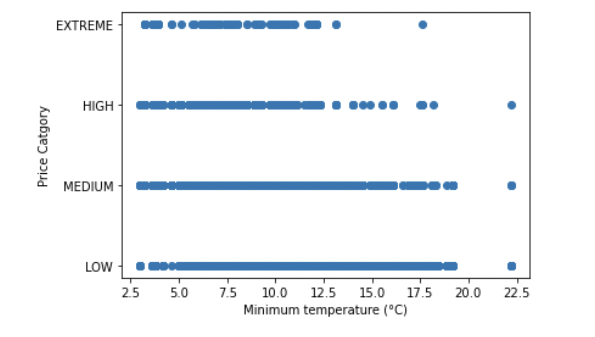
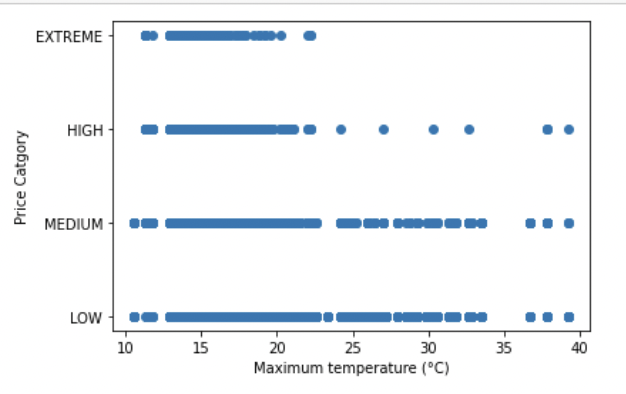
are most valuable for predicting energy usage/price?

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.



With the maximum daily price category model, the lower temperatures tend to have a higher price category comparing to higher temperatures. The solar energy production may reduce the price in hotter weather.



# feature selection

For the second model although we selected the features using Intuition (possible to evaluate “goodness” of each features), we analysed the chi2 methos 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 future?

There is not enough of data, only 8 months of price 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 of data, it’s hard split the data set for training and testing the model.

The maximum demand can be impacted by many other factors such as sunshine, evaporation, wind gust. Also, the sunshine can produce a significant amount of solar energy in specially summer which can reduce 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 energy production and wind energy production are not given in the dataset, which we think it as a limitation factor to analyse further. In future, this analysis can be improved by adding (or machine learning) the solar energy production per hour, wind energy production by hour data, which will significantly reduce the energy demand. Then, there is a possibility to develop a multidimensional model which can leads to a neural network.