

**Temperature and Energy Demand**

**By:** Pranav Kutty, Wen Wen, Nishandeep Brar, Yunru Zhu

# 

# 

# 

# Background

## Intentions

For this project we intended to find the correlation between energy consumption and the weather in Australia. We approached this by having each member look at an individual state so that we each had lighter datasets to look at. This allowed us to create models and compare and contrast said models. Thus, we were able to reach our desired conclusions through the models and analysis that we each did.

## Contributions

As mentioned in the “Intentions” section, we approached the coding by separating the workload by state. Pranav worked on the South Australia dataset, Wen worked on the New South Wales dataset, Nishan worked on the Victoria dataset and Yunru worked on the Queensland dataset. We initially had a fifth member, Zihan, who was meant to work on the Tasmania data, but unfortunately could not continue. For the presentation, Pranav worked on the agenda, background and conclusions, Wen worked on the cleaning, Nishan worked on the exploratory data analysis and Yunru did the modelling of demand. For the report, the background and conclusions were done by Pranav, the cleaning was done by Wen, the exploratory analysis was done mainly by Nishan, with some help from Pranav and Yunru, the modelling of demand was done by Yunru, the establishment of the three regression models was completed by Yunru with some help from Pranav, who used his data to make the second model.

## Form of Initial Data

To begin our project, we received two folders of data as our datasets. One folder contained temperature and weather data, consisting of 6 text files with 34 variables separated by region, along with two more text files containing notes and information. The second folder contained energy demand data, which came as 1,136 separate csv files, containing 5 variables.

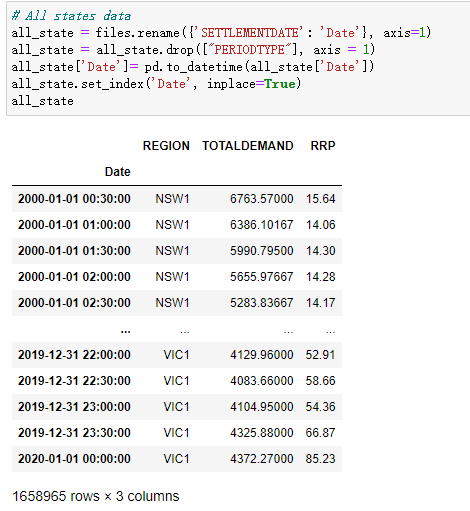
The energy demand data was gathered from the Australian Energy Market Operator and the temperature and humidity data comes from the Bureau of Meteorology Automatic Weather Station. Both sets feature data for the 20-year period between 2000 and 2019, and both sets have their values recorded in half hour intervals.

**Overview**

Without looking at the data we already know that there should be relationships between these two datasets. Hotter and colder weather both should lead to an increase in energy demand, as more people use energy to cool and heat their homes. Similarly, seasons like Autumn and Spring should see less demand for energy, due to the milder weather. Energy demand should also have a large impact on the time of day, as humans would tend to use more energy during the day than late at night.

# Data Manipulation and Modelling

## Cleaning and wrangling the data

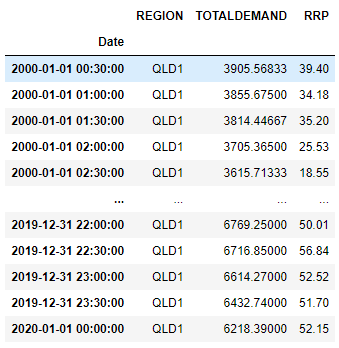
Clean ALL\_DEMAND DATA

* Dropped “PERIOD TYPE” column
* “SETTLEMENT DATE” column set as index
* format of the time changed

2000/01/01 00:30 -> 2000-01-01 00:30

* Columns containing numeric values converted to float
* Additional columns added

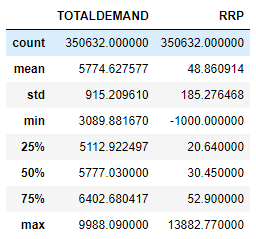
eg. “year”, “month”, “Season” to aid graphing



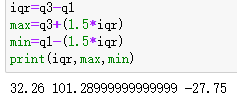
Select state by ‘REGION’

My state is Queensland, in data it's annotated as 'QLD1'

Remove outliers

By using .describe() to get 25% and 75% which is Q1,Q3

Calculate IQR, MAX,MIN

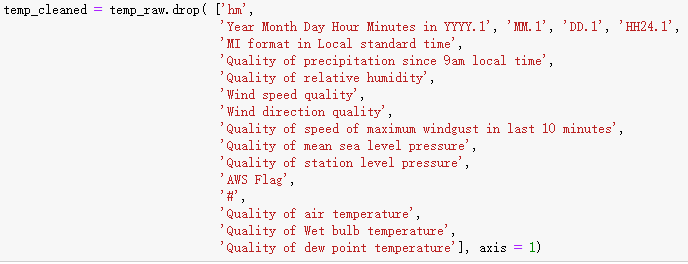
IQR = Q3-Q1

MIN = Q1 − 1.5 IQR

MAX = Q3 + 1.5 IQR

If the data is lower than minimum -23.95 or higher than maximum 102.16 we should remove them.

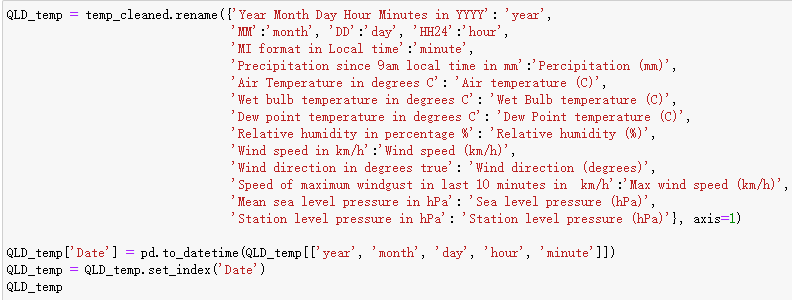
Clean Temperature Data

Read file for Queensland: 'TemperatureData\\HM01X\_Data\_040913\_999999999743964.txt'

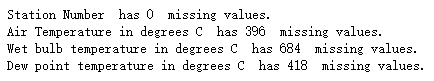
Delete Useless Columns

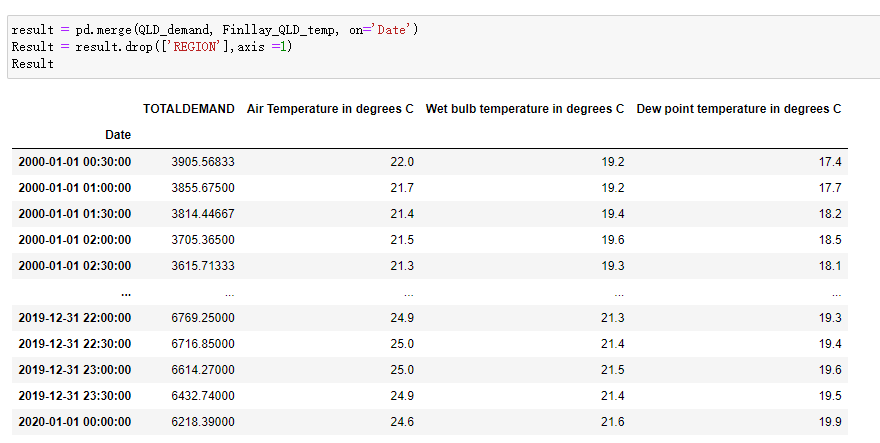
1.We only keep one of two different types of time in the data file.

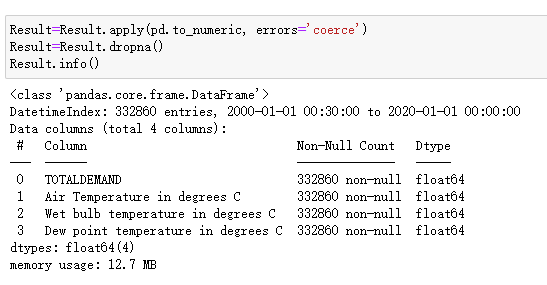
2.All the columns which were to do with quality such as “Quality of mean sea level pressure” contained the letter N and an empty string so they all had to be removed.

Rename Columns

Easier to understand the meaning of each column

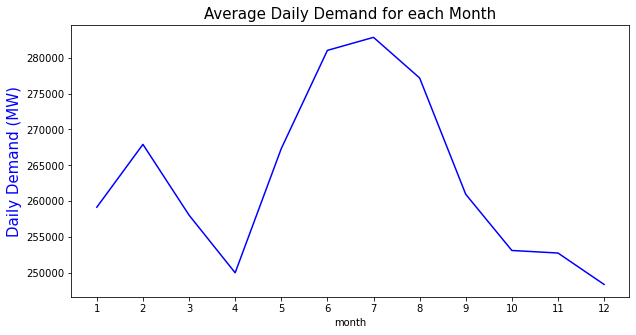
Find Missing Values In Data

Merge ALL DEMAND DATA and TEMPERATURE DATA into one table

Final Check

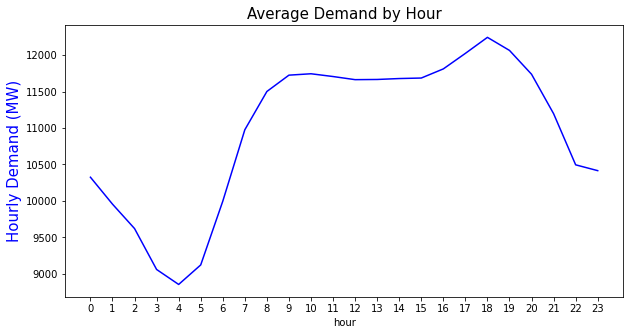
Check if there is Nan in the table, if it is dropped

Exploratory data analysis

How does daily demand change throughout the year?

This graph shows the average daily demand for each month.

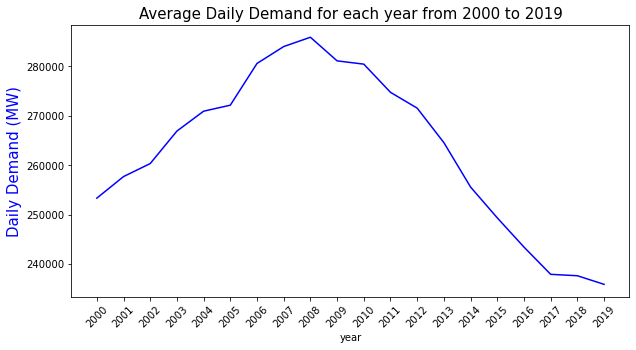
Daily demand is lowest in April and December at approximately 250 GW and highest in winter reaching a peak of over 280 GW in July.



How does demand change in the day?

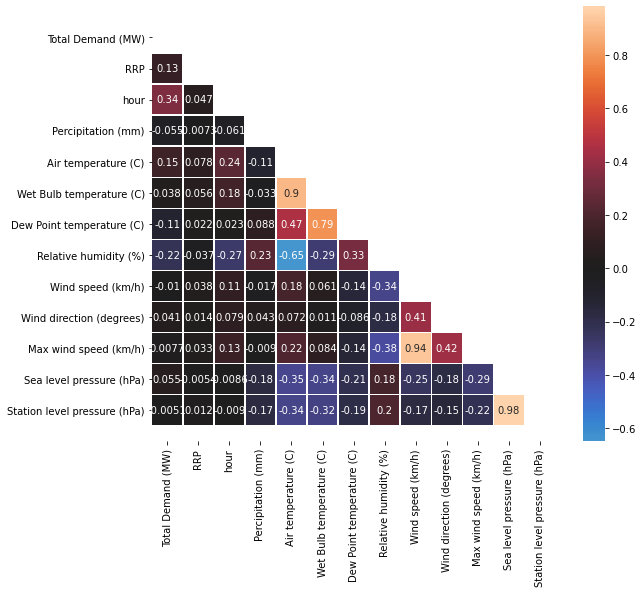
This graph shows the average hourly demand throughout the day.

We observe that demand is lowest around 4am and increases rapidly as people start waking up until finally stabilizing at around 11.5 GW by 9am. We see another peak at 6pm which may be because people are coming home from work. As the day ends we see the demand dip as people go to sleep.

What are the trends in demand from 2000 to 2019?

This graph shows the daily average demand for the years from 2000 to 2019.

Daily average demand peaked in 2009 at 280 GW and has been declining ever since up to 2019 when it was less than 240 GW.

**Heatmap**

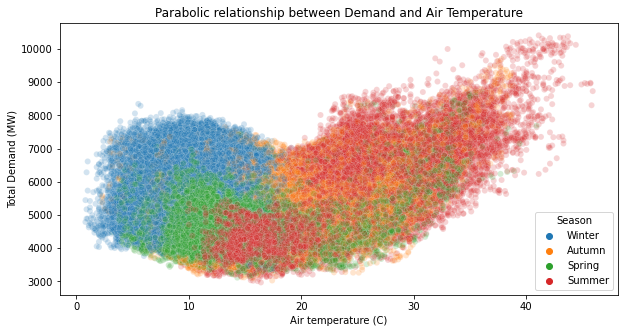
Looking at the first column, we see that the variables RRP, time of day, air temperature, dew point temperature and relative humidity all have an absolute correlation greater than 0.1 with total demand.

We should also take note of the very high correlations (> 0.5) between the following variables as they may result in multicollinearity when the model is created:

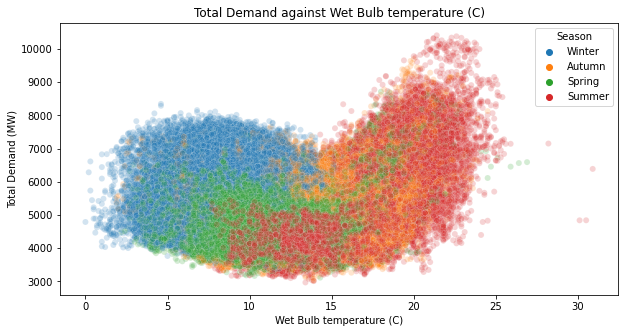
* Sea level pressure & station level pressure (0.98)
* Wind speed & max wind speed (0.94)
* Wet bulb temperature & dew point temperature (0.79)
* Air temperature & relative humidity (-0.65)
* Air temperature & wet bulb temperature (0.9)

**Exploring the relationship demand has with other variables**

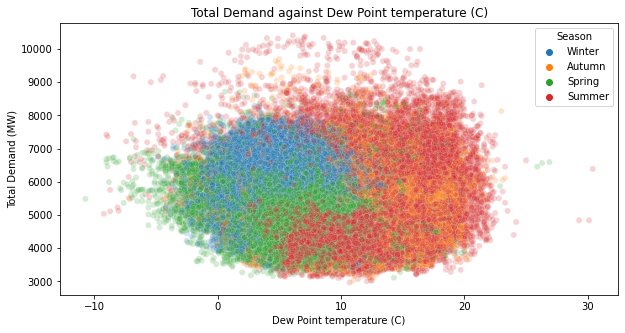
Scatter plots will be used to investigate the relationship between other variables and demand. Each point on the scatter plots represents a 30 minute period.

Relationship of Air Temperature and Wet Bulb temperature with Demand

We can see that both wet bulb temperature and air temperature have a parabolic relationship with demand.

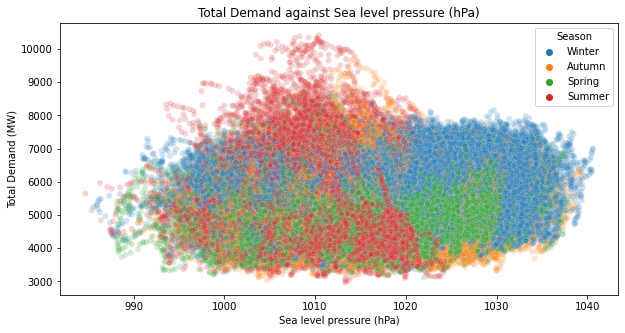
Therefore, it may be useful to include wet bulb temperature squared and air temperature squared in the model.

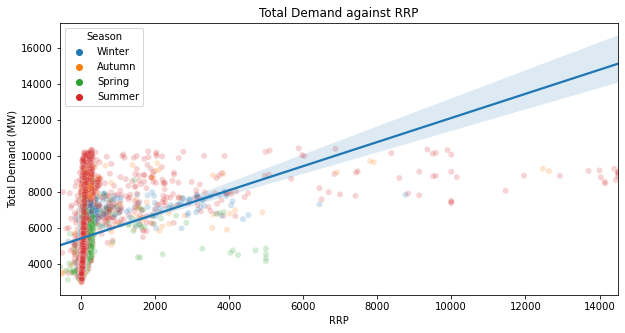
Relationship of Dew Point temperature with Demand

The scatter plot shows us that there is no clear relationship between dew point temperature and total demand.

Relationship of Station level pressure with Demand

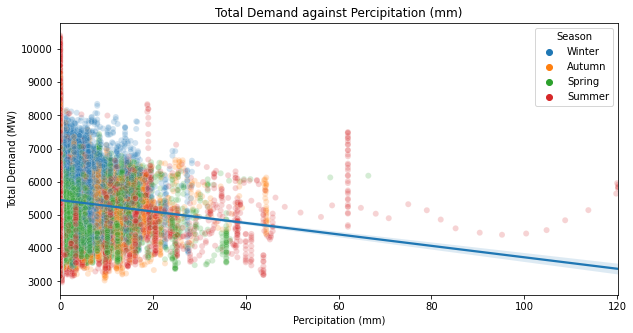
There also doesn’t appear to be a relationship between total demand and sea level pressure since as sea level pressure increases, there is no apparent change in total demand.

Station level pressure and sea level pressure have a correlation of 0.98 so we can conclude that station level pressure also has no relationship with total demand.

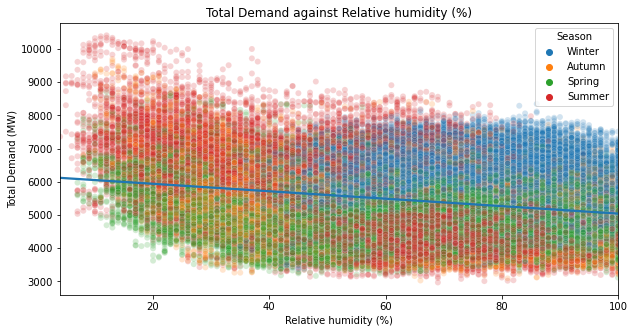
Relationship of RRP with Demand

RRP appears to have a positive linear relationship with total demand.

The exact linear relationship, which is represented by the blue line, is calculated by minimizing the sum of square errors. However, may be impractical because this chooses a model which is heavily influenced by the outliers. The vast majority of the data points have an RRP below 1000 so it may be better to use a different type of error function which produces a linear model with a higher slope. The current error function causes us to sacrifice a significant amount of accuracy for the majority of points so that the outlier won’t have large errors.

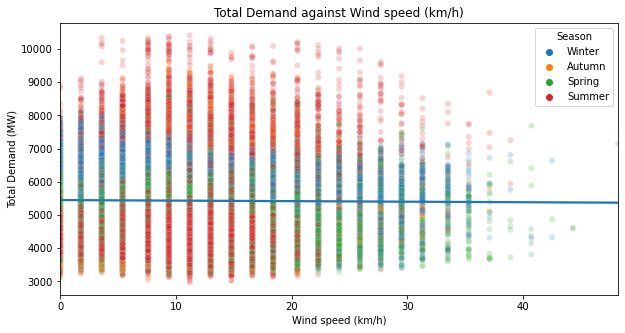
Relationship of Precipitation with Demand

From observation of the plot, the relationship between demand and precipitation appears to be negative and linear. The majority of the data can be seen between 0 and 40mm of precipitation, but the data does appear to be trending downwards and following the trendline, unlike the previous model. A linear model seems like the suitable choice for modelling this relationship, and this is evidenced by the data points following the linear trendline .

Relationship of Humidity with Demand

The relationship between total demand and relative humidity is a negative logarithmic relationship.

Therefore it may be better to use -ln(relative humidity) for the model instead.



Relationship between Wind speed and Demand

The scatter plot shows no relationship between total demand and wind speed. This is backed up by the linear regression model which produces a horizontal line indicating that the behaviour of total demand does not change on average as wind speed increases.

Since max wind speed and wind speed have a correlation of 0.94, we can conclude that max wind speed also does not have a relationship with total demand.

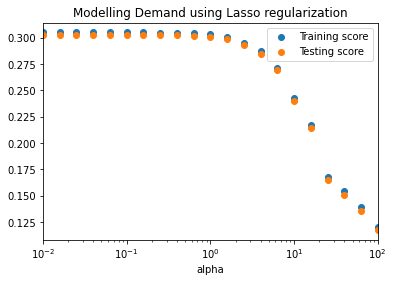
## 

## 

## 

## Modelling demand

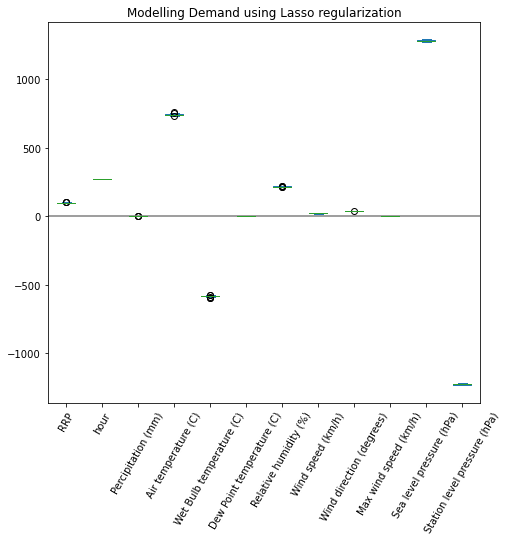
**Lasso Regularization**

We use lasso regularization in order to deal with the multicollinearity that occurs due to several variables being highly correlated such as sea level pressure and station level pressure.

Choosing appropriate alpha

We plot both the training score and testing score for several different alpha values. The value of 5 is chosen for the alpha so that there is a significant force to reduce multicollinearity but not too high as to reduce the accuracy of the model.

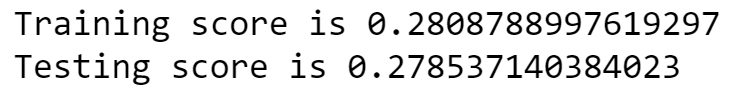
Testing the model’s consistency

In order to test how much the coefficients of the model change, we created 10 models each using a different subset of data. The variability of the coefficients from these models are displayed by the box plots.

The coefficients of these models are very similar thus the box plots look almost like horizontal lines.

However, there are clear signs of multicollinearity in the model which means the model does not capture the true relationship between some variables and demand.

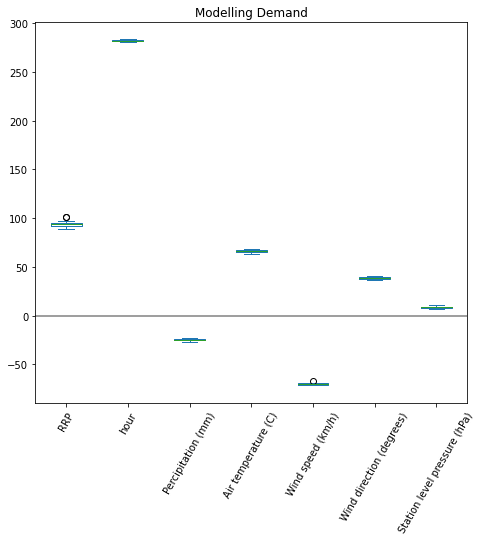
This includes the variables air temperature and wet bulb temperature where air temperature has a large positive coefficient and wet bulb temperature has a roughly equally large negative coefficient. Furthermore, sea level pressure and station level pressure are also affected by multicollinearity as sea level pressure’s coefficient is large and negative whereas station level pressure’s coefficient is large and negative.

Calculating score

The training score and testing score are very similar therefore, the model is not overfitting the data. The testing score tells us that this model can explain 28% of the variation in demand.

Since this model does not represent the true relationship of some variables due to multicollinearity, we will create another model which does not contain dependent variables which have an absolute correlation coefficient greater than 0.5.

**Model demand using dependant variables which aren’t highly correlated**

Testing the model’s variability

How well does the model fit the data?



From looking at this box plot of these variables we notice that the hour and RRP variables have the highest coefficients, meaning increases in demand should lead to considerable increases in these two variables. The precipitation and wind speed variables, on the other hand, have negative coefficients, meaning as precipitation and wind speed increase, demand is predicted to decrease. The variables of air temperature and wind direction both have moderate positive coefficients, and station level has a low positive coefficient.

When examining the training and testing data, we see their values are very close to each other, meaning that the model is not overfitting the data. A testing score of 0.13658 indicates that this model can explain about 14% of the variation in demand.

**MODEL SETTING :**

For the part of establishing the regression model, we will only consider the three variables related to temperature(Air temperature in degrees C，Wet bulb temperature in degrees C， Dew point temperature in degrees C), use them as independent variables, and use total demand as the dependent variable to verify three different types of regression models.

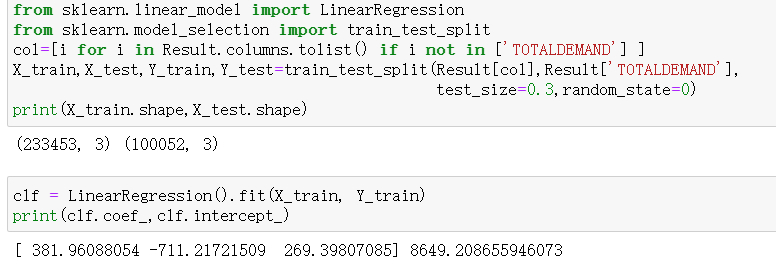
**Multiple regression model**

**1：Linear regression**

For linear regression, our ideal model would look something like this ：

**Y**(Total Demand) **= A1 \* X1**(Air temperature in degrees C) **+**

**A2 \* X2**(Wet bulb temperature in degrees C) **+**

**A3** **\* X3**( Dew point temperature in degrees C) **+ C**(interception)

Because of the large amount of data,

we first split the data into training set and

test set, and split the data according to the

proportion of training set and test set

accounting for 70% and 30% of the total

data. Print the shape of the training set and test set by code are (233453, 3), (100052, 3) respectively.

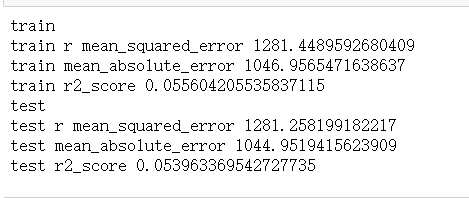
And then we found the correlation coefficients of the independent variables and the interception by fitting the train.

We got the Multiple linear regression model ：

**Y**(Total Demand) **= 381.961 \* X1**(Air temperature in degrees C) **-**

**711.217 \* X2**(Wet bulb temperature in degrees C) **+**

**269.398** **\* X3**( Dew point temperature in degrees C) **+ 8649.209**(interception)

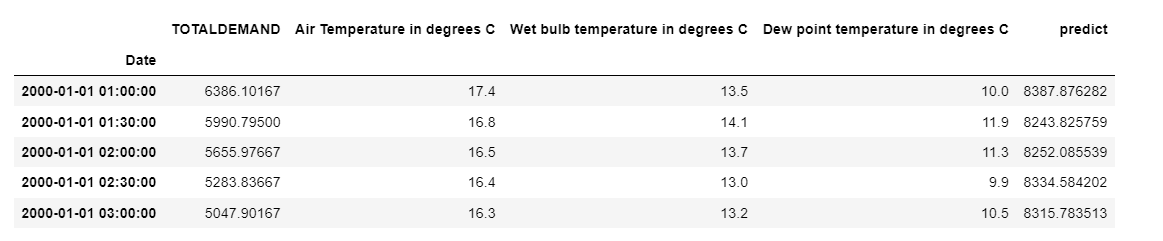


Then, MSE and MAE coefficients were calculated by Python to determine the difference

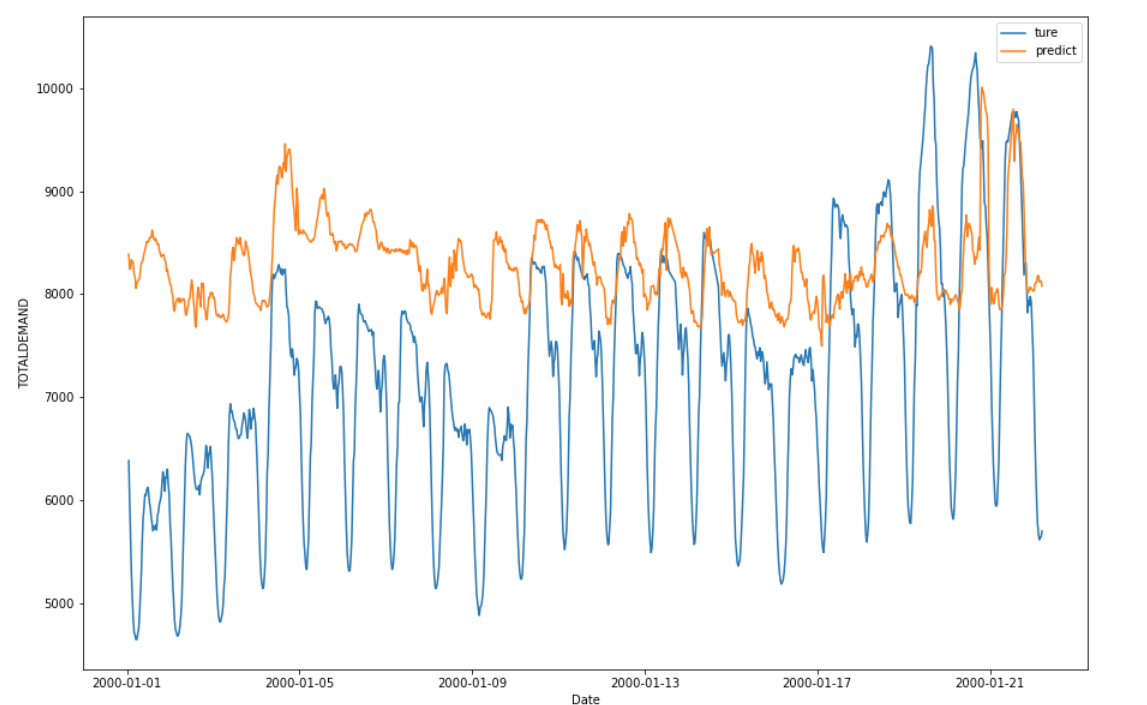
between the predicted value and the actual value, and the fitting degree of the r-square

a coefficient judgment model was calculated.

In order to intuitively show the difference between the expected value and the actual value, we can add a column of predicted values in the table, which can be calculated according to our model.



Due to the large amount of data, we present the first 1000 groups of data and intuitively compare the difference between the expected value and the actual value through a line plot.



In short, according to analysis of the influence of temperature on total demand in different seasons, we can roughly get that total demand and temperature are not a linear regression relationship. After determining the linear regression model above, by verifying the coefficients such as MAE, MSE, and r-squared, it can also be refined, and the regression model is not a linear model with high curvature.

**2：Quadratic regression**

In order to find a higher goodness of fit and verify our initial conjecture, we try to set a quadratic regression model.

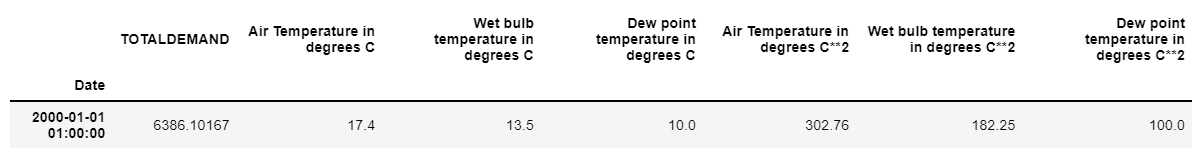
For Quadratic regression, our ideal model would look something like this:

**Y**(Total Demand) **= A1 \* X1^2**(Air temperature in degrees C) **+A2 \* X1**

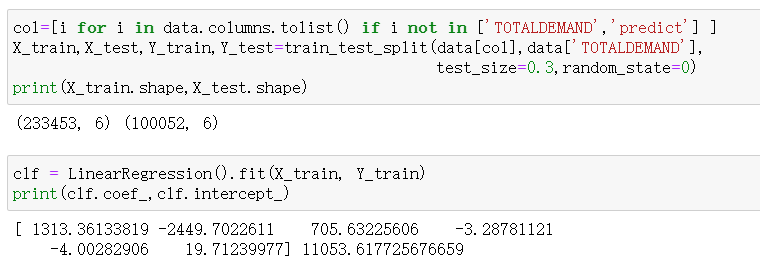
**B2 \* X2^2**(Wet bulb temperature in degrees C) **+ B2 \* X2**

**C3** **\* X3^2**( Dew point temperature in degrees C) **+ C3** **\* X3 + T**(interception)

Because we need to calculate more correlation coefficients here, we can't directly use the fit function to get the quadratic coefficients of the independent variables, so we add the square of the variable to the table.



Then repeat the same steps as above to split the data and find the correlation coefficients.

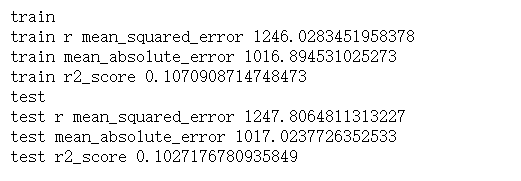


We got the Quadratic regression model ：

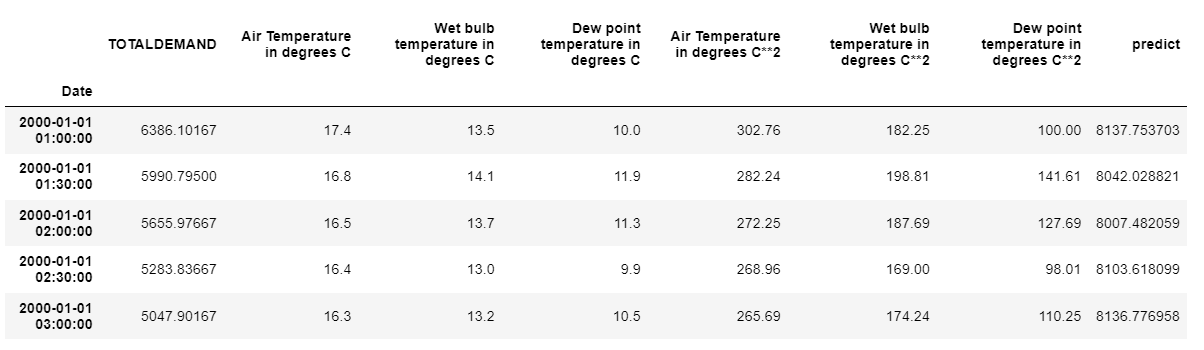
**Y**(Total Demand) **= -3.288 \* X1**(Air temperature in degrees C)**^2 + 1313.361 \* X1 -**

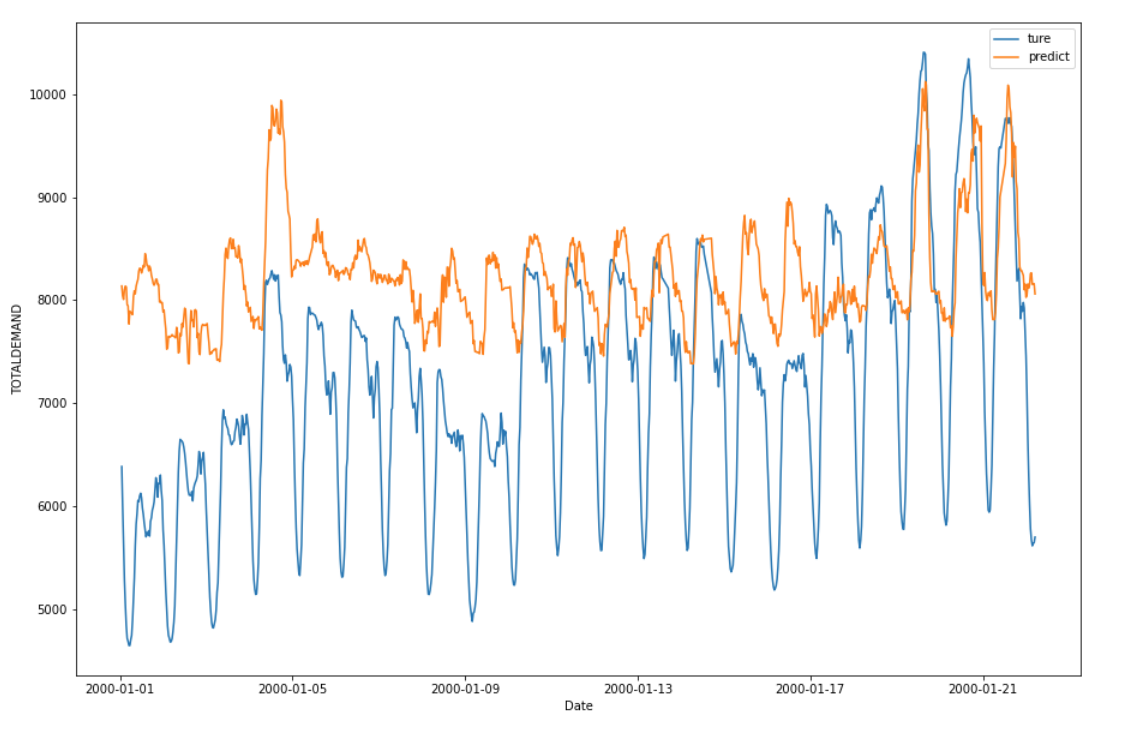
**4.003 \* X2**(Wet bulb temperature in degrees C)**^2 - 2449.702 \* X2 +**

**19.712 \* X3**( Dew point temperature in degrees C)**^2 + 705.632** **\* X3 + 11053.618**(interception)



After calculating the MAE, MSE, and R-square coefficients for the test set and training set, we can find that the quadratic regression model has a higher goodness of fit than the linear regression model by comparing these coefficients.

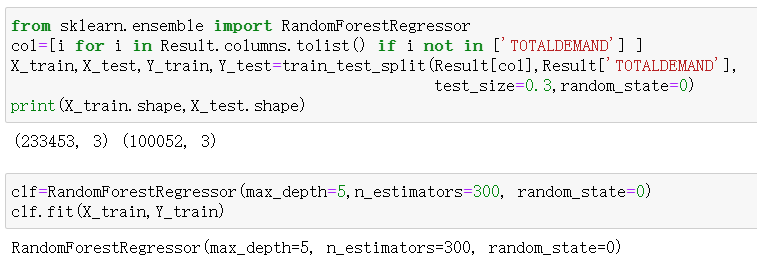


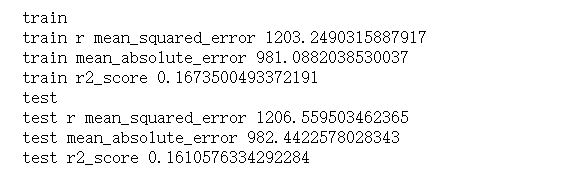
****

This is the comparison between the actual value and the expected value of the quadratic regression model in a line plot

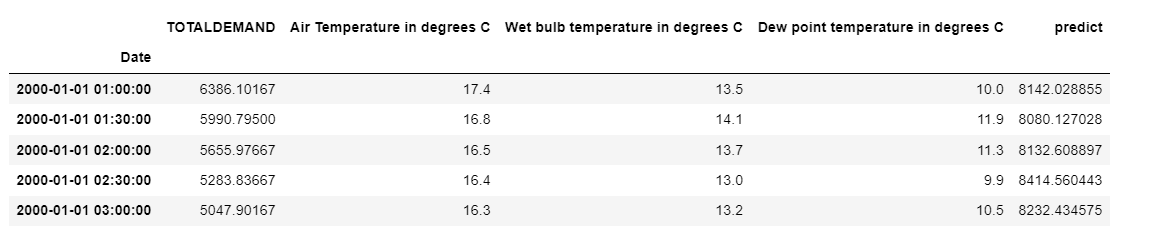
**Random forest (regression model)**

Although the quadratic regression is better than the linear regression model, by consulting the data, we know that the correlation coefficients of the regression model by random forest could be better for our data.



****

Comparing the R-squared coefficient of the regression model, this regression model has a higher r-squared coefficient than the above two models, which means that the model has a higher goodness of fit.





This is a comparison graph of the actual value and expected value of this model.

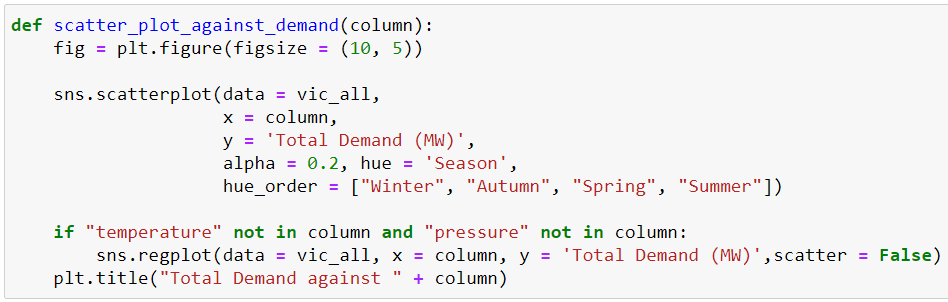
**Conclusions**

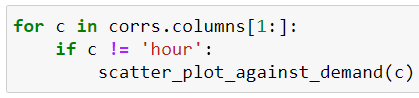
Did we end up meeting our initial aim of finding the correlation between energy usage and weather in Australia? From our analysis and modelling, we have found that demand has a low negative linear correlation with air temperature, however there is a clear parabolic relationship. What else did we find? We’ve found that the highest demand occurs in winter, whereas the 30-minute periods with the highest demand occur in Summer, that demand is lowest at around 4am and highest around 6pm, and how much impact exactly demand has on our other variables. We’ve also found that the accuracy of the model is not as high as you may expect it to be despite all these correlations. To conclude, from our work on this project we’ve found that although we can see some correlations, like demand increasing as temperature peaks and troughs and the relationship between demand and time of day, we can’t build a highly accurate model from just looking at the impact of temperature, as demand is impacted by a variety of different other variables as well.

**Appendix**

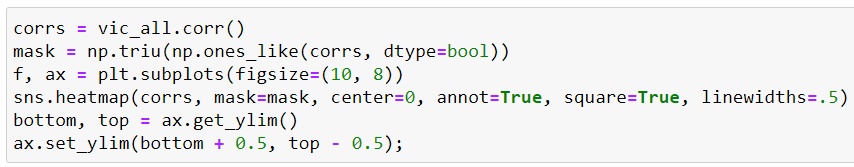
**Function to create line plots:**



**Function to create scatterplots:**



**Heatmap:**



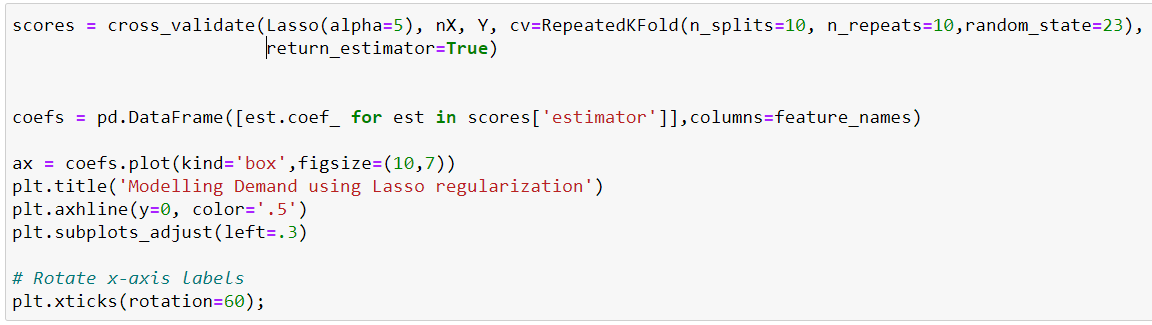
**Lasso regularization Model:**

Note: The code for the model without highly correlated dependent variables uses the same approach but without the regularisation and fewer variables.

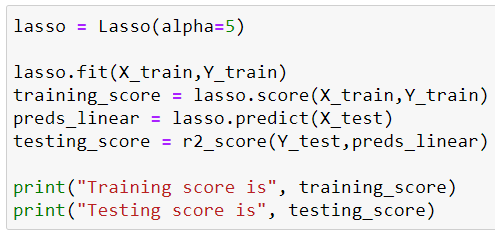
Finding optimal alpha parameter



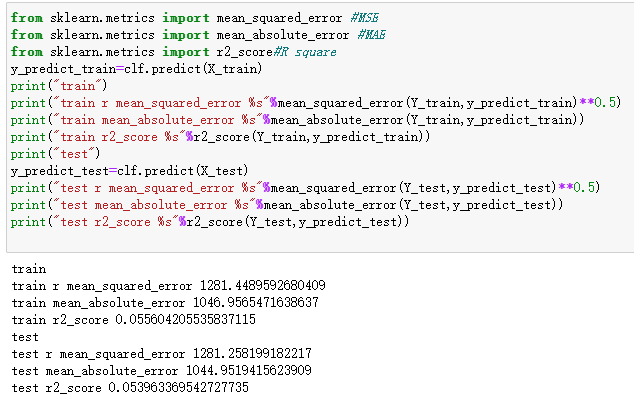
Plotting coefficients



Testing and training scores



Find the correlation coefficient



Plot line plots for expected and actual values

