Temperature and Energy Demand Project

ADS1002 Semester 2 2023
Final project report

	Name, Surname	Monash ID	Contribution (%)	Contribution (description)
1.	Aden Siau	33890277	20	Part 3
2.	Yohan nanayakkara	33890420	20	Part 2
3.	Angus Oldham	33154317	20	Part 4, part 5
4.	Raynal Goundar	34049932	20	Part 1
5.	Riley Harris			

Part 1: Project Description 220 - 250 words

Project Overview:

This comprehensive project delves deep into the intricate relationship between electricity demand and weather, focusing specifically on temperature fluctuations. The research utilises extensive datasets provided by the Australian Energy Market Operator and the Bureau of Meteorology's Automatic Weather Stations. The data is collected over two decades spanning from 2000 to 2019, this wealth of data serves as the foundation for rigorous analysis.

Project Background:

The project leverages the Australian Energy Market Operator's meticulous data, detailing energy demand in megawatts (MW) for each state. Simultaneously, temperature and humidity data, critical factors influencing energy consumption, were collected from the Bureau of Meteorology's Automatic Weather Stations. The dataset, recorded at a half-hour interval, paints a detailed picture of the dynamic relationship between weather conditions and electricity usage.

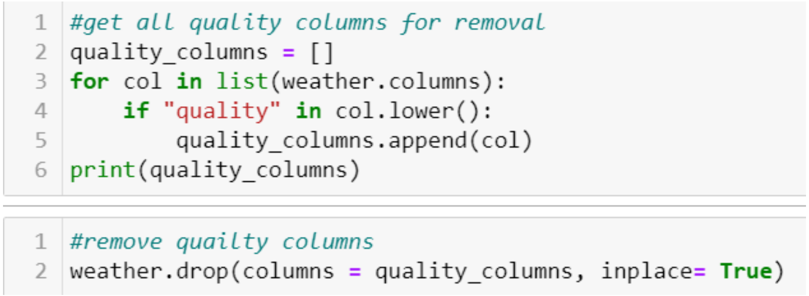
Project Approach:

Employing advanced statistical techniques, the study focuses on modelling to discern subtle shifts within the demand-temperature relationship. By identifying intricate patterns, the project aims to enhance the accuracy of predicting future energy demands, offering invaluable insights for energy planning. This predictive model holds significant implications for optimising energy distribution.

In addition to the existing datasets, the project incorporates detailed weather variables such as precipitation, air temperature, wet bulb temperature, dew point temperature, relative humidity, wind speed, wind direction, maximum wind gust speed, mean sea level pressure, and station level pressure. By integrating this granular weather data, the analysis aims to provide a holistic understanding of the multifaceted factors influencing energy demand patterns.

Part 2: Data preprocessing and manipulation:

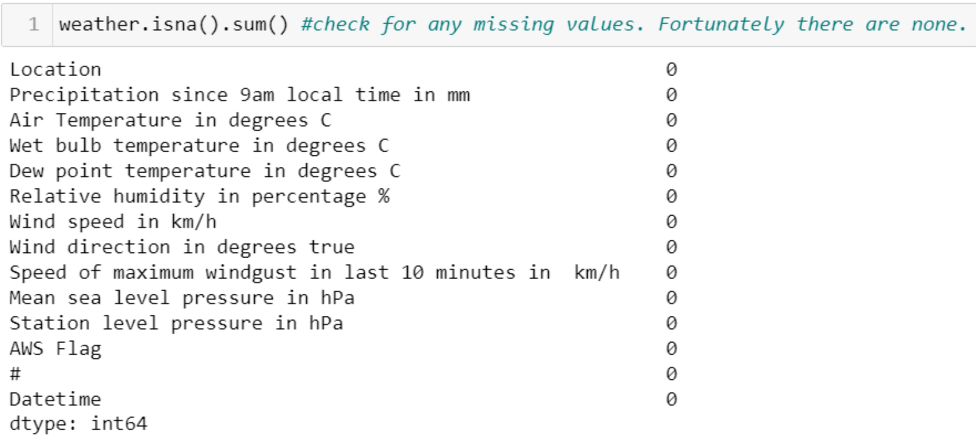
The main objective was to clean out both the weather dataset and the energy dataset in order to be able to smoothly merge the data frames and move on to exploratory analysis and modelling. During the preprocessing stage numerous tasks were undertaken such as removing unnecessary columns in the weather dataset such as ‘precipitation quality’ and ‘wind direction quality’ which had no relevance to the analysis undertaken.



```python
#get all quality columns for removal
quality_columns = []
for col in list(weather.columns):
    if "quality" in col.lower():
        quality_columns.append(col)
print(quality_columns)
```

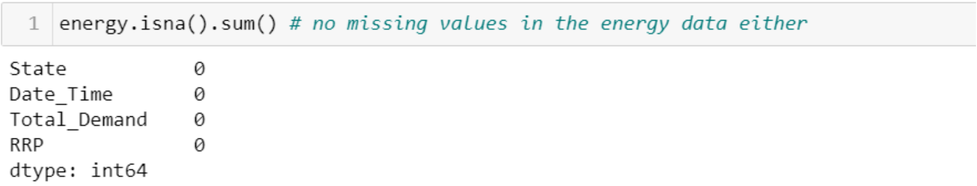
```python
#remove quailty columns
weather.drop(columns = quality_columns, inplace= True)
```

Checking for nan values in which we found out there were no nan values in all columns in both data frames which is a positive outcome. However, when attempting to convert some values to integers and floats, an error was encountered due to the presence of empty strings, so in order to move on these strings were identified and removed.



```python
weather.isna().sum() #check for any missing values. Fortunately there are none.
```

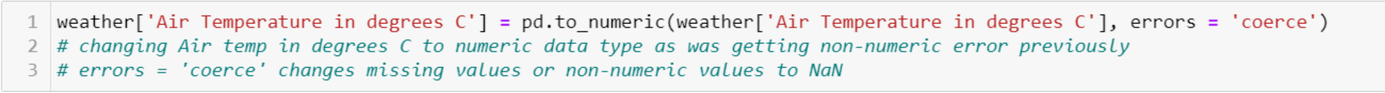
```
Location                                                0
Precipitation since 9am local time in mm                0
Air Temperature in degrees C                            0
Wet bulb temperature in degrees C                       0
Dew point temperature in degrees C                      0
Relative humidity in percentage %                       0
Wind speed in km/h                                      0
Wind direction in degrees true                          0
Speed of maximum windgust in last 10 minutes in  km/h   0
Mean sea level pressure in hPa                          0
Station level pressure in hPa                           0
AWS Flag                                                0
#                                                       0
Datetime                                                0
dtype: int64
```



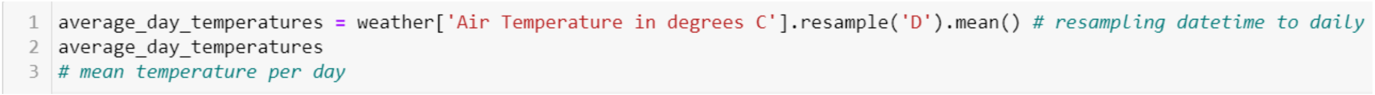
```python
energy.isna().sum() # no missing values in the energy data either
```

```
State           0
Date_Time       0
Total_Demand    0
RRP             0
dtype: int64
```

In the weather dataset, date and time columns were renamed and then merged into a single ‘Datetime’ column to allow time related analysis. Date type conversion was also utilised such as certain columns being converted to appropriate data types which includes conversion to numeric data types such as integers and floats, handling missing values by converting into nan values and converting columns representing time intervals to datetime objects to ensure data consistency.

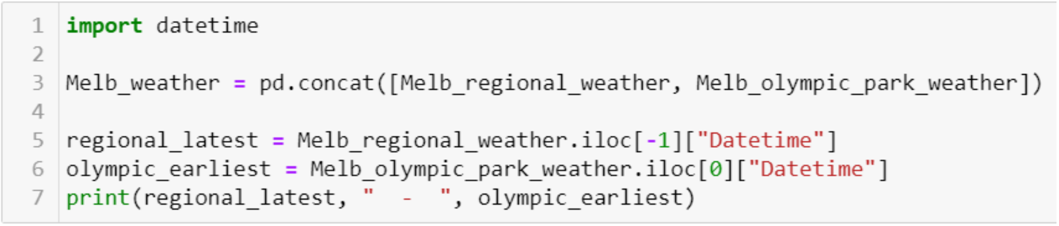


```python
weather['Air Temperature in degrees C'] = pd.to_numeric(weather['Air Temperature in degrees C'], errors = 'coerce')
# changing Air temp in degrees C to numeric data type as was getting non-numeric error previously
# errors = 'coerce' changes missing values or non-numeric values to NaN
```

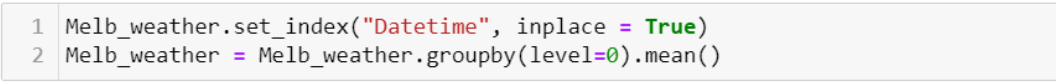


```python
average_day_temperatures = weather['Air Temperature in degrees C'].resample('D').mean() # resampling datetime to daily
average_day_temperatures
# mean temperature per day
```

Data manipulation was utilised in order to try to transform, alter and extract valuable information. We identified that there are two different weather stations named "Melbourne Regional " and "Melbourne Olympic Park," provided data over various dates, but their data was not continuous which was a problem. By conducting external research we discovered these stations were located approximately 2 km apart. So we combined these two stations and averages were calculated for overlapping timestamps. However, Melbourne’s data was rejected as it provided unnecessary complications when trying to merge and having to further subset the data.

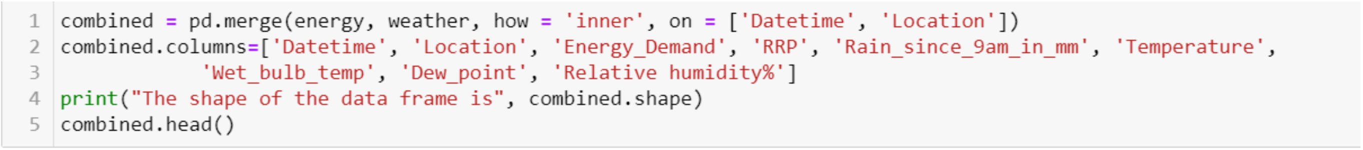


```python
import datetime

Melb_weather = pd.concat([Melb_regional_weather, Melb_olympic_park_weather])

regional_latest = Melb_regional_weather.iloc[-1]["Datetime"]
olympic_earliest = Melb_olympic_park_weather.iloc[0]["Datetime"]
print(regional_latest, "  -  ", olympic_earliest)
```



```python
Melb_weather.set_index("Datetime", inplace = True)
Melb_weather = Melb_weather.groupby(level=0).mean()
```

Now looking at the two data frames we got, one containing energy data organised by state and the other containing weather data organised by capital city, were merged by treating them as values from the capital city. This merge was possible because the data types and locations matched based on "Datetime" and "location”.



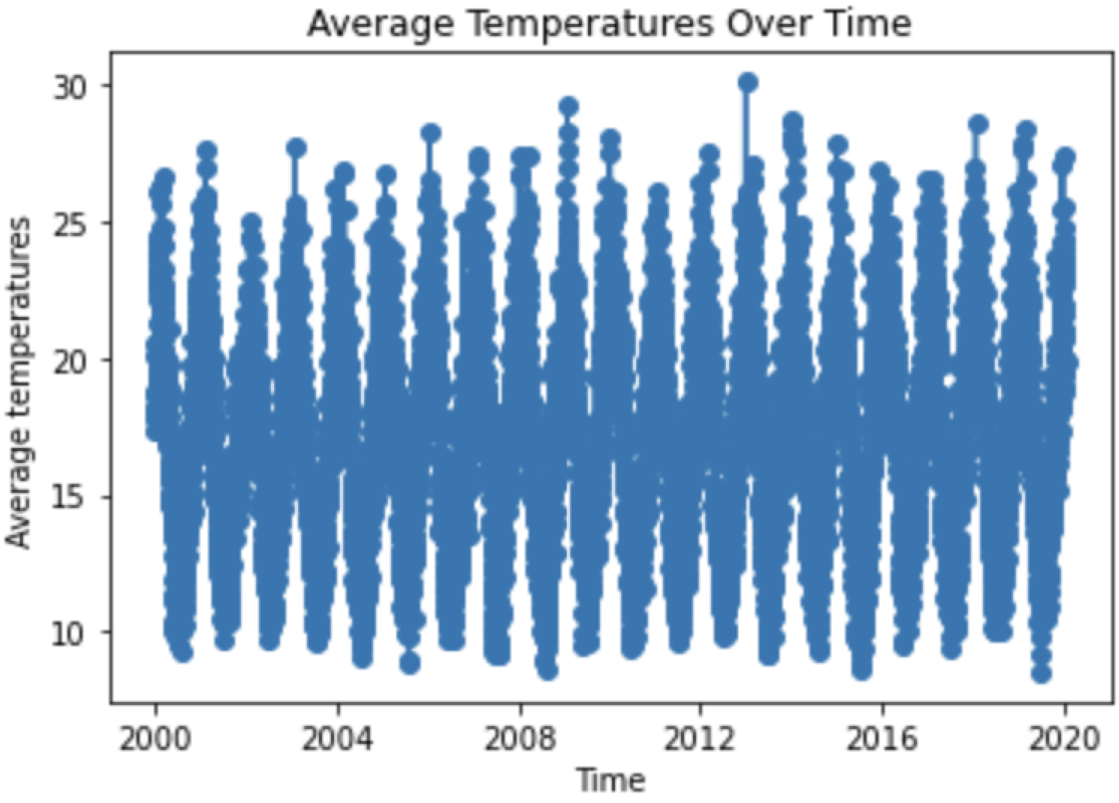
```python
combined = pd.merge(energy, weather, how = 'inner', on = ['Datetime', 'Location'])
combined.columns=['Datetime', 'Location', 'Energy_Demand', 'RRP', 'Rain_since_9am_in_mm', 'Temperature',
             'Wet_bulb_temp', 'Dew_point', 'Relative humidity%']
print("The shape of the data frame is", combined.shape)
combined.head()
```

In order to move onto exploratory analysis, the weather data was resampled to calculate average daily and 30-minute temperatures at specific time intervals. Data subsets were created by filtering rows based on specific conditions, such as selecting a particular season which would be beneficial when trying to visualise seasonality. Overall, the data preprocessing and manipulation techniques employed in this project helped effectively transform the datasets received into a smooth and usable format which can be used later on as the base for our exploratory analysis and modelling.

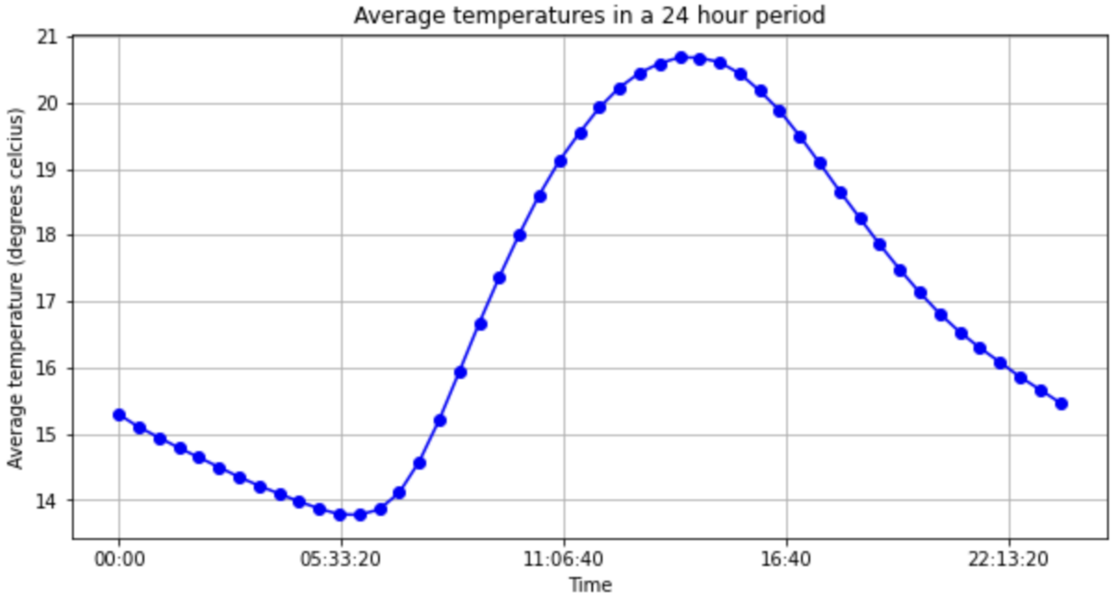
Part 3: Exploratory data analysis and significant conclusions:

Average daily temperature:

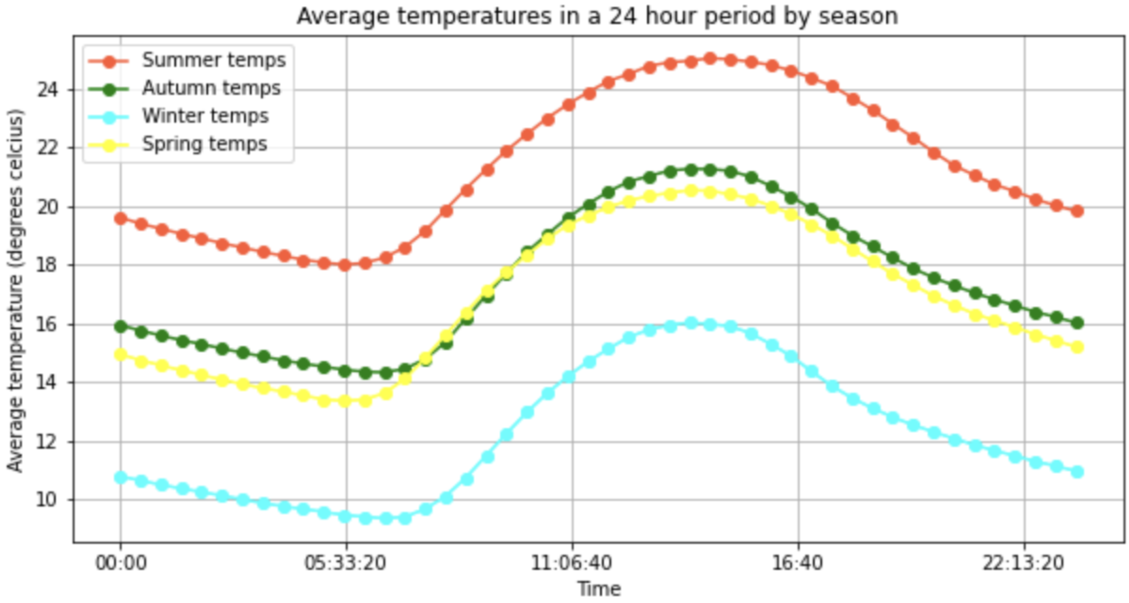
To investigate the average daily temperature, each Temperature from every data point was plotted against Time to gain an overall perspective of how temperatures have changed over time. From the graph, a seasonal pattern is evident as the peaks and troughs occur at the same intervals. However, due to the size of the data, the graph is extremely cluttered, making it difficult to obtain informative observations, and to observe patterns. Further, it was clear that Australian Temperature has remained relatively constant other than the seasonal fluctuations. This is evident as the average temperature has neither increased or decreased from when the data begins (2000) and when it ends (2020).

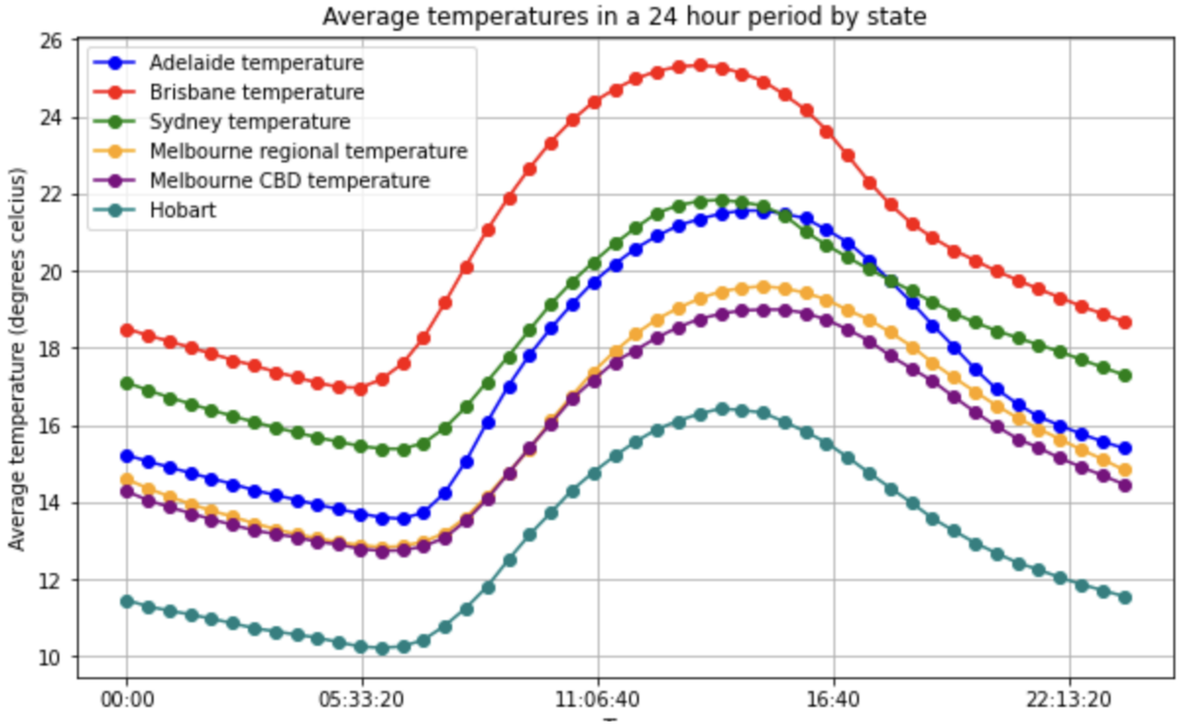


To gain a better understanding of how temperature changes according to time on a smaller scale (daily), an average temperature was calculated for each 30 minute time interval. This revealed a clear pattern for how the temperature fluctuates in a day, with the average being taken from all states and all seasons. Upon graphing these averages, a clear pattern was formed. On average, the minimum temperature in a day occurs at around 5:30 am, while the maximum temperature occurs at around 2:30 pm. (See graphs below)



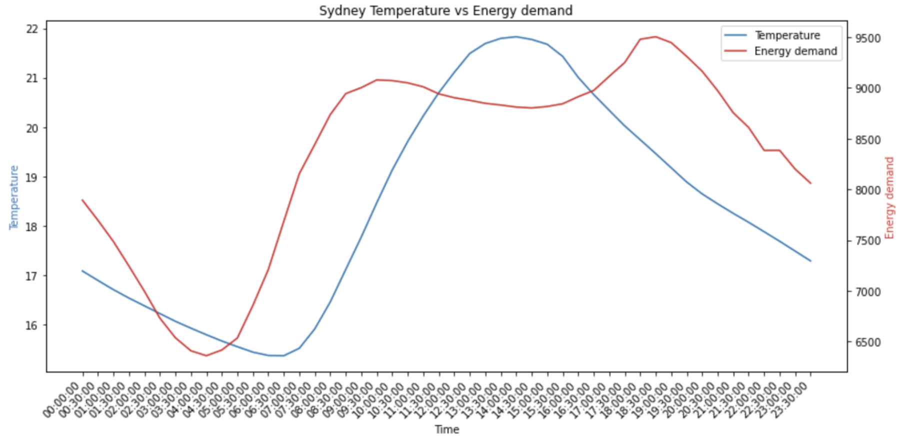
However, this average was deemed to be too broad, so the data was further subsetted by factors which would possibly have significant impacts on temperature. The data was subsetted into seasons and cities, which would give us an idea of how temperature changes according to time of the year, and geographically. From the graph subsetted by season, we can see that all 4 lines representing the temperature in different seasons follow the same shape as the overall average. However, the lines are translated up and down to coincide with the average temperatures in that season. This happened as expected, with Winter (blue line) having the lowest average temperatures, Summer (red line) having the highest average temperatures, and Autumn/Spring being in the middle range of average temperatures. In the graph subsetted by Cities, the temperatures clearly follow the same structure. The vertical differences between the lines were attributed to northern cities experiencing warmer temperatures, while Southern coastal cities had cooler temperatures. An extreme case of this was Hobart and Brisbane which had the lowest and highest average temperatures respectively, and are also the furthest South and North respectively. (See graphs below)

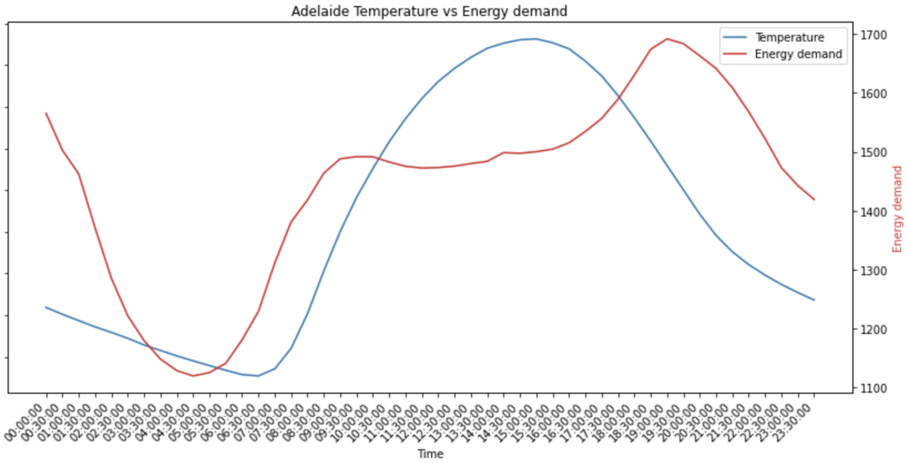


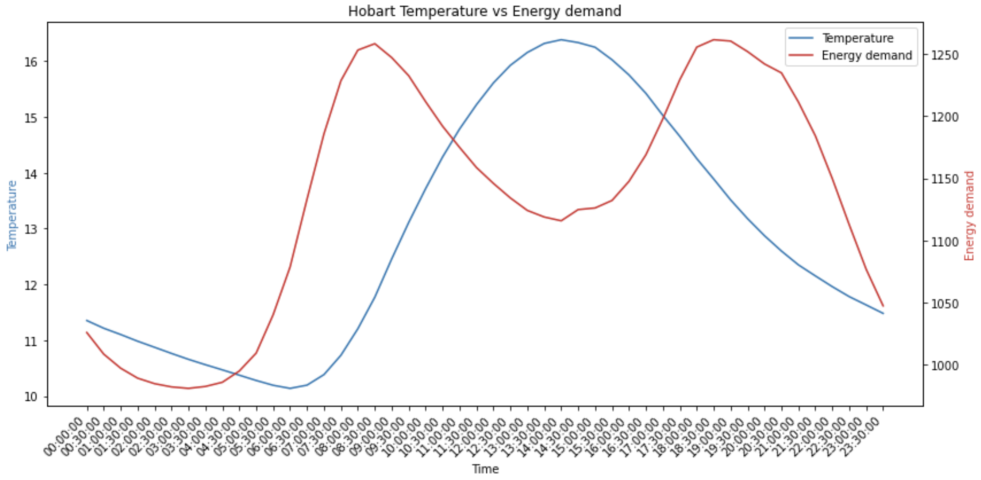


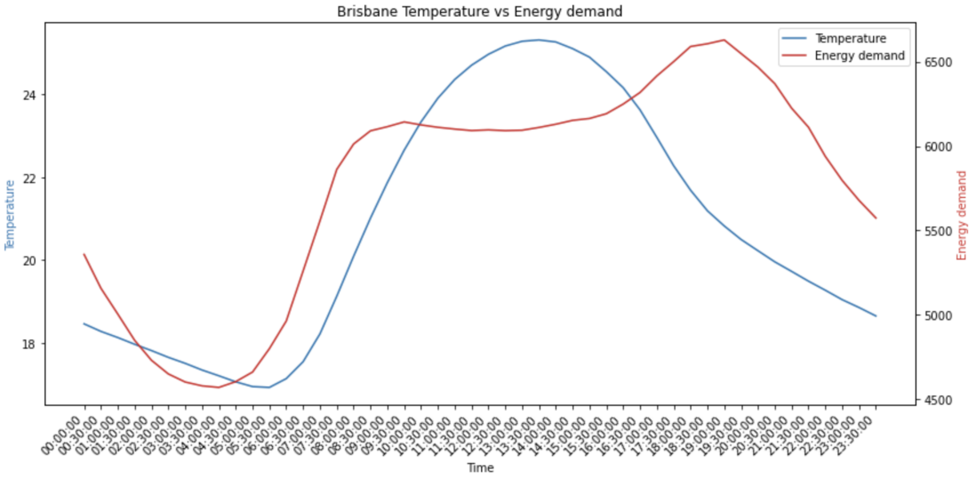
Temperature vs Energy demand: How does temperature affect Energy demand?

The main purpose of this project was to investigate how temperature affects the energy demand in Australia. To achieve this, the data was subsetted into cities once again. All 4 cities show similar graphs with energy demand and temperature following the same pattern. The only outlier to this was Hobart which had a significant drop in energy demand at around 2:30 pm. This could be due to Hobart's much lower population, and could also be due to its geographical location. (See graphs below)

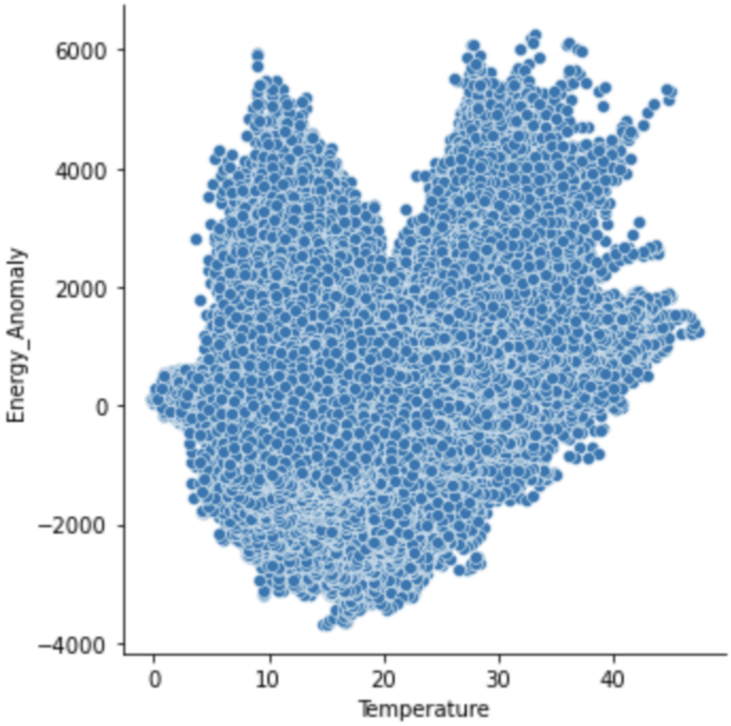

Sydney Temperature vs Energy demand
Temperature
Energy demand
22
21
20
19
18
17
16
9500
9000
8500
8000
7500
7000
6500
00:00:00 … 23:30:00
Time

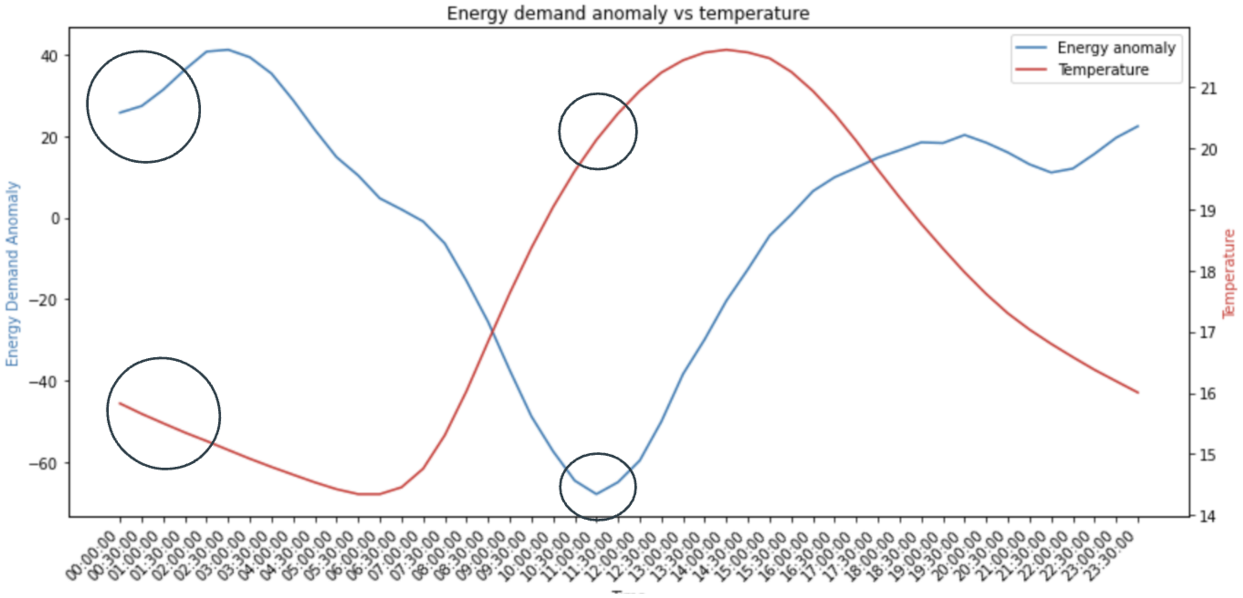

Adelaide Temperature vs Energy demand
Temperature
Energy demand
22
21
20
19
18
17
16
15
14
1700
1600
1500
1400
1300
1200
1100
00:00:00 … 23:30:00
Time


Hobart Temperature vs Energy demand
Temperature
Energy demand
16
15
14
13
12
11
10
1250
1200
1150
1100
1050
1000
00:00:00 … 23:30:00
Time


Brisbane Temperature vs Energy demand
Temperature
Energy demand
24
22
20
18
6500
6000
5500
5000
4500
00:00:00 … 23:30:00
Time

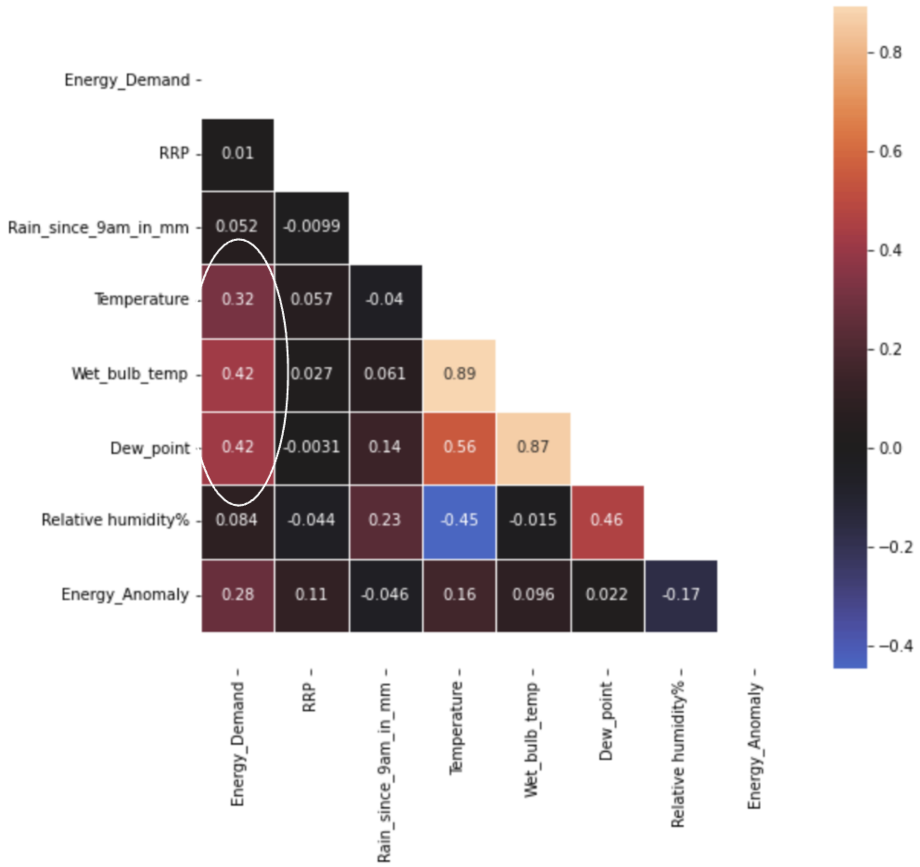
This same relationship was explored in a more general representation through graphing Average Energy Anomalies and Temperature over time (See graphs below). From the graph, it can be observed that the graphs are almost reflections of each other and that there is a clear U shape pattern. This tells us that when the temperature is low (lower than 20 degrees), the Energy Anomaly is high, telling us that the energy demand at that time is more than average. A good explanation for this observation is that when it is cold, houses and other buildings are all using heating appliances, using a significant amount of energy. This works both ways, with the Energy Anomaly also being high when the temperature is also high (due to more cooling appliances being utilised). Further, when the Energy Anomaly is at its lowest, the temperature is at around 20 degrees, which is a temperature at which most feel that neither cooling or heating is necessary.





Correlation heatmap:

A correlation heatmap was created to explore which variables had a strong relationship with Energy Demand and Energy Anomaly. From the heatmap, we can see that all correlations to Energy Demand and Energy Anomaly are quite poor (all below 0.5). However 3 variables which had the strongest correlation with Energy Demand were Temperature, Wet bulb temperature and Dew point. This suggests that the relationship between Energy demand and other possible variables is likely to not be linear.

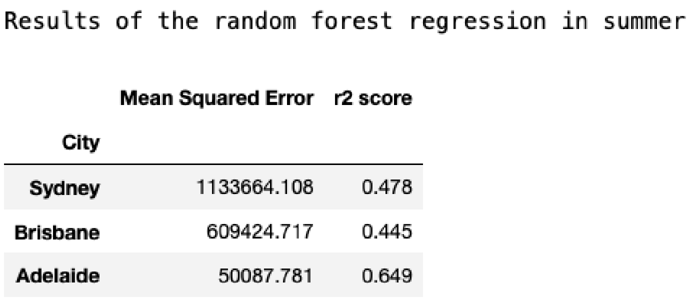


Part 4: Modelling Description

Using data from Sydney, Adelaide and Brisbane, a variety of models were attempted with the intention of predicting the energy demand anomaly based primarily on temperature and humidity. As seen during the exploratory data analysis, it is evident that temperature cycles greatly not only by season but also by city. In order to help combat this, separate models were created for each season and each city, using all data from all different times of the day, for every day of the year.

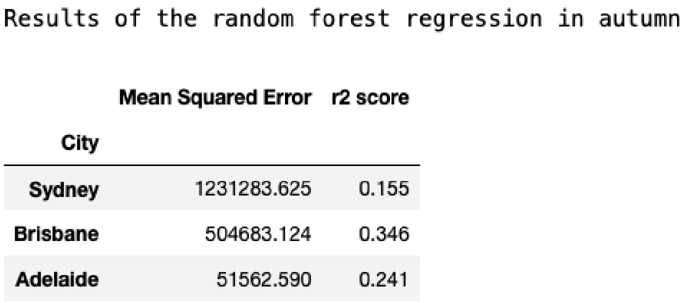
Random forests were overall the most successful model that was implemented. Using temperature and humidity, an indirect measurement of rainfall, the models varied greatly in success across the seasons, with summer being the easiest to predict and winter being the hardest. This is likely due to the fact that in Australia summers we see extremely high temperatures, where the demand for air conditioning is immense, then when a cool change arrives everyone turns the air conditioning off. This provides obvious trends and a large variation for the model to observe. In contrast, in winter the temperature in Australia is more constant, as although there is the odd warmer day, temperatures rarely reach extreme lows, causing temperature to have less significance on energy demand.

Upon investigating the feature importance, temperature was by far the most important, with values usually close above 0.8. This further reinforces the conclusion that temperature and energy demand are highly linked.



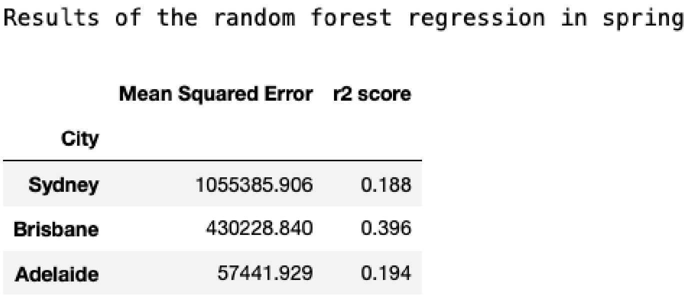
```
Results of the random forest regression in summer
```

City	Mean Squared Error	r2 score
Sydney	1133664.108	0.478
Brisbane	609424.717	0.445
Adelaide	50087.781	0.649



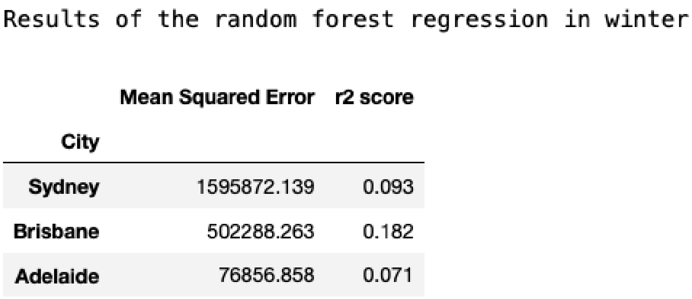
```
Results of the random forest regression in autumn
```

City	Mean Squared Error	r2 score
Sydney	1231283.625	0.155
Brisbane	504683.124	0.346
Adelaide	51562.590	0.241



```
Results of the random forest regression in spring
```

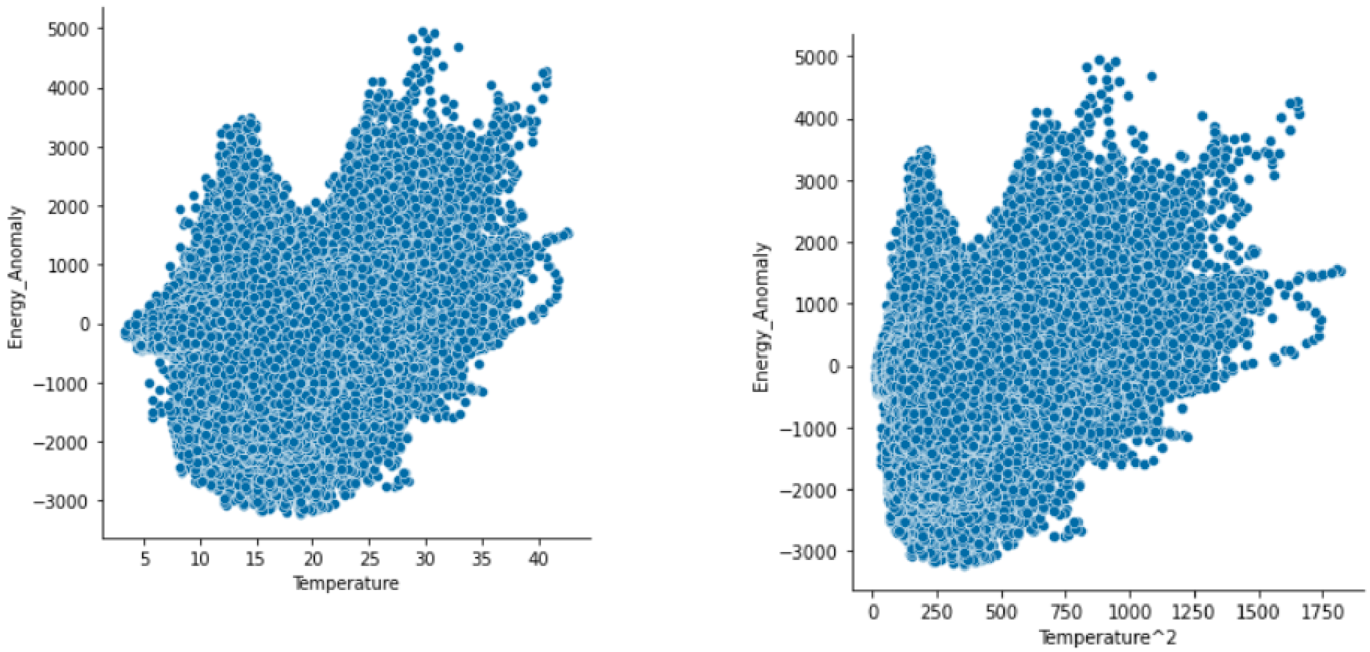
City	Mean Squared Error	r2 score
Sydney	1055385.906	0.188
Brisbane	430228.840	0.396
Adelaide	57441.929	0.194



```
Results of the random forest regression in winter
```

City	Mean Squared Error	r2 score
Sydney	1595872.139	0.093
Brisbane	502288.263	0.182
Adelaide	76856.858	0.071

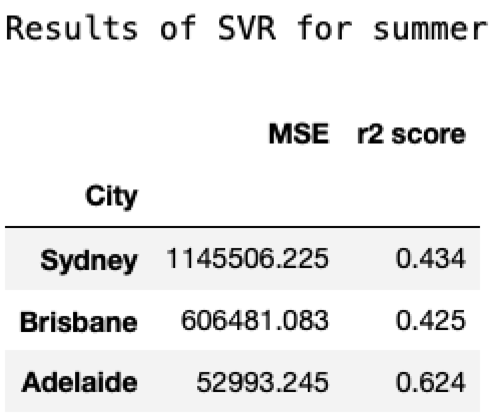
A rather ineffective model proved to be multilinear regression. Since temperature and energy demand have a quadratic relationship, simply fitting a linear coefficient to temperature would never produce an accurate model. To fix this, a temperature squared column was created, creating a linear relationship between this column and energy demand (*see below*). However, this process magnified any outliers in the data, as the data appears more spread out, decreasing the model's accuracy.



The spread of the temperature data (left) compared to the spread of the temperature^2 data (right). Note that neither shape follows a linear trend.

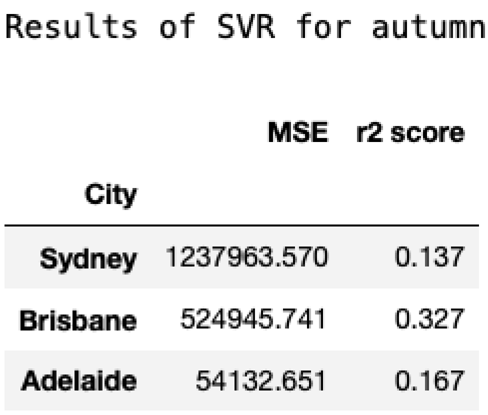
The final model attempted, which achieved reasonable success, was support vector regression. Once the data was subset by city and by season, each model had a dataset of over 80 000 rows, which proved far too much for SVR to handle. Therefore, the data was sampled considerably in order to achieve a reasonable run time, down to approximately 4 000 rows. This proved a much more manageable task for the model.

Ultimately, summer was once again the easiest to predict, likely for the same reasons and winter proved the hardest. While accuracies were comparable to random forest regression, it proved slightly less effective and the additional computation time did not produce greater results.



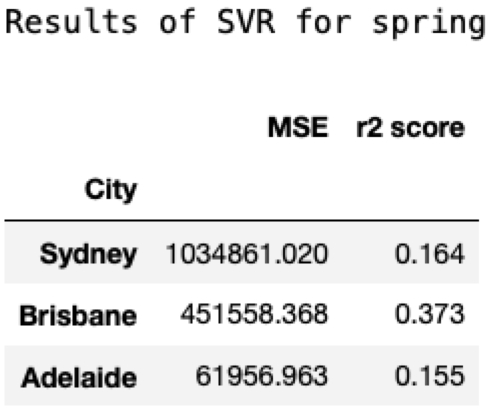
```
Results of SVR for summer
```

City	MSE	r2 score
Sydney	1145506.225	0.434
Brisbane	606481.083	0.425
Adelaide	52993.245	0.624



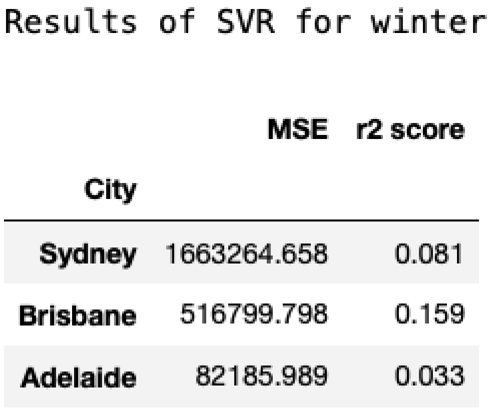
```
Results of SVR for autumn
```

City	MSE	r2 score
Sydney	1237963.570	0.137
Brisbane	524945.741	0.327
Adelaide	54132.651	0.167



```
Results of SVR for spring
```

City	MSE	r2 score
Sydney	1034861.020	0.164
Brisbane	451558.368	0.373
Adelaide	61956.963	0.155



```
Results of SVR for winter
```

City	MSE	r2 score
Sydney	1663264.658	0.081
Brisbane	516799.798	0.159
Adelaide	82185.989	0.033

Overall, to improve the modelling more careful subsetting of the data is needed earlier on in the process. This should reduce the spread of the data to better predict the demand anomaly. Moreover, additional features must be investigated to determine the extent of their effect, such as days of the week, school and public holidays, if population wide sporting events such as the Olympics are on, to name a few. By running a random forest using only these types of variables, without temperature, the feature importance will provide great detail about what other variables have an effect on energy demand.

Part 5: Conclusion to analysis

A clear parabolic relationship between temperature and energy demand has been established, linked primarily to the use of heaters in cooler temperatures and air conditioning in warmer temperatures. In addition, since they are closely related, and temperature follows a diurnal cycle, energy demand ebbs and flows throughout the day as the general population sleeps and wakes. Temperature also cycles seasonally, so the energy demand is always closer to the mean in more moderate climates. As a result of this, modelling during spring and autumn was difficult, as there was less of a trend for the models to find.

Through the modelling it was found that given the more comprehensive spread of data, meaning that both temperatures and the demand anomaly varied to greater extents, summer was by far the easiest season to predict, and the colder months were more difficult. In order to increase the accuracy during these times a more extensive dataset needs to be used in order to determine what other variables have an effect on energy demand.

References

BOM. (2019). *Australia's official weather forecasts & weather radar - Bureau of Meteorology*. Bom.gov.au. http://www.bom.gov.au