Data Analytics: Assignment 3

by Anthony Rizkallah

Table of contents

1 Python Packages

Pandas

Matplotlib

Numpy

Datetime

Scipy

Calendar

2 Question 1

2.1 Steps

After importing the EirGrid intraday 15-minute energy demand dataset, there are five missing demand values (due to load shedding) at 2014-03-30 01:00, 2014-03-30 01:15, 2014-03-30 01:30, 2014-03-30 01:45, 2014-08-14 14:15. Hence, to handle empty values, linear interpolation can be used

$$y = y_1 + (x - x_1)\frac{y_2 - y_1}{x_2 - x_1} \tag{1}$$

After handling NaN values, the time series can be plotted.

2.2 Results

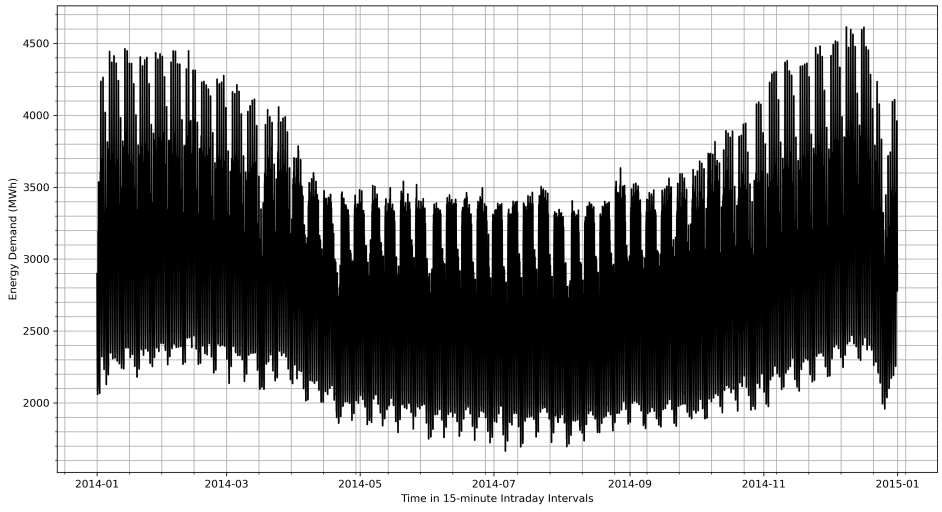

Figure 1. Energy Demand (MWh) over 15-minute Intraday Intervals

2.3 Inference

The time series in Figure (1), shows seasonality in energy demand over time, but due to the high volume of data, it is unclear how the seasonality is occuring. Hence, further exploration is required.

(See Section (3)).

3 Question 2

3.1 Steps

The Autocorrelation Function measures how values at time t in a time series are correlated with previous values. In other words, how data is correlated with its past data which would indicate an inherent seasonality attribute. For this case, the autocorrelation is being plotted for the actual energy demand, meaning that load values in the future will be correlated with past load values.

$$\text{Lags}_{10\,\text{Days}} = 10\,\text{days} \times \frac{4\,\text{datapoints}}{1\,\text{hour}} \times \frac{24\,\text{hours}}{1\,\text{days}} = 960\,\text{datapoints}$$

3.2 Results

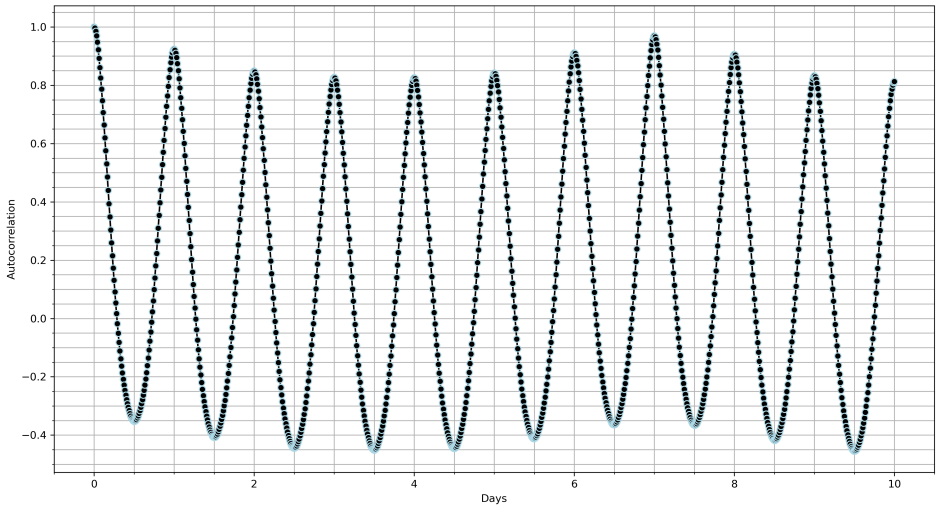

Figure 2. Autocorrelation of Energy Demand

3.3 Inference

The autocorrelation graph in Figure (2) shows daily seasonality. Graphically, the energy demand returns to approximately the same value every day with some exceptions.

4 Question 3

4.1 Steps

Creating a time of year variable is a useful method to identify demand at a certain percentage of the year by normalizing the time of the year between 0 and 1, and then plotting demand over the

normalized time. The method of calculating the time of year variable is as follows

$$\text{Day of Year} \;=\; d$$

where $d \epsilon [1, 365]$. Then,

$$\text{Fraction of Day} \;=\; \frac{h}{24} + \frac{m}{24 \times 60}$$

where h is the hour of day $[0, 24]$ and m is the minute $[0, 60]$. Finally, the time of year variable is

$$\text{Time of Year Variable} \;=\; \frac{\text{Day of Year} + \text{Fraction of Day} - 1}{365}$$

where the -1 is used to shift the date values by 1 due to day 1 of the year having a value of 1 and so on.

4.2 Results

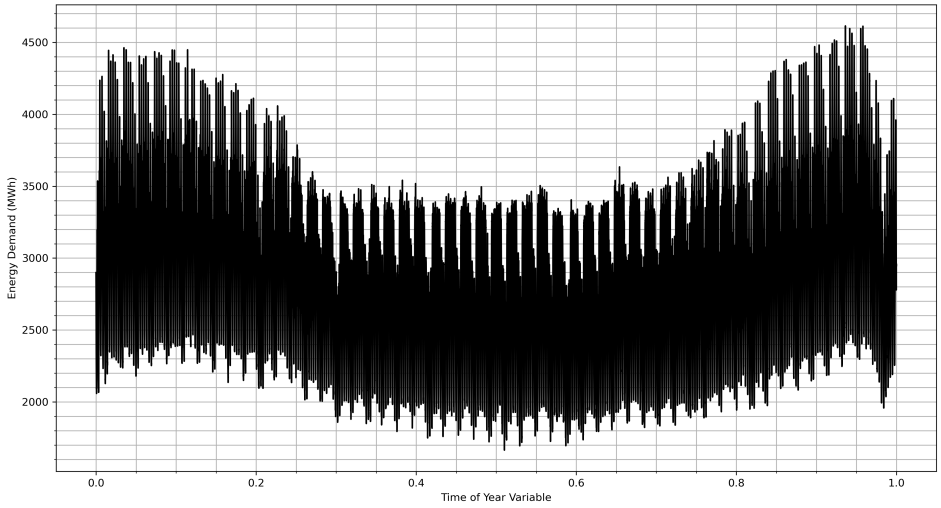

Figure 3. Energy Demand (MWh) over Time-of-Year Variable

4.3 Inference

The graph in Figure (3) is similar to the graph in Figure (1) with a different x-axis. The time of year variable is used to wrap time around a circle. However, the graph in this section does not represent that because in the case of graphing Time of Year Variable over Time in Years does not show demand.[1]

1. Justification of Lecture 6 A3Q3 explanation.

5 Question 4

5.1 Steps

To find the average demand for each month of the year, sum the demand over all the data points of each month and divide by the sum of days in that month

$$\text{Average}_{\text{Demand of Month}} = \frac{\sum D_i}{N}$$

where D_i is the demand at each data point i and N is the number of days in that month.

5.2 Results

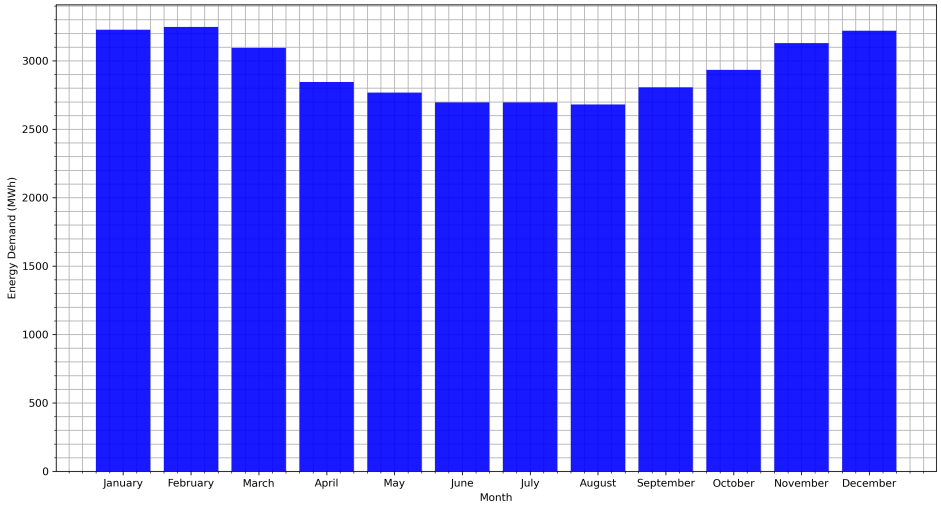

Figure 4. Average Monthly Energy Demand (MWh)

5.3 Inference

In November, December January, February, and March, energy demand is the highest in the year. The aformentioned months could be categorized as Winter demand. In April, May, June, July, August, September, and October, the demand is lower than the Winter and can be categorized as Summer.

The energy demand in the winter is higher than the summer due to cold weather increasing energy consumption for heating purposes.

6 Question 5

6.1 Steps

To find the average demand for each hour of the day, sum the demand over all the data points of

each day and divide by the sum of datapoints in each day

$$\text{Average}_{\text{Demand of Hour}} = \frac{\sum \text{DP}_i}{N}$$

where DP_i is the demand at each data point i in a day and N is the number of days in year (365 or 366 in a leap year).

6.2 Results

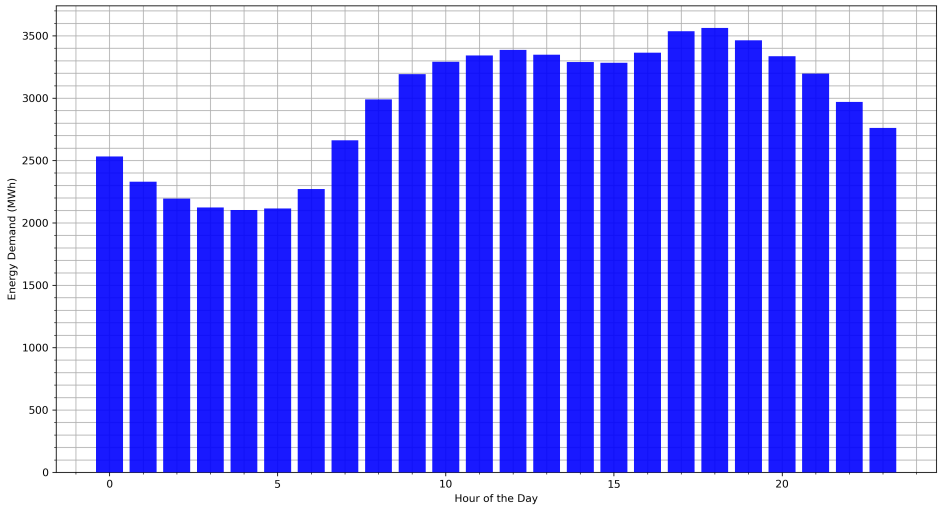

Figure 5. Average Hourly Energy Demand (MWh)

6.3 Inference

The energy demand varies over a 24-hour period such that overnight demand is lower than day demand, resembling the "Duck Curve". Moreover, as people wake up and head to work at around 6-7am, the energy starts ramping up. Then, at around 5pm (18:00), the energy demand peaks, likely due to people coming back from work and using more energy. Finally, during the evening, the energy demand decreases.

7 Question 6

7.1 Steps

To find the average demand for each day of the week, sum the demand over all the data points of each day of the week and divide by the sum of datapoints in each day of that week

$$\text{Average}_{\text{Demand of Day}} = \frac{\sum \text{Day}_i}{N_{\text{weeks}}}$$

where Day_i is the demand at each data point i in each day of the week and N is the number of weeks in a year- 52.

7.2 Results

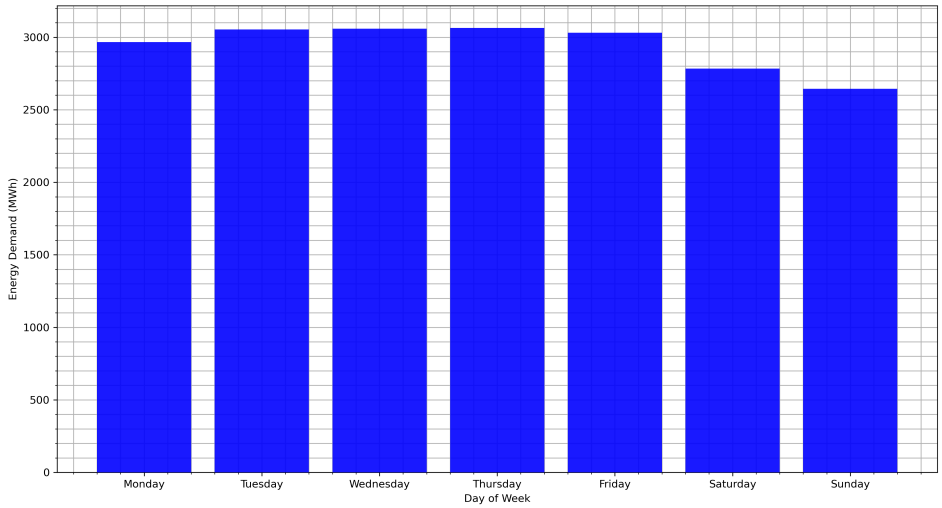

Figure 6. Average Energy Demand (MWh) of Each Day of the Week

7.3 Inference

The demand in Figure (6) shows the average demand on each day of the week. Observationally, the demand during weekdays is approximately uniform within a 100 MWh difference. However, during weekends, the demand drops approximately 200 MWh on Saturday and about 350 MWh on Sunday. That could be due to a drop in productivity (on a per capita basis), meaning, more time for activities, less office time, etc.

8 Question 7

8.1 Steps

To find the demand time-series for each day of the week, sum the demand over all the data points of each hour of the day of the week and divide by the sum of datapoints in each day of that week

$$\text{Average}_{\text{Demand of Day}} = \frac{\sum \text{Time}_{\text{Day}_i}}{N_{\text{weeks}}}$$

where Day_i is the demand at each data point i in each day of the week and N is the number of weeks in a year- 52, and Time is each datapoint.

8.2 Results

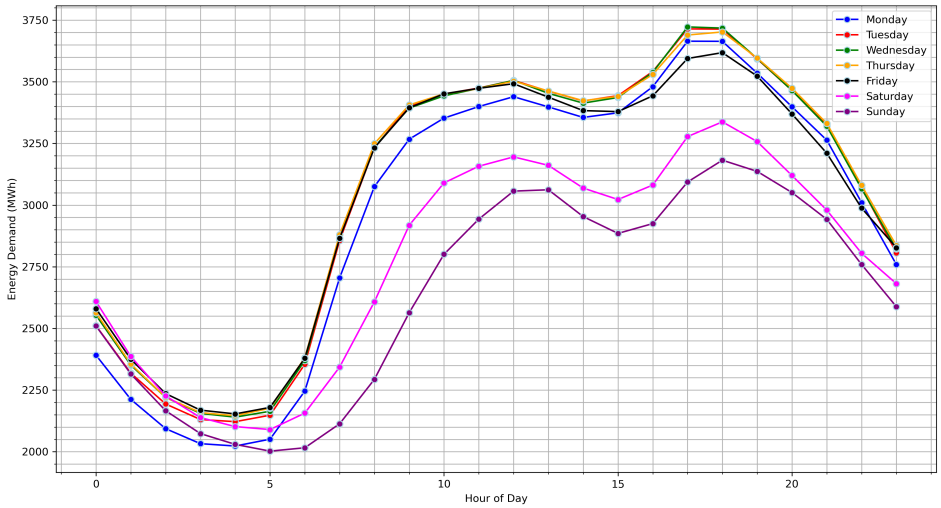

Figure 7. Energy Demand (MWh) Time-Series for Each Day of the Week

8.3 Inference

The graph in Figure (7) shows the demand time-series of each day of the week, sort of similar to the bar graph in Figure (5). From the inference of Section (6.2), the energy demand varies over a 24-hour period such that overnight demand is lower than day demand, resembling the "Duck Curve". Moreover, as people wake up and head to work at around 6-7am, the energy starts ramping up. Then, at around 5pm (18:00), the energy demand peaks, likely due to people coming back from work and using more energy. Finally, during the evening, the energy demand decreases.[2]

For the difference between weekdays and weekends, the inference from Section (7.2) - Figure (6) can be verified in this section. The energy demand during weekdays is higher than that of weekends (7).

9 Question 8

9.1 Steps

To determine whether there is a statistically significant difference between demand during the weekend (Saturday and Sunday) and during the working week (Monday through Friday), a statistical hypothesis test, such as a t-test, can be used to reach a conclusion. So, first

H_0 : There is no difference between demand during weekend and weekdays

H_1 : There is a difference between demand during weekend and weekdays

Next, a two-tailed test is appropriate because the goal is whether there is or isn't a difference

2. Section (6.2)

between the two energy demand vectors, and this could happen on either side ($\pm Z$). The t-test is

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Moreover, for a two-tailed test, there are twice as many critical regions because it is in both directions. This means that $\frac{\alpha}{2}$ should be considered rather than α.

Therefore, to prove/disprove the Null Hypothesis, the idea is

$$\begin{aligned} p &< \frac{\alpha}{2} \\ 2p &< \alpha \\ 2P(\bar{x} > \mu_0) &< \alpha \end{aligned}$$

9.2 Results

P-Value	0.0

Table 1. P-Value of Difference Between Weekday and Weekend Energy Demand (T-test)

9.3 Inference

The Null Hypothesis H_0 that there is no difference between demand during weekend and weekdays as evident by the p-value of the two-tailed t-test. Hence, the alternative hypothesis H_1 is accepted, and the difference between weekday and weekend energy demand is statistically significant.

10 Question 9

10.1 Steps

The persistence model is a simple yet effective benchmark for time-series forecasting, particularly for data that exhibits minimal short-term fluctuations. It assumes that the best predictor for a future value is the most recent observed value, making it a useful baseline for evaluating the performance of more complex forecasting models. The persistence forecast for a given time step t and lead time k is given by:

$$\hat{y}(t+k) = y(t) \tag{2}$$

where $\hat{y}(t+k)$ is the predicted value and $y(t)$ is the actual observed value at time t. To implement the persistence model, the dataset is first divided into two halves: the first half is used for training, while the second half is used for evaluation. For each forecast horizon up to a full day ahead (96 steps in a 15-minute interval dataset), predictions are generated using the last available observation. The forecasted values are then compared to the actual values to compute performance metrics.

In this case, Mean Absolute Error (MAE) are calculated across all forecast horizons. MAE is the average absolute difference between predicted and actual values. Finally, the error values are plotted against forecast horizons to analyze how prediction accuracy degrades over time.

10.2 Results

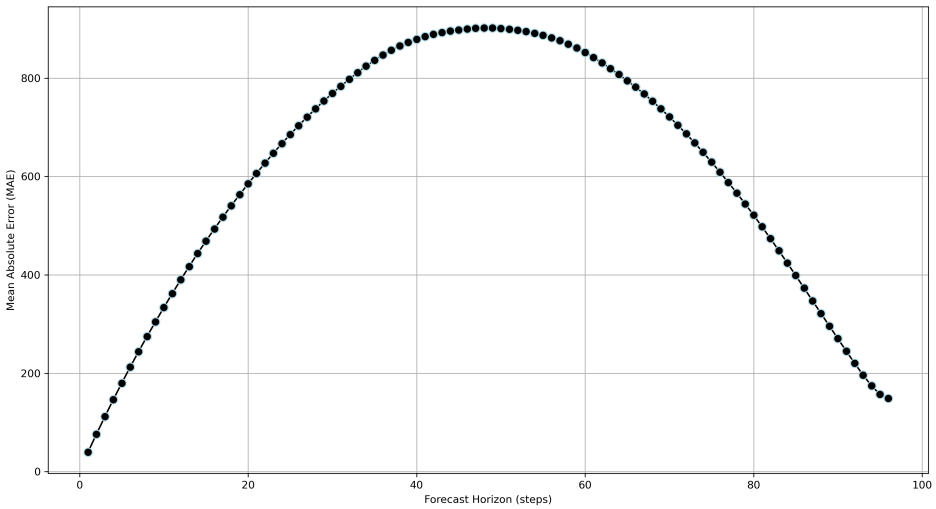

Figure 8. MAE Over Forecast Horizons (steps)

11 Question 10

11.1 Steps

In this case, Mean Absolute Percentage Error (MAPE) are calculated across all forecast horizons. MAPE expresses this error as a percentage of the actual values. Finally, the error values are plotted against forecast horizons to analyze how prediction accuracy degrades over time.

11.2 Results

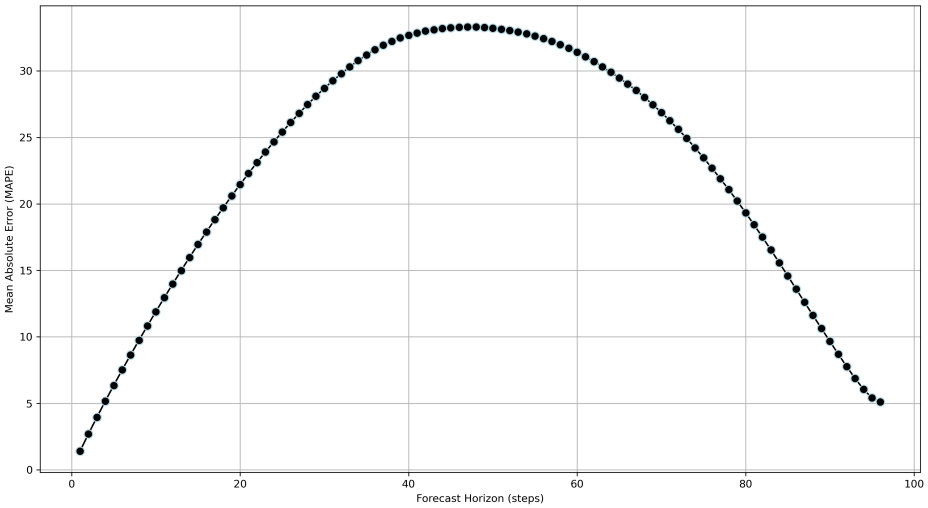

Figure 9. MAPE Over Forecast Horizons (steps)

11.3 Inference

First, the obvious observation is the hump-shaped curve of MAE and MAPE that starts at a value close to 0 and then after half the forecast horizon MAE and MAPE decrease again. From Figure (2), there was an obvious daily seasonality in the time series, and this is evident over the 96 step-horizon in Figures (8) and (9), which are also showcasing daily seasonality. This means that a persistence model is accurate in predicting daily demand with relatively low error.