Exploratory Data Analysis and Visualisation

CID: 06000562, CID: 06035448, CID: 06013608

March 4, 2025

1 Univariate EDA

Upon checking for any records with missing information in the dataset, we find only one record with missing information out of all 1,461 records (date = 2024-02-29) likely due to the system not collecting data on leap year days. We will therefore omit this record from the full dataset prior to further analysis.

In this section, we focus on national_demand. Similar analysis was done for other columns in the complete Rmd report where plot each time series in the dataset to gain a preliminary understanding of its structure and patterns. Figure 1 shows that national_demand has a seasonal pattern with a negative overall trend.

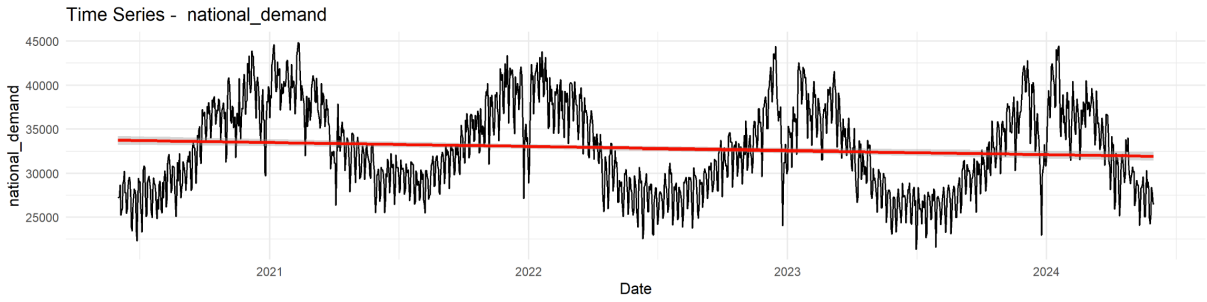

Figure 1: National demand time series

Figure 2 and 3 highlights distinct variations in demand across months and weekdays respectively, with significantly higher consumption on the winter months compared to summer months, and weekdays compared to weekends.

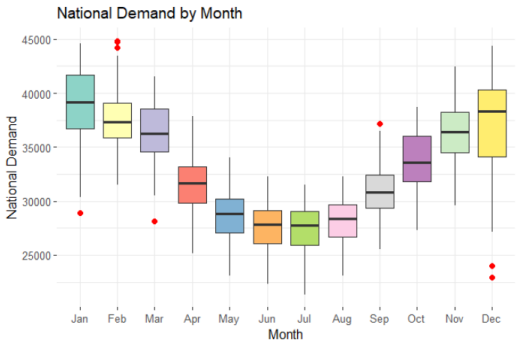

Figure 2: National demand by Month

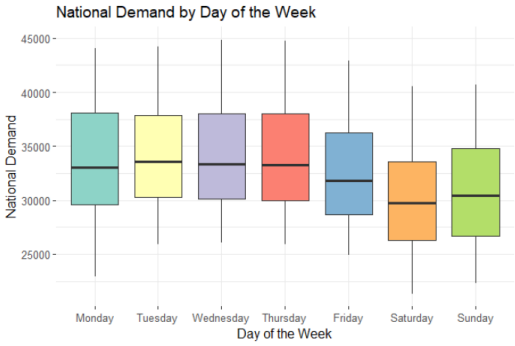

Figure 3: National demand by day of the week

Figures 4 and 5 illustrate autocorrelation function plots for national demand and further highlight the yearly and weekly seasonal patterns and there is clear oscillations of ACF that go outside the blue dashed horizontal lines which represent the 95% confidence interval for the autocorrelation at each lag.

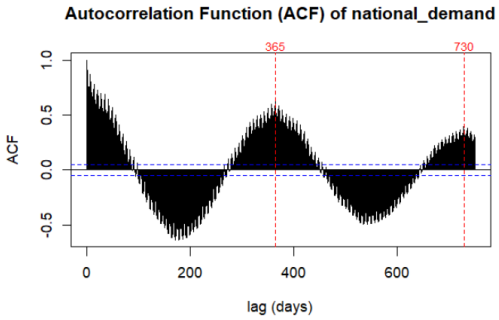

Figure 4: Autocorrelation - max lag = 750

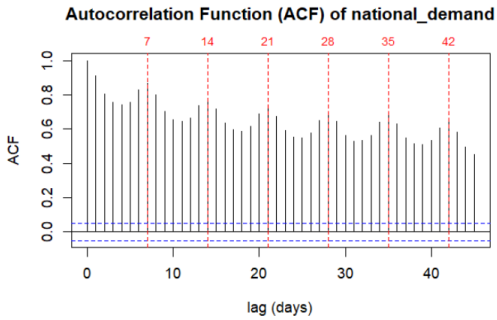

Figure 5: Autocorrelation - max lag = 45

We used spectral analysis to identify seasonal patterns in the national demand time series and the results can be shown in Figure 6. Significant spikes in the periodogram indicate dominant seasonal patterns. A primary periodicity of 375 days (suspected not to be 365 due to a calculation discrepancy in the spectrum R function) and secondary periodicity of 7 days have been found, which are inline with the yearly and weekly patterns we have already observed in the autocorrelation plots.

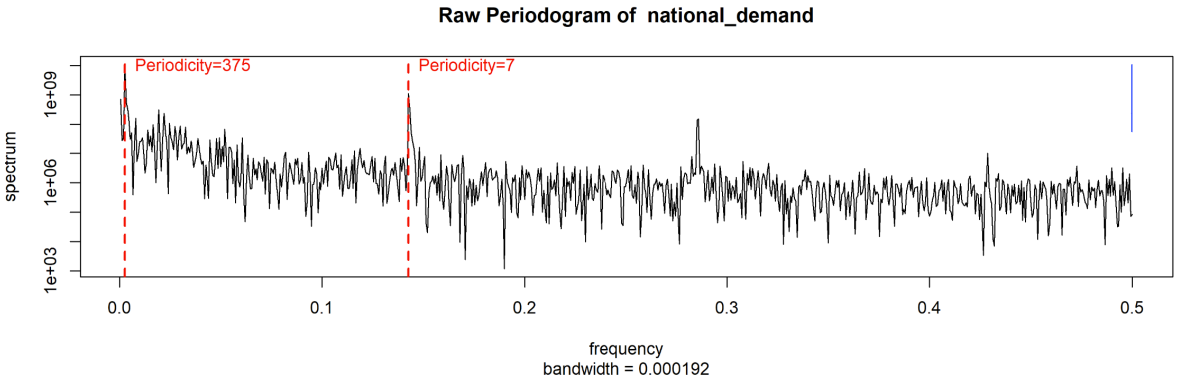

Figure 6: Spectral analysis

In order to analyse whether a time series is stationary, we can use a test called augmented Dickey-Fuller. From this test, we have found the p-value for the national demand time series to be **0.0385**, which is not significant at the 1% level meaning it is found to be non-stationary. Therefore a transformation will need to be carried out so that this time series can be used in further analyses.

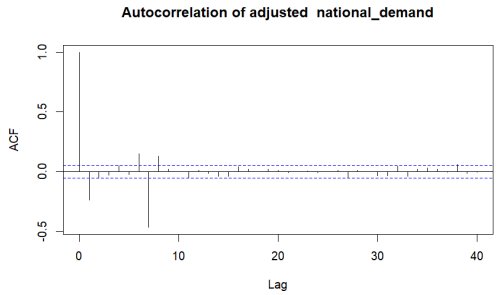

Figure 7: ACF of adjusted national demand (Q=0)

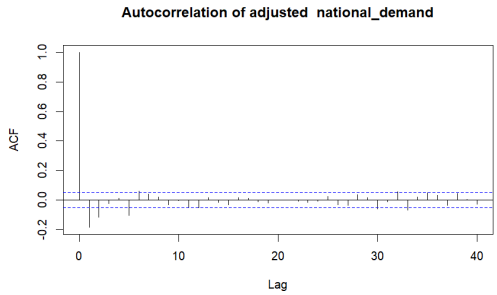

Figure 8: ACF of adjusted national demand with (Q=1)

To achieve this, we apply an ARIMA model with first-order differencing (d=1) to remove trends and first-order seasonal differencing (D=1) to account for seasonality. The residuals, plotted in Figure 7, reveal persistent autocorrelation at lag 7, indicating that the seasonal component was not fully removed. However, this autocorrelation disappears at lags 14 and 21, suggesting that a first-order seasonal moving

average (Q=1) is appropriate. In Figure 8, we present the autocorrelations of the residuals after incorporating the seasonal moving average term, demonstrating improved stationarity. We plot the adjusted time series with Q=1 in figure 9. After using the Augmented Dickey-Fuller Test we confirm that the new p-value is < 0.01.

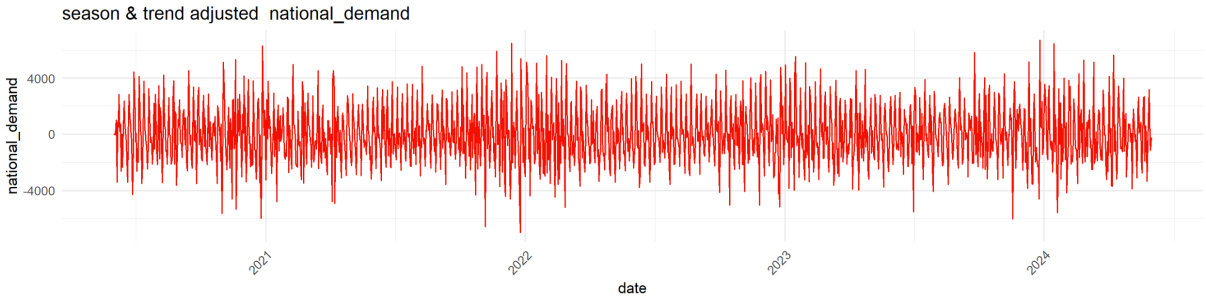

Figure 9: season and trend adjusted national demand

2 Multivariate EDA

We have decided to explore the variables national_demand, wind_generation, solar_generation, max_temp and wind_speed. In order to examine the relationships between these variables we computed a correlation matrix (figure 10). The matrix shows Pearson correlation coefficients which indicate the strength and direction of linear relationships between variables, where a strong correlation is close to -1 or +1, while a weak correlation is close to 0. To test whether these relationships persist after removing seasonality we identified which variables appeared to have non-stationary time series (national_demand, solar_generation, max_temp) and applied a stationary transformation, exactly like what was done for national demand in the univariate EDA analysis section. This process is often called deseasonaliation and will allow us to examine the true underlying dependencies between variables.
Using augmented Dickey-Fuller tests, we found solar_generation and max_temp to have a P-value of **0.5648** and **0.0439** respectively, both of which are not significant at a 1% level. Therefore we applied the same transformations described in the univariate EDA section to these two variables, in addition to the national_demand variable.

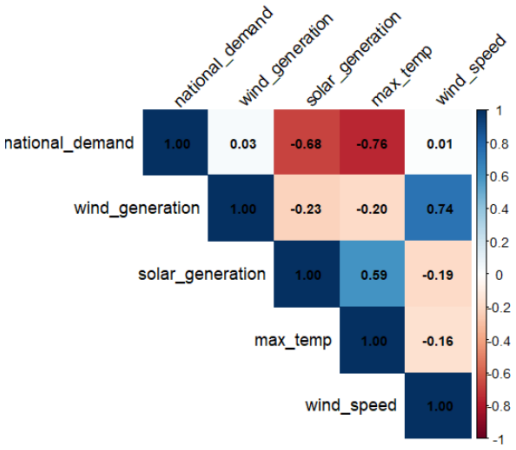

Figure 10: Before transformation

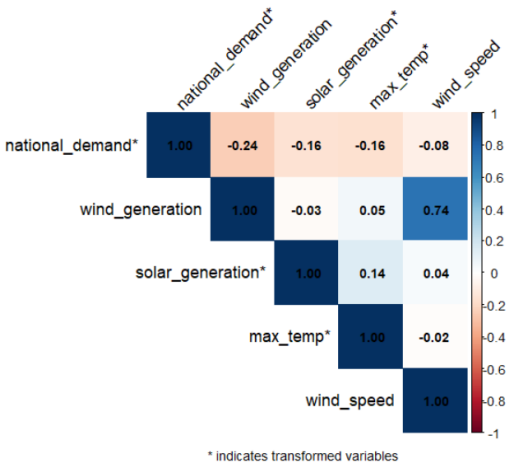

Figure 11: After transformation

After the above transformation (indicated by the asterisk *), we recomputed the correlation matrix (Figure 11) and noticed that the majority of the correlation coefficients weakened. This phenomenon highlights that seasonality was a significant driver of high correlations, and is possibly a "confounder" variable, since they shared similar periodicities. Our new correlation scores reflect the stationary relationships between these variables, allowing us to explore the data without arriving at misleading conclusions due to the influence of seasonality.

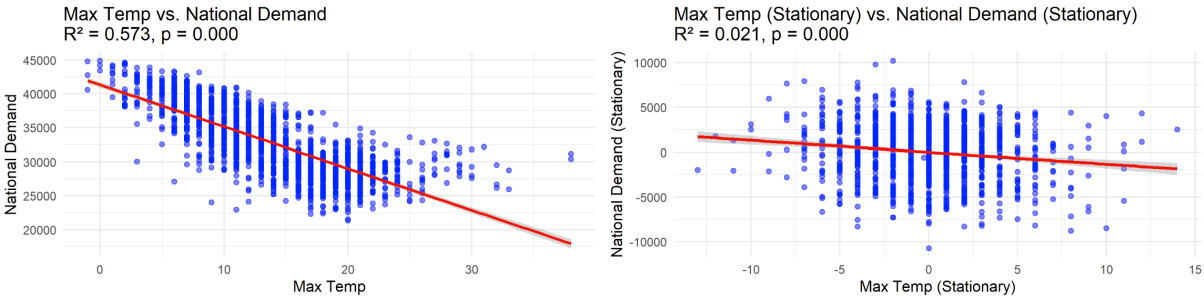

Figure 12: National Demand against Maximum Temperature

One of the most notable drops was present in the variable national demand against max temperature, which experienced a drop in R^2 drop from 0.573 to 0.021 as we apply the transformations. Before we performed de-seasonalisation it appeared that warmer temperatures significantly reduced demand. However, after removing seasonal effects the relationship weakened significantly indicating that the initial correlation was likely driven by the natural yearly cycle rather than a direct cause and effect relationship between these two variables.

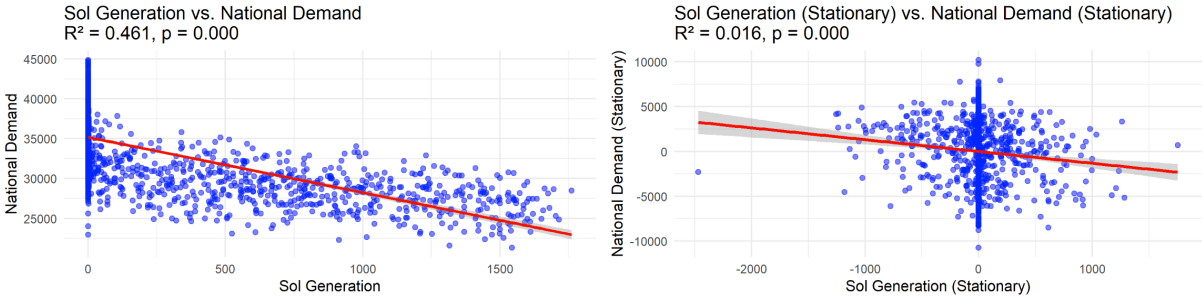

Figure 13: National Demand against Solar Generation

Another drop in correlation due to the de-seasonalisation transformation is national demand against solar generation, where the R^2 went from 0.461 to 0.016. The data suggests that higher solar generation was only linked with lower demand due to seasonal variations not because it directly displaces conventional energy usage. It's likely that national demand follows a separate trend independent of solar availability.

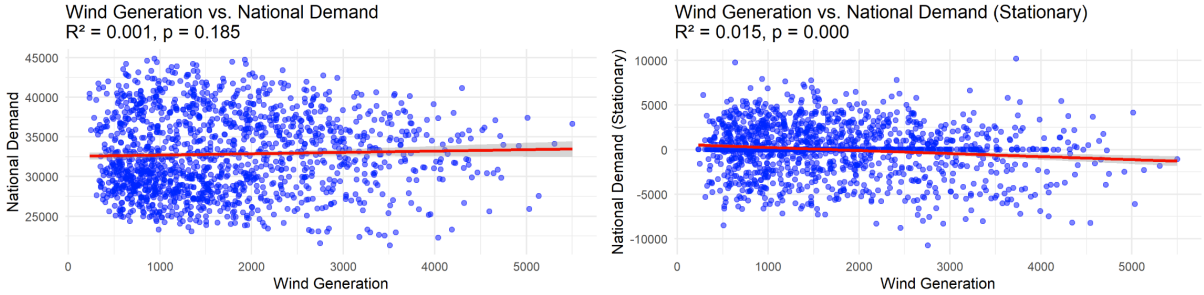

Figure 14: National Demand against Wind Generation

Interestingly, following the de-seasonalitation of the national demand variable, the correlation it had with wind generation actually increased, in contract to other variables. The R^2 went from 0.001 to 0.015 which, although is an increase in correlation, still indicates almost no relationship between these two variables.