

Modelling Peak Electricity Demand in Great Britain

Group 1, Jingren Zhou (s2477464), Qikang Fan (s2472647), Jiayang Zhang (s2312754)

Word Count: 2990

1 Introduction

Accurately forecasting daily peak electricity demand is critical for NESO, especially during winter. This report develops a linear regression model using historical winter data, predicting peak demand based on wind, solar generation, temperature, seasonal, weekly and annual factors. By carefully preprocessing data and rigorous cross-validation, our model delivers improved accuracy, enabling NESO to better manage infrastructure planning and supply risks.

2 Data description and exploratory analysis

2.1 Data sources and preprocessing

We analyse two datasets (`scs_demand_modelling.csv` and `scs_hourly_temp.csv`) provided by NESO.

Dataset Overview: The datasets span winter months (Nov-Mar, 1991-2014) and include temporal indicators (`year` , `monthindex` , `wdayindex` , `DSN` , `start_year`), weather variables (`temp` , `TO` , `TE`), generation factors (`wind` , `solar_S`), and the target demand (`demand_gross`). The `Date` variable identifies the exact day of each record, and hourly temperature data aggregated at 6 pm validates the existing daily temperature (`temp`).

Preprocessing Steps: We create variables `TE^2` and `TO^2` to capture nonlinear temperature effects. Additionally, we remove rows with missing data and contain `temp_peak_hot` and `temp_peak_cold` into our dataset to represent the daily maximum and minimum temperatures. Next, we calculate the difference between these peak temperatures, denoted by `temp_peak_diff` , and the daily average temperature, represented by `temp_average` . For convenience in following analysis, we redefine the variables `year` and `start_year` by subtracting 1991 (i.e., `year = year - 1991` and `start_year = start_year - 1991`). We also use `wdayindex` to distinguish weekdays from weekends (stored in `workingday`). These preprocessing steps will be shown to be meaningful in later analysis. Specifically, the first few rows of the processed dataset are shown below:

	X	Date	wind	solar_S	demand_gross	temp	wdayindex	monthindex	year	TO
1	1	1991-11-01	0.764529	0	44356.5	11.52955764	5	11	0	12.1570585875
2	2	1991-11-02	0.592676	0	38432	9.91422873	6	11	0	10.359490855
3	3	1991-11-03	0.402806	0	36807.5	4.831513948	0	11	0	5.82018314025
4	4	1991-11-04	0.65587	0	48244.5	7.006967808	1	11	0	7.0687541515
5	5	1991-11-05	0.359595	0	48525.5	2.167777053	2	11	0	3.49389927425
6	6	1991-11-06	0.77221	0	48845.5	8.407002165	3	11	0	8.19670665425
7	7	1991-11-07	0.765461	0	47388	10.70184707	4	11	0	11.1924888475
8	8	1991-11-08	0.549788	0	45521.5	6.158275025	5	11	0	7.01651731875
9	9	1991-11-09	0.2633	0	39743.5	2.902364284	6	11	0	4.2064046405
10	10	1991-11-10	0.830943	0	38181	7.641862107	0	11	0	7.23848164525

Showing 1 to 10 of 50 entries

Previous

1

2345Next

	TE	start_year	DSN	TE_sq	TO_sq	workingday	temp_peak_hot	temp_peak_cold	temp_average	temp_peak_diff
1	11.3652007500528	0	1	129.1677880890007	147.7940734999075	weekday	14.04427085	10.43405541	12.04415840583333	3.610215440000001
2	10.8623458025264	0	2	117.9905563336629	107.3190507748286	weekend	12.30437082	8.599576543	10.13934499675	3.704794277000001
3	8.34126447138821	0	3	69.57669298164322	33.87453178605035	weekend	9.622447816	3.155859227	6.626016417875	6.466588588999999
4	7.70500931144411	0	4	59.36716848944044	49.96728525434849	weekday	7.417286437	3.184627708	5.602913615041667	4.232658729
5	5.59945429284705	0	5	31.35388837768326	12.20733213860468	weekday	6.085967405	0.540208101	3.391143501666667	5.545759304
6	6.89808047354853	0	6	47.58351421955152	67.18599997582623	weekday	9.663587183	0.87547771	5.868631467416667	8.788109473
7	9.04528466052426	0	7	81.81717458991548	125.2718066014119	weekday	11.94167858	9.851931457	10.73427287395833	2.089747123
8	8.03090098963713	0	8	64.49537070535465	49.23151528431869	weekday	9.812119574	4.174085957	7.853337830458333	5.638033617
9	6.11865281506857	0	9	37.43791227134653	17.69383999961993	weekend	7.568687586	2.047554513	4.174375237875	5.521133073
10	6.67856723015928	0	10	44.6032602477574	52.39561652862115	weekend	8.52132318	2.119608086	5.3026479105	6.401715094

Showing 1 to 10 of 50 entries

Previous

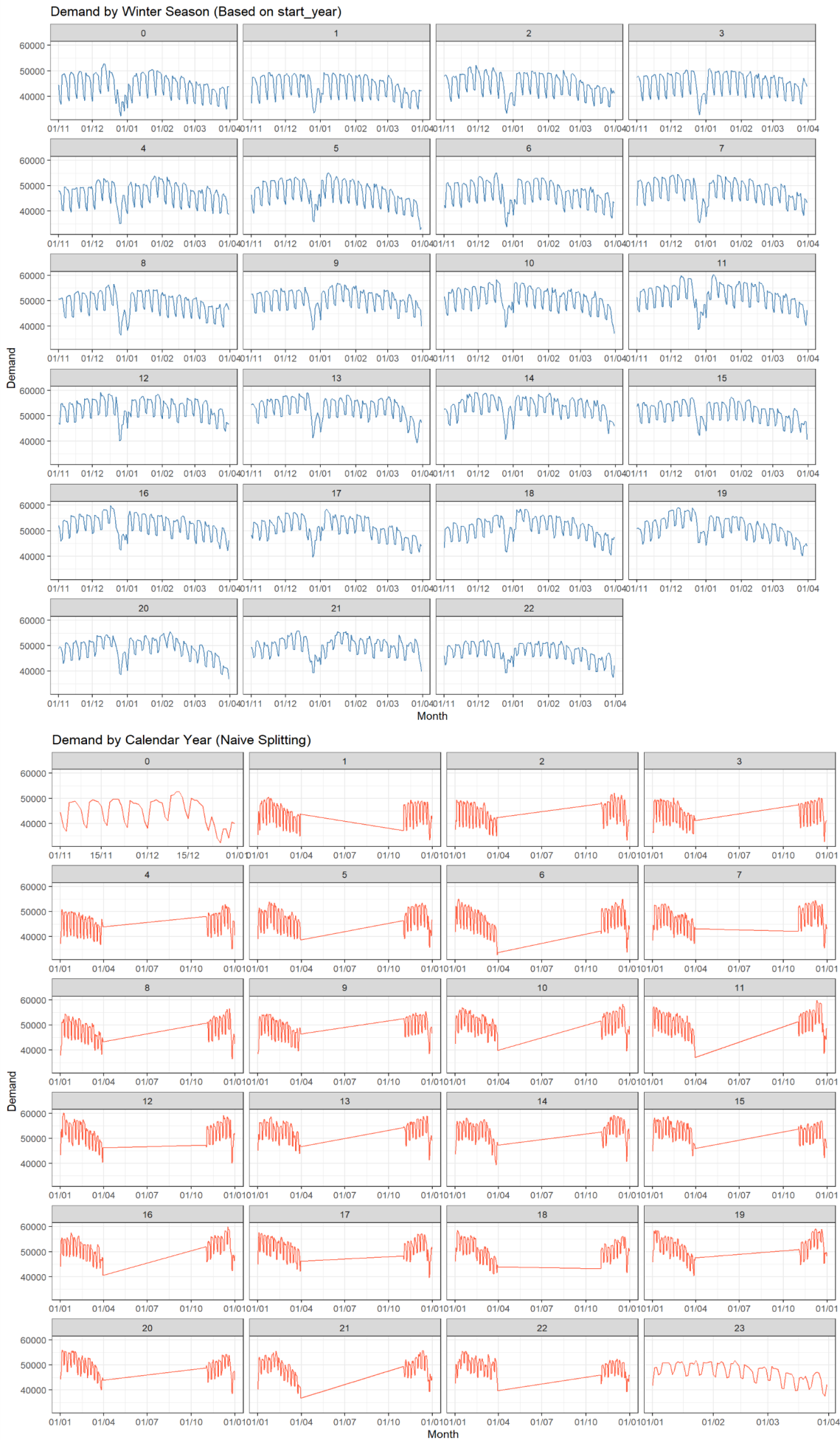
1

2345Next

2.2 Exploratory data analysis

2.2.1 Demand vs. Time

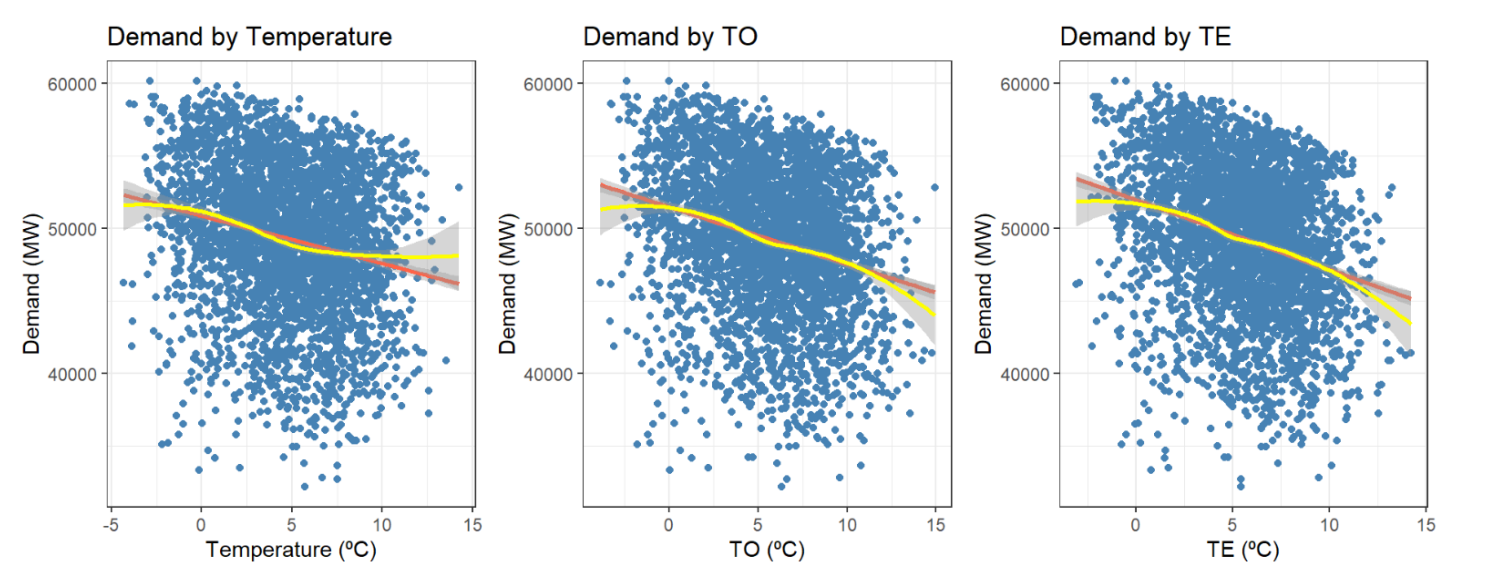
First, we plot the variations of `demand_gross` with respect to `start_year` and `year` separately, as shown below:



From these plots, we observe that the demand demonstrates a certain regular pattern (cyclical variation) over time. However, when using the variable `year`, noticeable gaps appear in the data. This occurs primarily due to the `year` variable includes data from both the previous and the following winter seasons, whereas `start_year` captures complete data for individual winters. Therefore, in our subsequent model selection, we will carefully consider whether to use `start_year` or `year`. Based on these preliminary observations, `start_year` appears to be a more suitable candidate.

2.2.2 Demand vs. Temperature

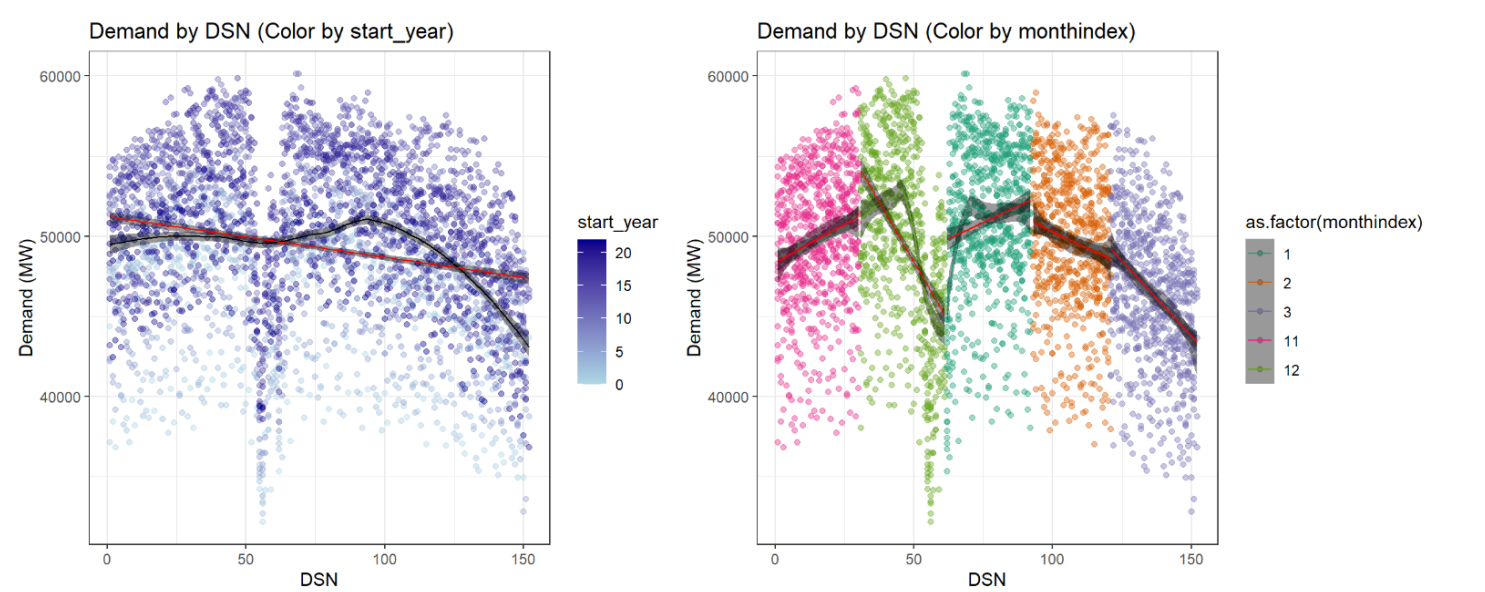
Now, we analyse the relationships between the `demand_gross` and the variables `temp`, `TO`, and `TE`.



Based on the graphical analysis shown above, we observe that `temp`, `TO`, and `TE` all have effects on the demand. These impacts may include both linear and nonlinear components, prompting us to introduce the variables `TE^2` and `TO^2` to capture this nonlinear behavior. Similarly, variables such as `temp_peak_hot`, `temp_peak_cold`, `temp_peak_diff`, and `temp_average` may also influence the demand. This is intuitive because daily maximum and minimum temperatures, their differences, and average daily temperatures are commonly experienced factors that affect consumer behavior and energy usage patterns. Therefore, our model needs carefully consideration when incorporating these temperature-related variables as key predictors.

2.2.3 Demand vs. DSN, start_year and monthindex

Then, we analyse the relationship between the `demand_gross` and the variable `DSN`, using two plots: one with `DSN` on the x-axis, `demand_gross` on the y-axis, and colour intensity representing different start years; the other plot also has `DSN` on the x-axis and `demand_gross` on the y-axis, but uses distinct colours to represent different months.

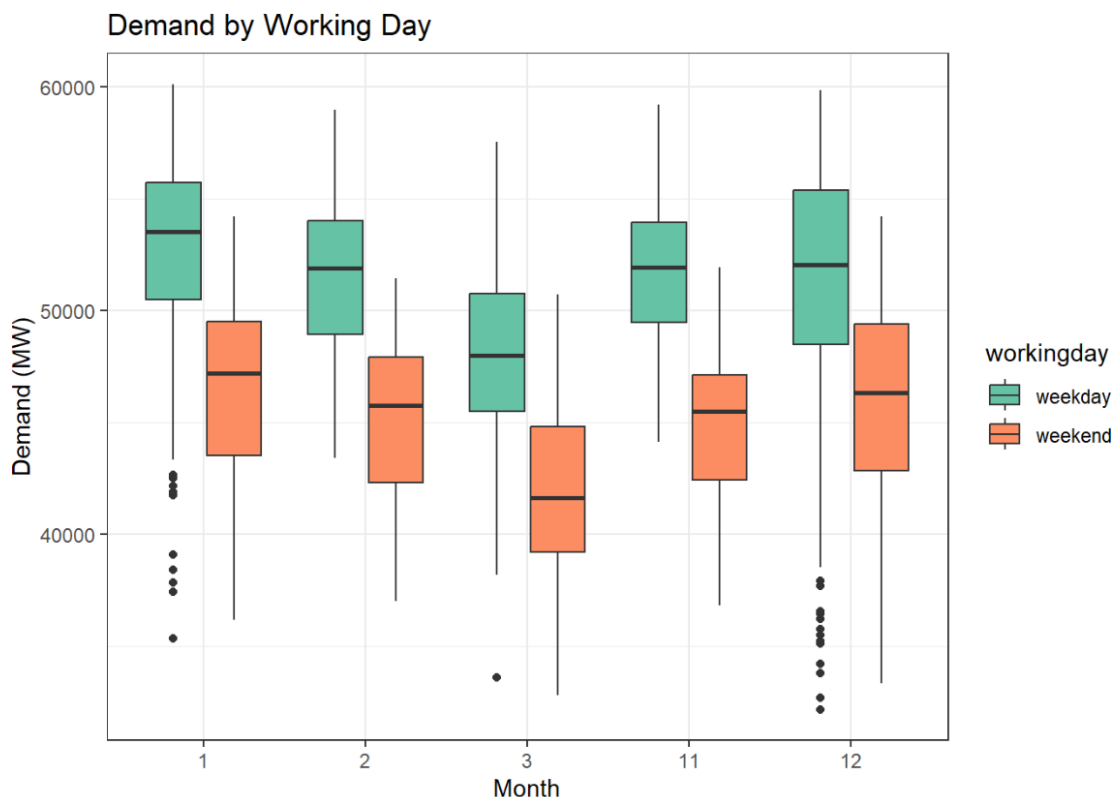


Based on the plots above, we observe a clear increasing trend in demand over successive years, indicated by the colour gradient from lighter (earlier years) to darker (later years). This suggests it is highly valuable to include temporal factors (such as `year` or `start_year`) in our modeling framework.

From the plot on the right (coloured by month), we can clearly see that different months exhibit distinct impacts on the demand. Incorporating variables that represent such intra-winter timing (`monthindex`, `DSN`, etc.) could further enhance the predictive capability of our model.

2.2.4 Demand vs. weekends, weekdays

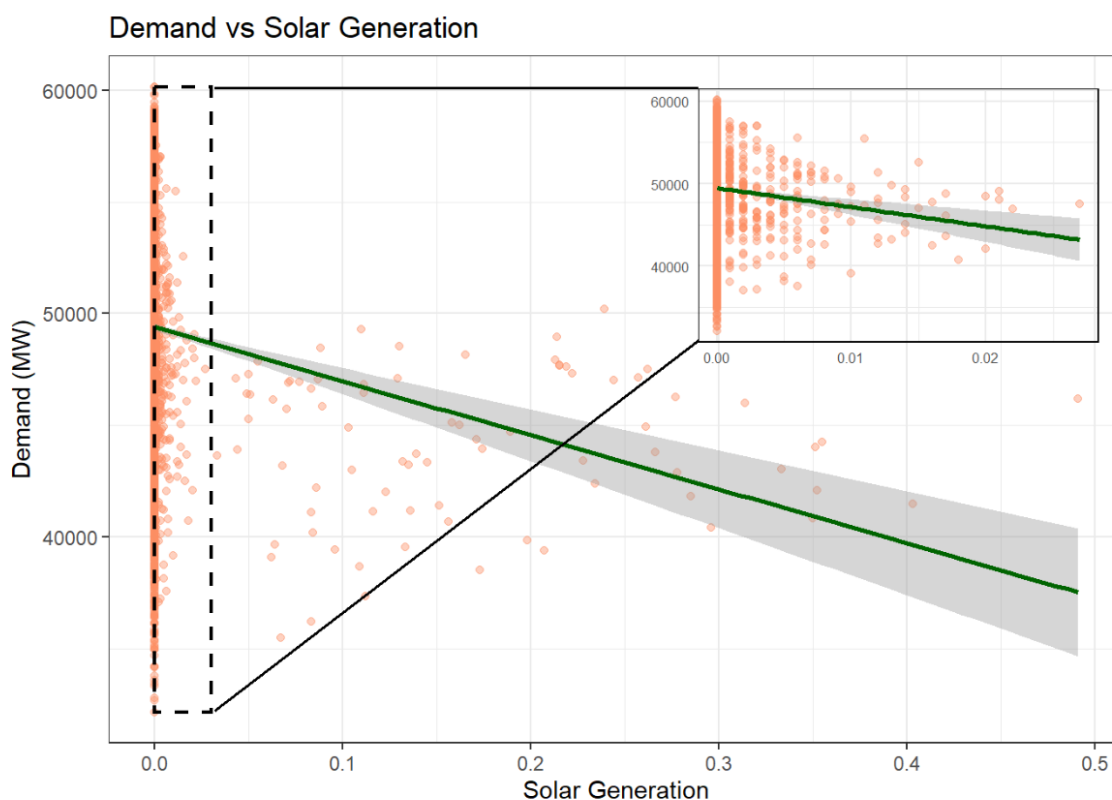
Furthermore, it is important to consider the impact of weekends and weekdays. Here, we use `workingday` to represent this distinction. The following plot shows the relationship between `demand_gross` and `monthindex` for weekdays and weekends:



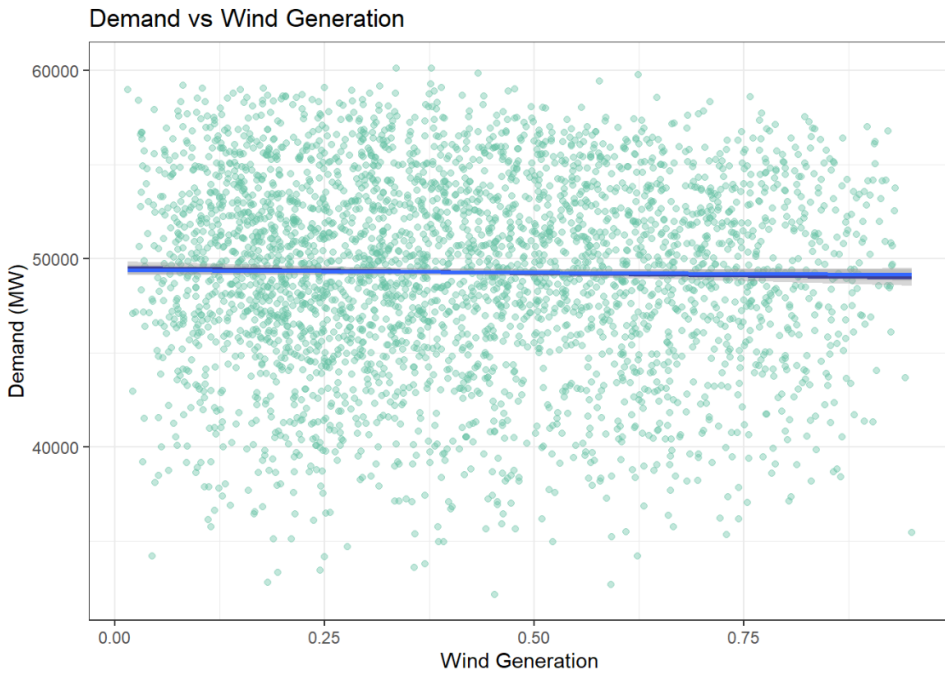
From the boxplot above, we observe a distinct difference in electricity demand between weekdays and weekends across all winter months. Specifically, weekday demand exceeds weekend demand consistently, highlighting a systematic variation related to the day type. Thus including the categorical variable `workingday` is crucial for capturing weekly variations, as it provides a clear and stable pattern.

2.2.5 Demand vs. Natural Resources

Additionally, we analyse the relationships between `demand_gross` and the variables `solar_s` and `wind`. The following plots illustrate these relationships:



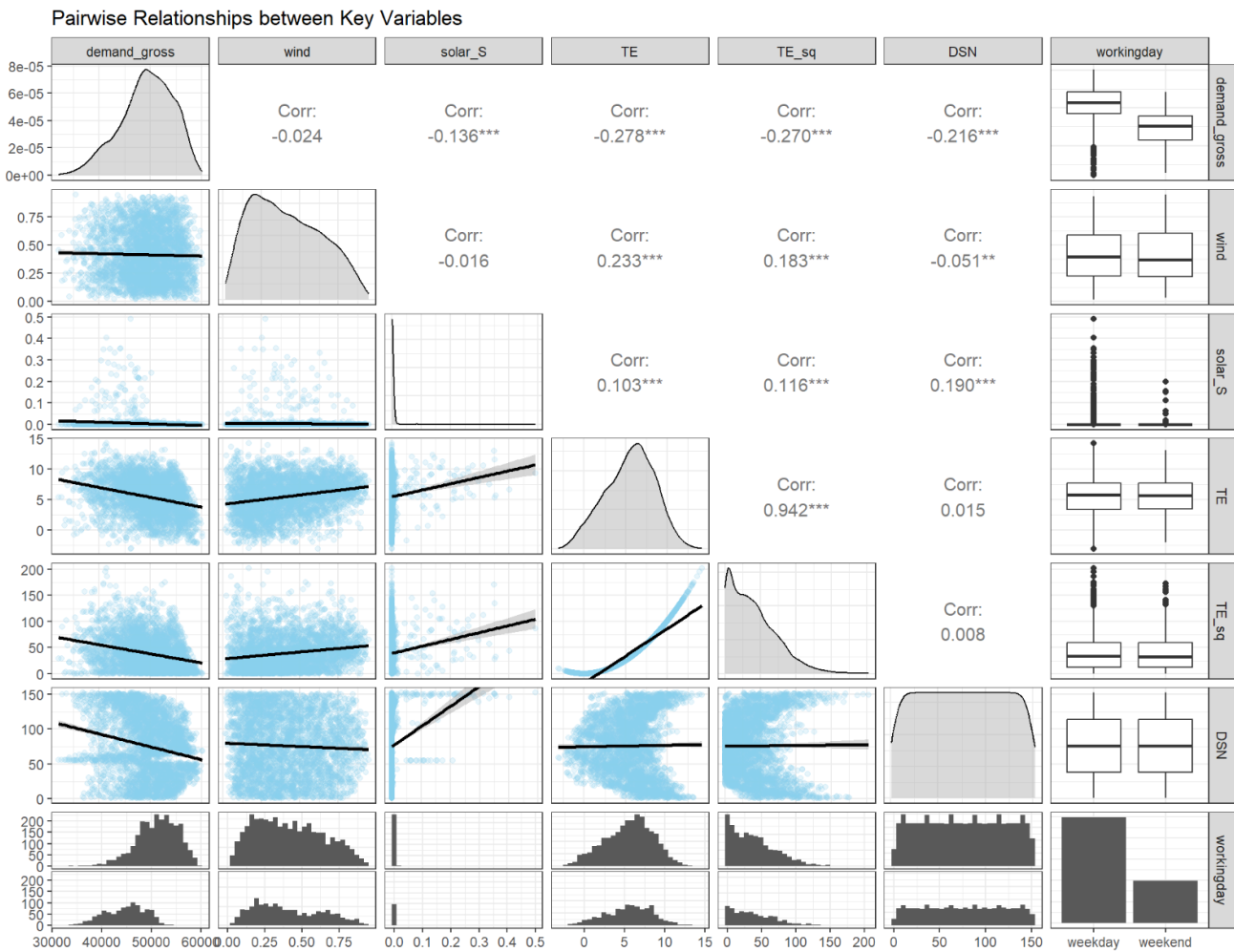
The scatter plot above illustrates the relationship between electricity demand and the solar capacity factor, with the main plot showing the full range of solar capacity values and the inset providing a detailed view of observations clustered near zero. The plot indicates a negative relationship between electricity demand and the solar capacity factor, suggesting that demand decreases as solar generation increases. However, a large proportion of observations cluster near zero, reflecting minimal solar generation during winter months. Thus, it could be beneficial to consider incorporating the solar capacity factor into our model to account for this observed relationship.



The scatter plot above shows the relationship between demand and wind generation. We observe a very slight negative linear trend, indicating that demand marginally decreases as wind generation increases. Despite the minimal magnitude of this trend, it may still be beneficial to consider incorporating wind generation into the model to capture this subtle relationship during winter months.

2.2.6 Pairwise Relationships

Finally, in order to systematically explore and visualise the pairwise relationships amongst key variables, we use the `ggpairs` function, generating a comprehensive scatterplot matrix and correlation analysis graph:



From the scatterplots and correlation matrix above, we clearly observe the relationships between `demand_gross` and each key variable. The variable `wind` demonstrates a very weak negative correlation (approximately `-0.024`) with the demand, suggesting only a slight decline in the demand as wind generation increases. However, it may still be worth include the variable `wind` in the model. In contrast, the variable `solar_s` shows a modest negative relationship (`-0.136`), indicating that increased solar generation correlates with lower electricity demand. Obviously, this aligns with the intuitive understanding, where lower solar radiation conditions lead to increased lighting and heating demands, which is an effect the model captures and explains effectively.

Variables related to temperature, `TE` and its quadratic term `TE_sq`, exhibit significant negative correlations (`-0.278` and `-0.270`, respectively) with the demand, suggesting that these linear or nonlinear variables may be included. Additionally, `DSN` (the sequential numbering of days starting from 1st November of each year) presents a negative correlation (`-0.216`) with the demand, consistent with the expectation of gradually decreasing electricity demand as winter progresses toward warmer periods.

The categorical variable `workingday` visually distinguishes clear differences between weekday and weekend electricity demand, reinforcing its importance in modeling weekly fluctuations.

With this, we complete our exploratory data analysis. Next, starting from a basic reference model, we will build upon it to develop a more effective predictive model.

3 Model fitting and cross-validation results

We begin to explore how we can improve upon our baseline model:

$$M_0 : Y_i = \beta_0 + \beta_1 \text{wind}_i + \beta_2 \text{solar}_S + \beta_3 \text{temp}_i + \beta_4 \text{wdayindex}_i + \beta_5 \text{monthindex}_i$$

Given our previous analysis of the data, intuitively, we consider adding the variable `DSN` to represent the daily progression within each winter season, capturing seasonal variations in the demand. Additionally, incorporating `start_year` or `year` could capture variations in the demand across different years, potentially accounting for broader economic factors or other year-specific influences.

The stepwise model fitting process is as follows:

3.1 Model Exploration

First, we add the variables `start_year` and `year` respectively to the baseline model, as these can help us to capture the effect of the year in the model. We construct the models and calculate their Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), R^2 , adjusted R^2 , F-statistics, and p-values to analyse and compare their performances.

Comprehensive Comparison of Models

Model	R_squared	Adj_R_squared	F_statistic	P_value	AIC	BIC
M0: M0	0.4656	0.4636	232.2	1	67145.12	67237.44
M1: M0 + factor(year)	0.7747	0.7724	328.84	1	64185.41	64419.28
M2: M0 + factor(start_year)	0.7778	0.7756	344.44	1	64135.16	64362.87

From the table above, we can observe that model M2 consistently performs better across all metrics. Specifically, it has the highest R^2 (0.7778) and the adjusted R^2 (0.7756), as well as the lowest AIC (64135.16) and BIC (64362.87) values. These statistical measures indicate that model M2 provides the best fit amongst the three models considered. Therefore, we choose model M2, as the inclusion of the `start_year` variable significantly improves the model's explanatory power regarding yearly variations.

Next, we consider the impact of changes in the time scale on the demand within the same winter season (from early winter in November to late winter in March). Both `DSN` and `monthindex` represent subdivisions of this time period but with slight differences. The original model uses `monthindex`; we propose replacing it with `DSN` and performing cross-analysis, including evaluating the AIC, BIC, and R^2 metrics, to assess their comparative performances.

Comprehensive Comparison of Models

Model	R_squared	Adj_R_squared	F_statistic	P_value	AIC	BIC
M2: M2	0.7778	0.7756	344.44	1	64135.16	64362.87
M3: M2 , factor(monthindex) -> DSN	0.7597	0.7574	340.4	1	64402.68	64611.94

The corresponding Cross-Validation by winter months (Nov-Mar) is as follows:

Cross-Validation Results by Winter Months (Nov-Mar)

intercept	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD	Model
TRUE	2436.2511	0.7734	1563.6093	77.36202	0.0115626	37.51242	M2
TRUE	2533.3808	0.7548	1669.5438	79.01830	0.0132031	33.26903	M3

From the table above, we see that model M2 performs better than model M3. M2 has higher R^2 (0.7778 vs. 0.7597) and adjusted R^2 (0.7756 vs. 0.7574), and lower AIC (64135.16 vs. 64402.68) and BIC (64362.87 vs. 64611.94). This shows that M2 fits the data better, capturing seasonal changes more effectively. Therefore, we select M2. Cross-validation results also show that M2 clearly performs better, with lower RMSE and MAE, and higher R^2 . Thus, using `monthindex` in M2 is better than using `DSN` in M3 for predicting seasonal patterns.

Since M2 performs well, we plan to improve it further by replacing `factor(wdayindex)` with `factor(workingday)`. We make this change because electricity usage patterns usually differ between weekdays and weekends, and considering this difference might increase the model's accuracy. We conducted comparative and cross-validation analysis between this new model and M2, with results shown in the two tables below.

Comprehensive Comparison of Models

Model	R_squared	Adj_R_squared	F_statistic	P_value	AIC	BIC
M2: M2	0.7778	0.7756	344.44	1	64135.16	64362.87
M4: M2 , factor(wdayindex) -> factor(workingday)	0.7651	0.7631	374.4	1	64318.93	64515.88

Cross-Validation Results by Winter Months (Nov-Mar)

intercept	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD	Model
TRUE	2431.8406	0.7736	1566.0302	218.6996	0.0316532	110.01581	M2
TRUE	2496.8069	0.7614	1653.2949	180.1820	0.0245770	86.93143	M4

The comparison of models indicates that M2 performs slightly better than M4 (modified model), with higher R-squared values and lower AIC/BIC scores. Cross-validation during winter months further supports M2's superior predictive accuracy, reflected in lower RMSE and MAE.

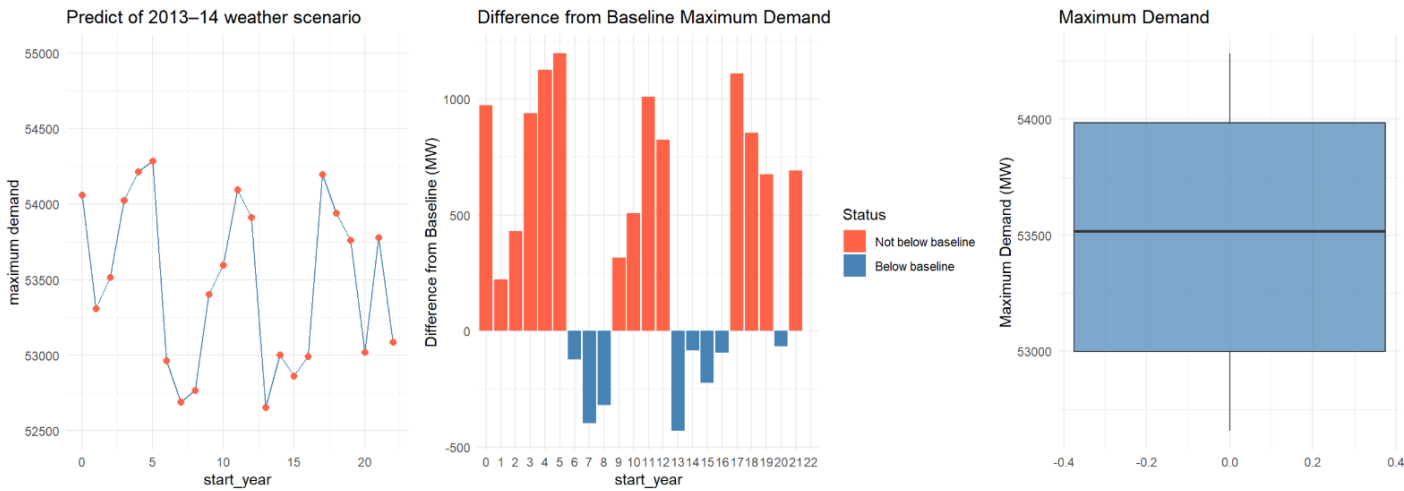
Our next step will now be to analyse temperature (`temp`) and the environmental factors (`solar_s` and `wind`).

3.2 Model Refinement: Weather Effects

We use scenario analysis to explore how historical weather conditions affect fluctuations in winter peak electricity demand relative to the baseline from the 2013-14 winter season.

MaxDemand Summary Statistics

Metric	Value
Min MaxDemand	52656.8745
Max MaxDemand	54285.0557
Range (max - min)	1628.1812
Mean MaxDemand	53485.0748
SD MaxDemand	548.8078



The figures clearly illustrate that weather conditions from different historical years significantly impact peak electricity demand. In some scenarios, notably colder winters (such as `start_year = 5`) increase the demand by over 1,000 MW compared to the baseline, whereas milder conditions (e.g., `start_year = 13`) can reduce it by several hundred MW. Overall, weather variations alone cause peak demand fluctuations of around 1,700 MW (ranging from approximately 400 MW below to 1,300 MW above baseline). This substantial variation emphasises the importance of accounting for weather-driven uncertainty in electricity system planning.

Additionally, these extreme scenarios highlight critical planning implications. The most severe winter conditions which have cold temperature and low wind generation, thus to ensure system reliability, an extra capacity reserve of around 1,000~1,2000 MW is required. Operators should thus include this weather sensitivity into emergency response planning, capacity reserves, and risk assessments with caution. Consequently, incorporating wind and solar generation into our model becomes essential, as it helps to capture and reflect the variability in the demand driven by weather changes.

3.3 Model Refinement: Temperature Effects

It is time to make a comparative analysis across multiple regression models to identify which best captures the variability of winter peak electricity demand, incorporating temperature-related variables.

Comprehensive Comparison of Models

Model	R_squared	Adj_R_squared	F_statistic	P_value	AIC	BIC
M2	0.7778	0.7756	344.44	1	64135.16	64362.87
M5: M2, temp -> TE	0.7914	0.7893	373.26	1	63915.79	64143.5
M6: M2, temp -> TO	0.7834	0.7812	355.82	1	64046.88	64274.59
M7: M2, temp -> temp_peak_hot	0.7862	0.7841	361.8	1	64001.39	64229.1
M8: M2, temp -> temp_peak_cold	0.7772	0.7749	343.18	1	64145.07	64372.78
M9: M2, temp -> temp_peak_diff	0.7436	0.741	285.31	1	64633.81	64861.53
M10: M2, temp -> temp_average	0.7841	0.7819	357.32	1	64035.37	64263.09

Cross-Validation Results by Winter Months (Nov-Mar)

intercept	RMSE	Rsquared	MAE	RMSED	RsquaredSD	MAESD	Model
TRUE	2432.6861	0.7731	1566.5622	233.2515	0.0385299	76.02516	M2
TRUE	2356.8063	0.787	1438.8442	247.5168	0.0398695	81.97979	M5
TRUE	2401.6195	0.7788	1517.9608	237.0185	0.0388221	78.31010	M6
TRUE	2385.2326	0.7817	1479.8416	250.0258	0.0412301	76.52688	M7
TRUE	2436.0135	0.7723	1571.7344	249.3217	0.0415941	90.98618	M8
TRUE	2615.9539	0.7379	1774.7242	221.1215	0.0391058	81.45418	M9
TRUE	2397.2357	0.7795	1510.5527	248.0732	0.0407515	81.51814	M10

It's important to highlight that model `M5` (using `TE` instead of `Temp`) demonstrates superior performance. Among all examined models, it has the greatest R^2 and adjusted R^2 values and best prediction accuracy, as shown by the lowest RMSE and MAE in cross-validation results. This indicates that replacing the original temperature variable with `TE` significantly improves model performance. Specifically, the use of `TE` allows the model to better capture demand variability during winter months.

3.4 Model Diagnostics

Based on the analysis presented above, we select the following as our final model:

$$\text{Final Model : } Y_i = \beta_0 + \beta_1 \text{wind}_i + \beta_2 \text{solar_S}_i + \beta_3 \text{TE}_i + \beta_4 \text{wdayindex}_i + \beta_5 \text{monthindex}_i + \beta_6 \text{start_year}_i + \varepsilon_i$$

where $\varepsilon_i \sim \text{Normal}(0, \sigma^2)$, Y_i is the demand gross at time i , $\beta_j, j = 1, \dots, 6$ are the regression coefficients and ε_i is the error term at time i .

We perform a comprehensive analysis comparing our final model with the baseline model. This comparison aims to demonstrate that our selected model is well-constructed and provides improved predictive performance compared to the original model.

```
##
## Call:
## lm(formula = M5, data = demand_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16065.4   -602.0    270.5   1153.1   5007.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    42565.07     244.71  173.938 < 2e-16 ***
## wind           1061.25     190.46   5.572 2.71e-08 ***
## solar_S       -13950.24    1433.58  -9.731 < 2e-16 ***
## TE             -481.71      16.97  -28.390 < 2e-16 ***
## factor(wdayindex)1    7094.76     149.43  47.479 < 2e-16 ***
## factor(wdayindex)2    7229.70     149.09  48.493 < 2e-16 ***
## factor(wdayindex)3    7159.72     149.09  48.022 < 2e-16 ***
## factor(wdayindex)4    6933.73     149.17  46.483 < 2e-16 ***
## factor(wdayindex)5    5525.25     149.31  37.005 < 2e-16 ***
## factor(wdayindex)6     784.35     149.00   5.264 1.50e-07 ***
## factor(monthindex)2   -943.23     128.11  -7.363 2.25e-13 ***
## factor(monthindex)3  -2880.72     139.43 -20.660 < 2e-16 ***
## factor(monthindex)11    396.03     137.92   2.871 0.00411 **
## factor(monthindex)12  -1305.36     124.76 -10.463 < 2e-16 ***
## factor(start_year)1   -181.46     269.95  -0.672 0.50151
## factor(start_year)2    155.07     270.43   0.573 0.56638
## factor(start_year)3   1231.97     270.66   4.552 5.51e-06 ***
## factor(start_year)4   2064.35     270.16   7.641 2.77e-14 ***
## factor(start_year)5   2558.49     270.20   9.469 < 2e-16 ***
## factor(start_year)6   3499.66     270.99  12.914 < 2e-16 ***
## factor(start_year)7   3970.46     270.09  14.700 < 2e-16 ***
## factor(start_year)8   4943.27     269.83  18.320 < 2e-16 ***
## factor(start_year)9   6213.78     269.95  23.019 < 2e-16 ***
## factor(start_year)10  6982.98     270.43  25.822 < 2e-16 ***
## factor(start_year)11  8106.72     270.37  29.983 < 2e-16 ***
## factor(start_year)12  8697.59     269.71  32.248 < 2e-16 ***
## factor(start_year)13  8843.60     270.33  32.714 < 2e-16 ***
## factor(start_year)14  8850.46     270.19  32.756 < 2e-16 ***
## factor(start_year)15  7864.72     271.39  28.980 < 2e-16 ***
## factor(start_year)16  8208.08     269.87  30.415 < 2e-16 ***
## factor(start_year)17  6418.61     270.16  23.759 < 2e-16 ***
## factor(start_year)18  6402.22     270.56  23.663 < 2e-16 ***
## factor(start_year)19  6476.80     270.75  23.922 < 2e-16 ***
## factor(start_year)20  5109.25     270.59  18.882 < 2e-16 ***
## factor(start_year)21  4914.43     270.91  18.141 < 2e-16 ***
## factor(start_year)22  3713.24     270.59  13.723 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Residual standard error: 2349 on 3443 degrees of freedom
Multiple R-squared: 0.7914, Adjusted R-squared: 0.7893
F-statistic: 373.3 on 35 and 3443 DF, p-value: < 2.2e-16

The overall regression model demonstrates strong explanatory power, with a multiple R^2 of 0.7914 and an adjusted R^2 of 0.7893, indicating that it explains nearly 79% of the variance in peak electricity demand. The F-statistic of 373.3 (p-value < 2e-16) confirms the model's high significance, and most predictors (including wind, solar_S, TE, plus relevant categorical factors) show p-values well below 0.001, reflecting their robustness in capturing winter peak demand variability.

In contrast, start_year1 (1991) and start_year2 (1992) yield large p-values, suggesting weak significant difference from the reference category. Despite these isolated non-significant effects, the model remains highly reliable and predictive, as shown by its high R^2 and significant F-test. Subsequent analysis will compare the baseline model (M_0) with this final model using R^2 , AIC, and BIC to confirm performance and guide further refinements—such as exploring non-linearities or introducing additional variables to account for annual variability.

Comprehensive Comparison of Models

Model	R_squared	Adj_R_squared	F_statistic	P_value	AIC	BIC
M0	0.4656	0.4636	232.2	1	67145.12	67237.44
Final Model	0.7914	0.7893	373.26	1	63915.79	64143.5

Cross-Validation Results by Winter Months (Nov-Mar)

intercept	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD	Model
TRUE	3752.6779	0.4608	2973.2365	126.6423	0.0194961	89.73274	M0
TRUE	2362.5839	0.7867	1440.766	141.4658	0.0162348	53.76422	Final Model

Relative to the baseline model, the final model demonstrates better performance across all key metrics. Especially, it achieves a higher R^2 and adjusted R^2 , indicating an improved fit. Its lower AIC and BIC highlight enhanced explanatory power, while cross-validation shows reduced RMSE and MAE, confirming superior predictive accuracy. Consequently, these findings confirm the final model's effectiveness in capturing the variability of winter peak electricity demand and providing reliable forecasts.

ANOVA analysis further supports the final model's superiority over the baseline:

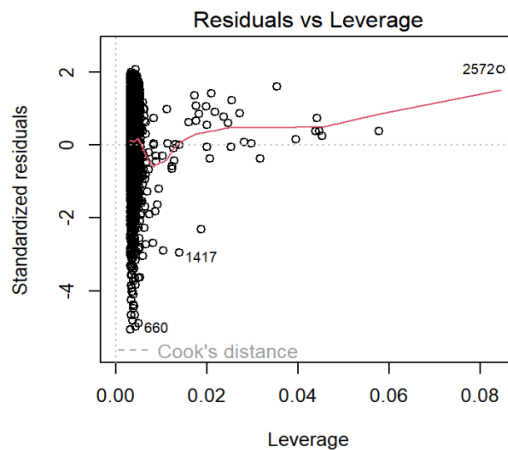
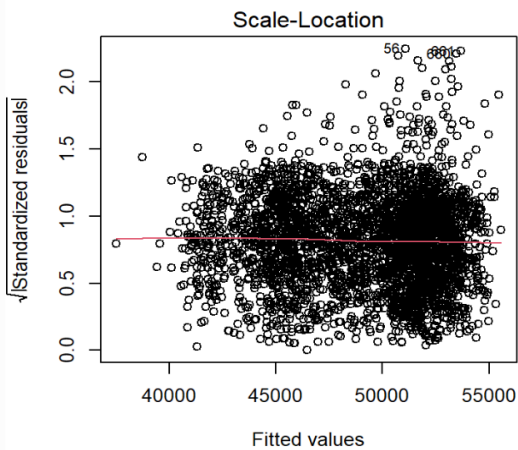
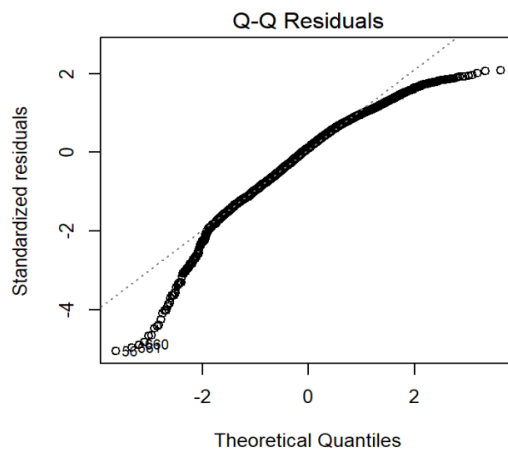
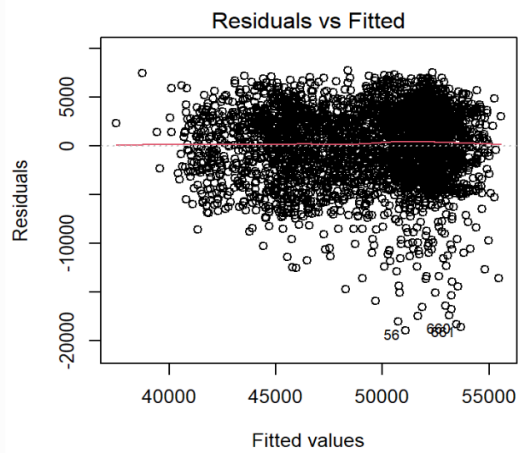
```
## Analysis of Variance Table
##
## Model 1: demand_gross ~ wind + solar_S + temp + factor(wdayindex) + factor(monthindex)
## Model 2: demand_gross ~ wind + solar_S + TE + factor(wdayindex) + factor(monthindex) +
##   factor(start_year)
##   Res.Df      RSS Df Sum of Sq    F    Pr(>F)
## 1    3465 4.8660e+10
## 2    3443 1.8991e+10 22 2.9669e+10 244.49 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the ANOVA table above, we performed an F-test to compare the nested models, confirming that the final model provides a highly significant improvement (p < 2.2e-16) over the baseline. This statistical evidence, together with the outstanding fit and predictive metrics, indicates the final model's added parameters are justified, and it is the preferred choice for forecasting winter peak electricity demand.

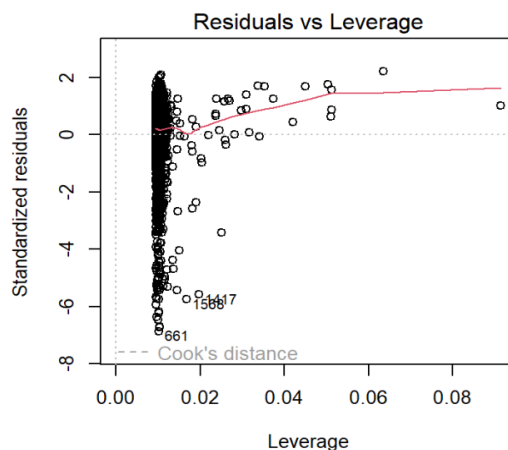
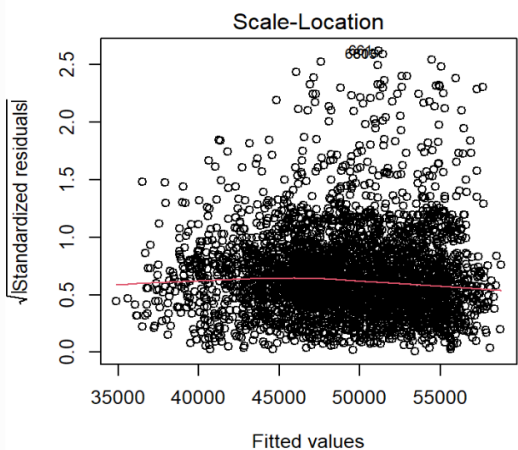
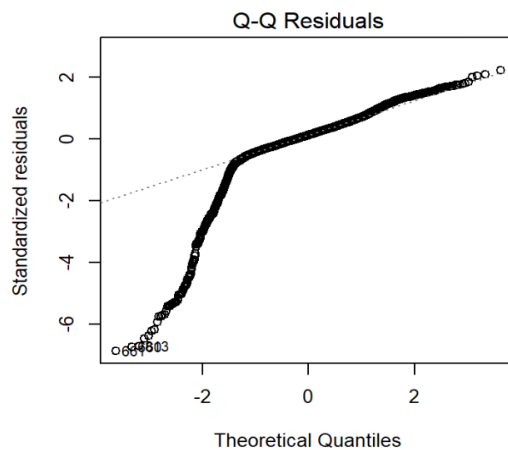
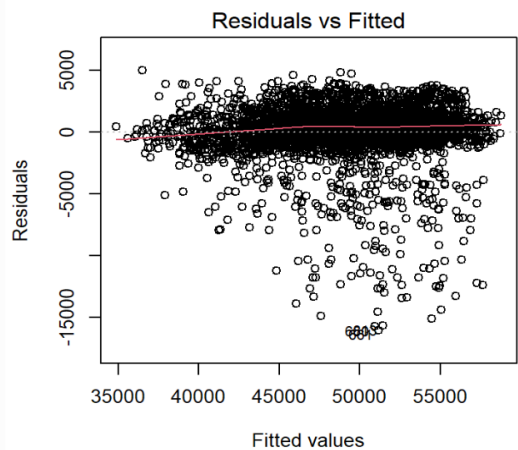
3.5 Model Diagnostics: Residuals

We conduct residual diagnostics to assess the final model's fit and identify any potential issues, such as nonlinearity, heteroskedasticity, or influential outliers.

The graphical analysis for the residuals of the baseline model:



The residual plots for the final model are as follows:



From a residual diagnostics perspective, the revised model demonstrates clear improvements over the initial version. In the Residuals vs. Fitted plot, the residual cloud appears more randomly scattered, indicating a reduction in the subtle patterns or curvature seen in the baseline model's residuals.

In addition, the Q–Q plot for the final model reveals a tighter alignment with the diagonal for the bulk of the data, suggesting that most residuals adhere more closely to normality. Likewise, the Scale-Location plot exhibits a more uniform spread of standardised residuals across the range of fitted values, pointing to diminished heteroskedasticity. Although both models feature a small number of points with high leverage and sizable residuals(e.g., `start_year1` corresponding to 1991 and `start_year2` corresponding to 1992), the overall distribution in the final model is more centralised, indicating a milder influence of outliers.

3.6 Variable Inflation Factor (VIF) Analysis

To assess the severity of multicollinearity among explanatory variables in the final model, we calculated the Variance Inflation Factor (VIF). High VIF values indicate strong correlations between predictors, potentially leading to unstable coefficient estimates and reduced model interpretability. The table below presents the VIF results for each variable in the final model:

VIF Results				
	GVIF	Df	GVIF^(1/(2*Df))	Adjusted_VIF
wind	1.165	1	1.079	1.079
solar_S	1.073	1	1.036	1.036
TE	1.617	1	1.272	1.272
factor(wdayindex)	1.012	6	1.001	1.001
factor(monthindex)	1.489	4	1.051	1.051
factor(start_year)	1.166	22	1.004	1.004

The VIF results presented above indicate minimal concerns regarding multicollinearity within the selected variables. Typically, a VIF value exceeding 5 is considered problematic, while values below this threshold are acceptable. Here, all adjusted VIF values are considerably lower than 5, with the highest being 1.272 for the variable `TE`. Consequently, we conclude that multicollinearity is not an issue in our final model, affirming the reliability of the estimated coefficients and stability of the overall analysis.

4 Conclusion

Our final linear regression model provides a transparent and interpretable method for predicting daily peak electricity demand, effectively supporting NESO's long-term planning objectives. The exploratory data analysis identified temperature-related factors (`TE`), renewable generation (`solar` , `wind`), seasonal (`monthindex`), and weekday (`wdayindex`) variations as significant predictors. Specifically, we observed a notable negative correlation between peak demand and solar generation, validating the inclusion of renewable sources (`solar_S` and `wind`) as key explanatory variables.

Although weekday demand consistently surpassed weekend levels—highlighting the systematic weekly variation—the final model favoured `wdayindex` over the simpler binary `workingday` variable due to its superior statistical performance in terms of R^2 , AIC, and BIC. Additionally, seasonal variations (`monthindex`) and evolving annual dynamics (`start_year`) effectively captured temporal trends, further reinforcing the robustness and predictive power of our modeling framework.

5 Limitations and Recommendations

5.1 Limitations

- Historical Data and Future Changes:** Historical data (only from 1991-2014) might not fully capture future structural changes such as demographic shifts, economic growth, technological advancements, or climate change. Therefore, predictions based solely on historical trends may be unreliable during extreme scenarios.
- Variable Selection and Omitted Bias:** The current model mainly considers `TE` , `wind` , `solar` , and `monthindex` , but may neglect other significant drivers of electricity demand, e.g., the adoption of electric vehicles and smart grid developments. Additionally, the report lacks a method for regularly updating model parameters to adaption to these ongoing changes.

- **Limitations of Linear Assumptions:** Linear regression provides clear interpretability but may inadequately represent non-linear relationships, interaction effects, or situations where variability of errors (heteroscedasticity) exists.
- **Ignoring Regional Differences:** The model does not account for spatial differences across UK regions, such as the distinct temperature and electricity demand between England and Scotland. For instance, northern areas experience colder winters with higher heating demands. Without geographic variables, global predictions might deviate from local realities.
- **Exclusion of Holidays and Special Events:** The model overlooks holidays (e.g., Christmas), major events, and unexpected occurrences, which can significantly alter electricity usage patterns. For example, increased remote work during the 2020 pandemic changed consumption trends, which the current model cannot effectively capture.
- **Error Distribution and Outliers:** The model assumes normal distribution of residuals; however, actual electricity demand data often includes outliers and heavier tails, potentially reducing predictive accuracy.

5.2 Recommendations

- Conduct scenario analysis for varying future situations, such as increased electric vehicle use, energy policy shifts, and extreme weather events, to quantify prediction uncertainties.
- Incorporate high-frequency weather forecasts or segmented weather data to enhance sensitivity to extreme climate events.
- Apply Monte Carlo simulation techniques to expand scenario analyses, particularly for assessing risks associated with extreme demand levels.

6 Code Appendix

```
#
# Place the code needed in the Report_project02.Rmd, including documentation.
#

#' compare_models
#'
#' Compare multiple linear regression models, computing and displaying R^2,
#' adjusted R^2, F-statistic, p-value, AIC, and BIC.
#'
#' @param models      A list of linear regression models (lm objects).
#' @param model_labels A character vector of labels corresponding to each model.
#'                    If `NULL`, no custom labels are displayed.
#'
#' @return A formatted table (using `kableExtra`) comparing statistical metrics across models.
compare_models <- function(models, model_labels = NULL) {
  r_squared      <- sapply(models, function(m) summary(m)$r.squared)
  adj_r_squared  <- sapply(models, function(m) summary(m)$adj.r.squared)
  f_stats        <- sapply(models, function(m) summary(m)$fstatistic[1])
  p_values       <- sapply(models, function(m) {
    fs <- summary(m)$fstatistic
    pf(fs[1], fs[2], fs[3])
  })
  aic_values     <- sapply(models, AIC)
  bic_values     <- sapply(models, BIC)

  max_r2         <- max(r_squared)
  max_adj_r2     <- max(adj_r_squared)
  max_f          <- max(f_stats)
  min_pval       <- min(p_values)
  min_aic        <- min(aic_values)
  min_bic        <- min(bic_values)

  comparison_df <- data.frame(
    Model = model_labels,
    R_squared = cell_spec(
      round(r_squared, 4), "html",
      bold      = r_squared == max_r2,
      color     = ifelse(r_squared == max_r2, "white", "black"),
      background = ifelse(r_squared == max_r2, "steelblue", "transparent")
    ),
  ),
```

```

Adj_R_squared = cell_spec(
  round(adj_r_squared, 4), "html",
  bold      = adj_r_squared == max_adj_r2,
  color      = ifelse(adj_r_squared == max_adj_r2, "white", "black"),
  background = ifelse(adj_r_squared == max_adj_r2, "skyblue", "transparent")
),
F_statistic = cell_spec(round(f_stats, 2), "html"),
P_value = cell_spec(
  format.pval(p_values, digits = 4), "html"
),
AIC = cell_spec(
  round(aic_values, 2), "html",
  bold      = aic_values == min_aic,
  color      = ifelse(aic_values == min_aic, "white", "black"),
  background = ifelse(aic_values == min_aic, "orchid", "transparent")
),
BIC = cell_spec(
  round(bic_values, 2), "html",
  bold      = bic_values == min_bic,
  color      = ifelse(bic_values == min_bic, "white", "black"),
  background = ifelse(bic_values == min_bic, "mediumpurple", "transparent")
)
)

kbl(comparison_df, escape = FALSE, format = "html",
     caption = "Comprehensive Comparison of Models") %>%
  kable_styling(full_width = FALSE)
}

#' compare_winter_models_multi
#'
#' Perform cross-validation on winter data (month range can be customized) to compare multiple regression models.
#' This function uses the `train` function from the `caret` package to fit and validate the models.
#'
#' @param demand_data A data frame containing demand (or other dependent variables),
#'                    month information (`month`), and explanatory variables used in the models.
#' @param model_formulas A list of `formula` objects, specifying the structure of each model to be compared.
#' @param model_names A character vector of labels for the models, matching the length of `model_formulas`.
#'
#' @return A formatted table (using `kableExtra`) displaying the cross-validation results for each model,
#'         including RMSE, R^2, MAE, etc., with highlights for the best values.
compare_winter_models_multi <- function(demand_data, model_formulas, model_names = NULL) {

  train_control <- trainControl(
    method = "cv",
    number = 5,
    index = createFolds(as.factor(demand_data$month), k = 5, returnTrain = TRUE)
  )

  all_results <- list()

  for (i in seq_along(model_formulas)) {
    current_formula <- model_formulas[[i]]
    current_model_name <- model_names[i]

    fit <- train(
      form      = current_formula,
      data      = demand_data,
      method    = "lm",
      trControl = train_control
    )

    df_res <- fit$results
    df_res$Model <- current_model_name
  }
}

```

```

    all_results[[i]] <- df_res
  }

final_df <- do.call(rbind, all_results)

# Highlight RMSE
if ("RMSE" %in% colnames(final_df)) {
  min_rmse <- min(final_df$RMSE, na.rm = TRUE)
  final_df$RMSE <- cell_spec(
    round(final_df$RMSE, 4),
    "html",
    bold      = final_df$RMSE == min_rmse,
    color      = ifelse(final_df$RMSE == min_rmse, "white", "black"),
    background = ifelse(final_df$RMSE == min_rmse, "darkgreen", "transparent")
  )
}

# Highlight R^2
if ("Rsquared" %in% colnames(final_df)) {
  max_r2 <- max(final_df$Rsquared, na.rm = TRUE)
  final_df$Rsquared <- cell_spec(
    round(final_df$Rsquared, 4),
    "html",
    bold      = final_df$Rsquared == max_r2,
    color      = ifelse(final_df$Rsquared == max_r2, "white", "black"),
    background = ifelse(final_df$Rsquared == max_r2, "steelblue", "transparent")
  )
}

# Highlight MAE
if ("MAE" %in% colnames(final_df)) {
  min_mae <- min(final_df$MAE, na.rm = TRUE)
  final_df$MAE <- cell_spec(
    round(final_df$MAE, 4),
    "html",
    bold      = final_df$MAE == min_mae,
    color      = ifelse(final_df$MAE == min_mae, "white", "black"),
    background = ifelse(final_df$MAE == min_mae, "tomato", "transparent")
  )
}

kbl(
  final_df,
  caption  = "Cross-Validation Results by Winter Months (Nov-Mar)",
  escape   = FALSE,
  format    = "html",
  row.names = FALSE
) %>%
  kable_styling(full_width = FALSE)
}

```