Modeling Power Plant Energy Output Using Nonlinear Regression

Assignment Report

Submitted By:
Rajesh Kumar Dhimal
Coventry University Student ID: 16542745
Softwarica Student ID: 250072
Faculty of Engineering, Environment and Computing
Coventry University,
Softwarica College of IT and E-commerce

Supervised By:
Lecturer Hikmat Saud
7089 CEM: Introduction to Statistical Methods for Data Science

May 23, 2025

May 23, 2025

Abstract

We were requested to figure out better performing model to predict net hourly electrical energy output of a Combined Cycle Power Plant (CCPP) from given five non-linear regression models. We were requsted to use Ordinary Least Squares Method to estimate optimum model parameters for the given models. The models were evaluated using various performance metrics like Residual Sum of Squares (RSS), log-likelihood (LL), Akaike Information Criterion (AIC), Bayesian Information Criterian (BIC); other regression metrics like Root Mean Squared Error(RMSE), R-squared (R^2) along with analysis of residual distribution.The selected model was then validated by 70-30 train-test split with 95% confidence intervals.

Finally, Approximate Bayesian Computation (ABC) was used to estimate the posterior distributions of key parameters, allowing us to quantify uncertainty in predictions and better understand the system dynamics.

Introduction

In Combined Cycle Power Plants (CCPPs), heat generated from the gas turbine is reused to generate electricity by powering the steam turbine. Doing so enhances the efficiency of the power houses and hence maximizes the output for the same investment.

Our focus in this study was to predict the net energy output of a CCPP using given nonlinear regression models that capture relationships between the given variables and electrical energy output.

Dataset Description

The data consists of more than 9500 hourly observations, each capturing the plant's operating conditions. The data reflects varying environmental and operational conditions like the humidity in the air, vaccum level inside the machine, tempratue of the place, making it suitable for building predictiive models.

The dataset has four input variables (environmental) and one output variable:

- **x1 (T)**: Temperature: outside air temperature.
- **x3 (AP)**: Pressure: inside the gas turbine.
- **x4 (V)**: Exhaust Vacuum: pressure in the condenser of the steam turbine.
- **x5 (RH)**: Relative Humidity: the moisture content of air.
- **y (EP)**: Electricity output (hourly): our target variable.

Task 1: Preliminary Data Analysis

Task 1.1: Data Preparation

The data was uploaded into the system and primary works like data cleaning, data exploration, handeling missing values and handeling duplicates were performed to make it suitable for modeling.

To support further time series analysis, a new 'time' column was added, and columns were re-arranged so the output column `Energy Output (x2)` becomes the last column and `Time` becomes the first; a general convension in machine learning community.

Task 1.2: Time Series Analysis

Time series plots were created for each variable by placing time series values along x-axis and their corresponding readings along y-axis. Time series plots visually represent how the data varies and behaves across the entire collection period.

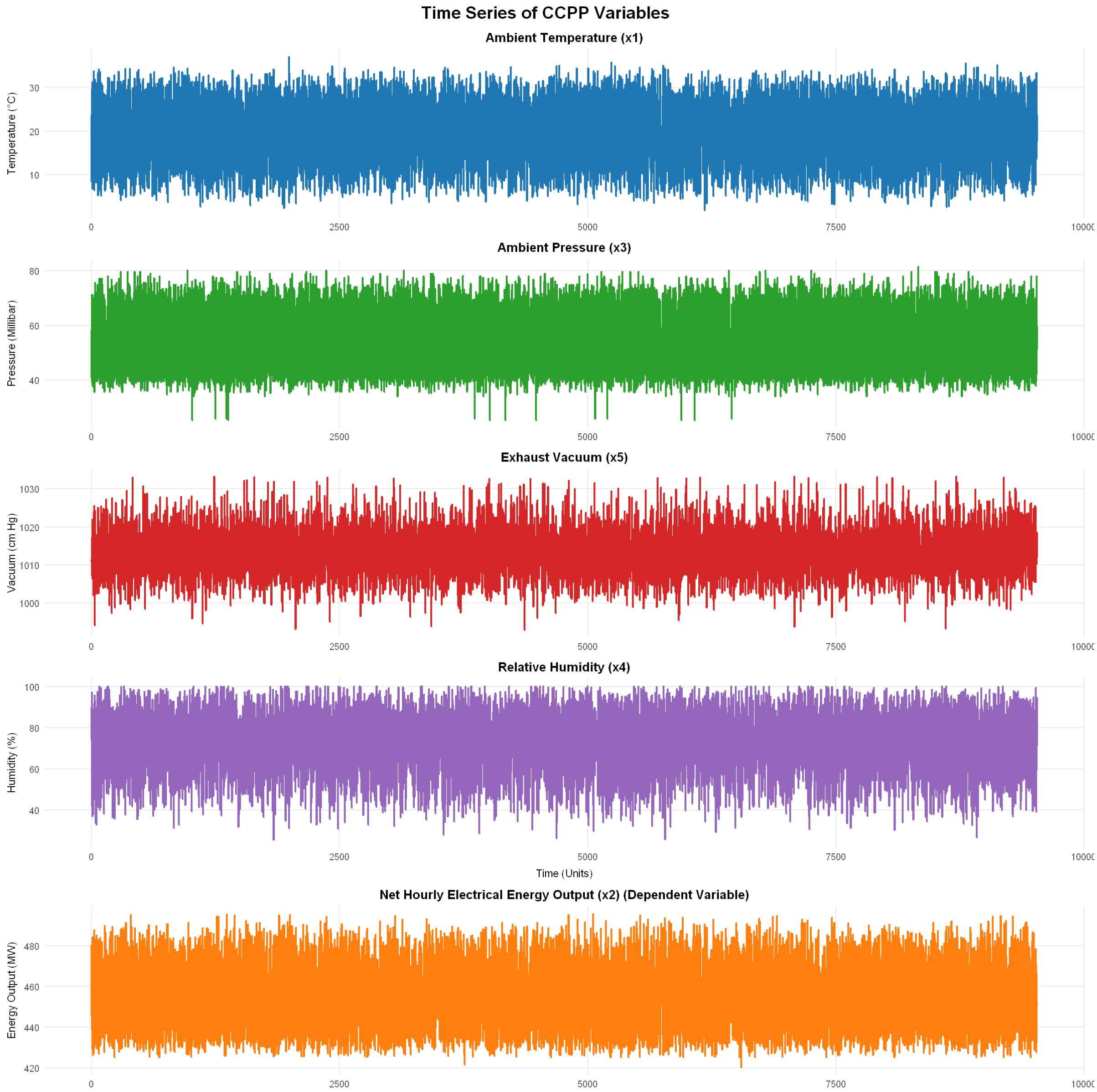


Figure 1: (Time series plots of all five variables stacked vertically: Ambient Temperature (x_1), Ambient Pressure (x_3), Exhaust Vaccum (x_4), Relative Humidity (x_5), and Net Hourly Electrical Energy Output (x_2) plotted against time (in hours))

These plots illustrate how each variable fluctuates over the six-years period from 2006 to 2011. Due to the very high frequency in data, over 9,500 hourly records, all the plots appear very dense and noisy, making it difficult to spot meaningful trends. To address this, a rolling average with a window of 100 observations was applied to each variable.

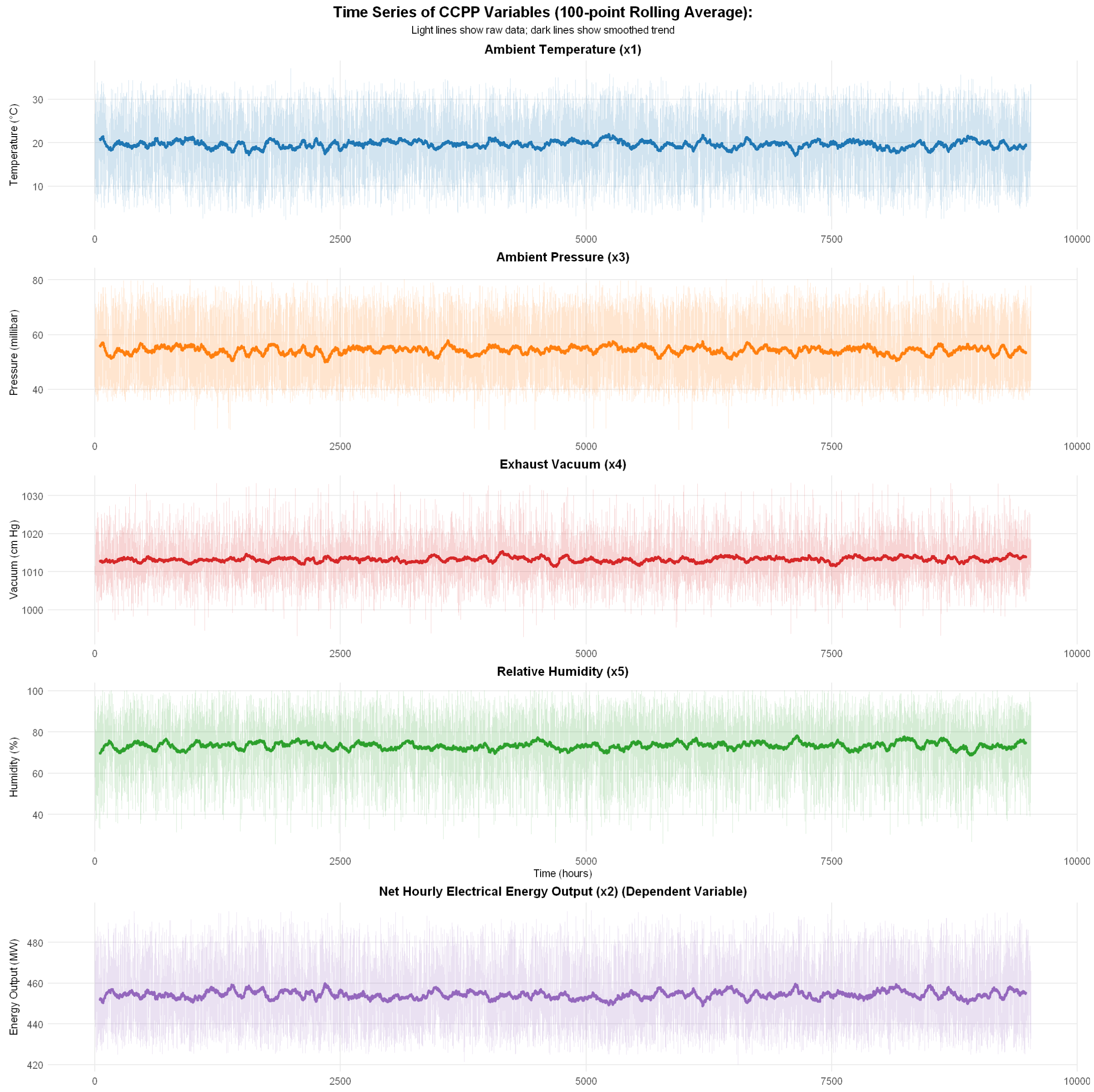


Figure 2: (Time series plots with rolling window of 100 observations)

The smothing rolling average lines in helps better visualize the overall trends but still high clutter is present. To get a more simplified and high-level view of how these variables behave over time, the dataset was divided into segments, where each segment contains 200 consecutive observations. The average of each segment was then calculated and plotted, resulting in a cleaner and more interpretable visualization.

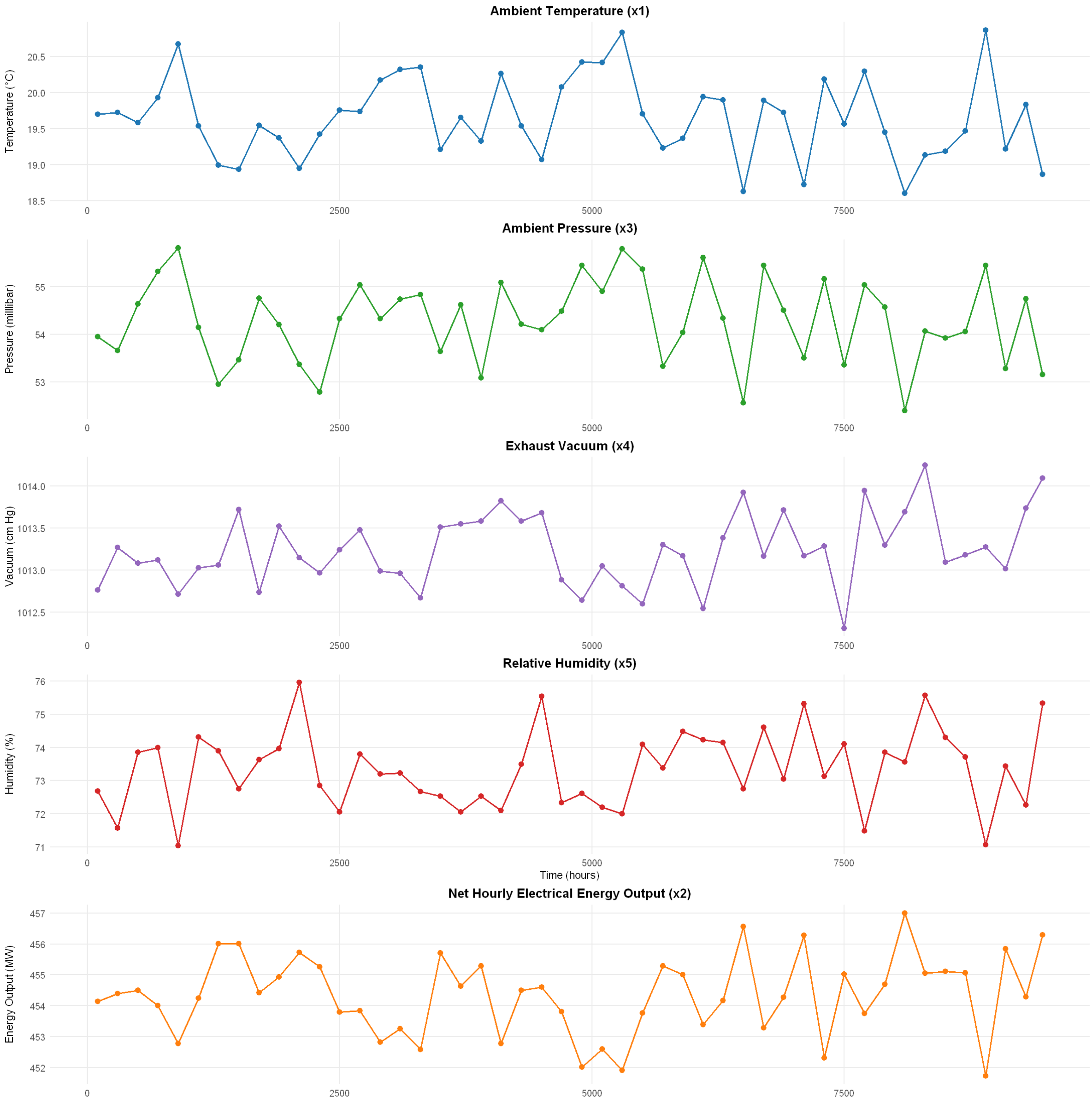


Figure 3: (Time series plots with 200 samples segmentation)

Figure-3 provides a much clear view of the overall trends where data is segmented and plotted, rather than plotting every single point. The ambient temperature (x_1) shows a strong seasonal trend with mean temprature of 19.66 degree celcius, with regular peaks and dips which reflacts yearly weather changes. Exhaust vaccum (x_4) also shows consistent changes over time ranging from 990 cmg to 1140 cm-hg with mean of about 1013 cm-hg, which may be related to other environmental conditions. Ambient pressure (x_3) is relatively stable having mean of 54.29 millibar, it may indicates less influence on energy output. The energy output (x_2) itself also fluctuates, it is likely due to the combined effect of all these variables.

Task 1.3: Distribution Analysis

Histogram is a type of bar chart where data is divided into different segments called bins or buckets and shows how many of the observations fall into each bin. If more data points are concentrated in cretain binning range then this leads to taller bars whereas short bart will be the result of very few observation into that interval. They helps in identifying unusual patterns in the dataset.

Density curves, also called kernal density curves are the smoothened version for histograms. Total area under the density curve is always one representing the whole dataset. The shape of the curve defines the distribution of valaues within the dataset.

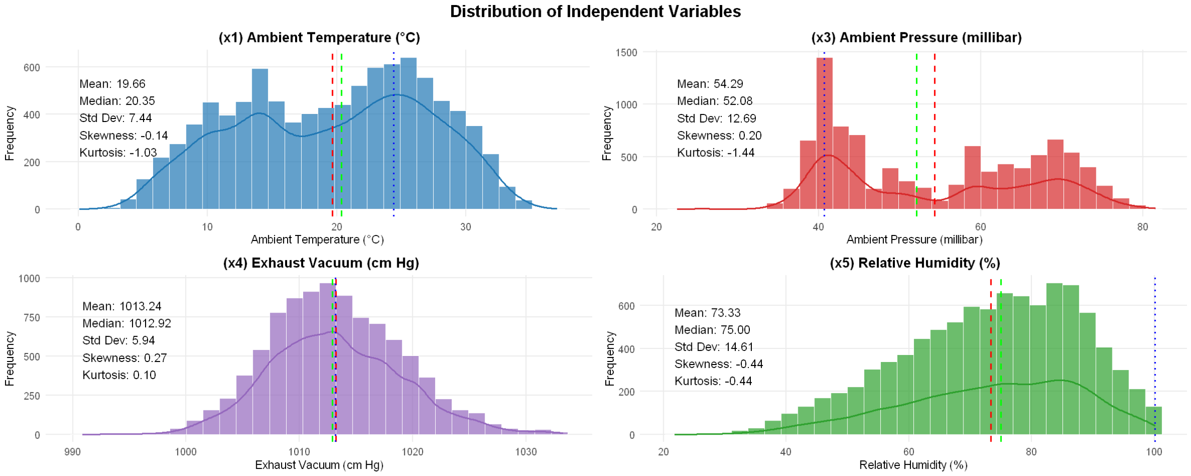


Figure 4: (Distribution of Independent Variables)

Ambient Temperature ($x1$) and Relative Humidity ($x3$) show slight left skewness ($skewness = -0.14$ and -0.44 respectively), which means more values lies along the longer tail on the left side. Ambient Pressure ($x3$) shows high variability ($mean = 54.29$, $skewnwss = 0.20$) without approaching to any of the well known symmetry patterns. Exhaust Vaccum ($x4$) is nearly normally distributed ($mean = 1013.24$, $skewness = 0.27$) showing very little right skewness.

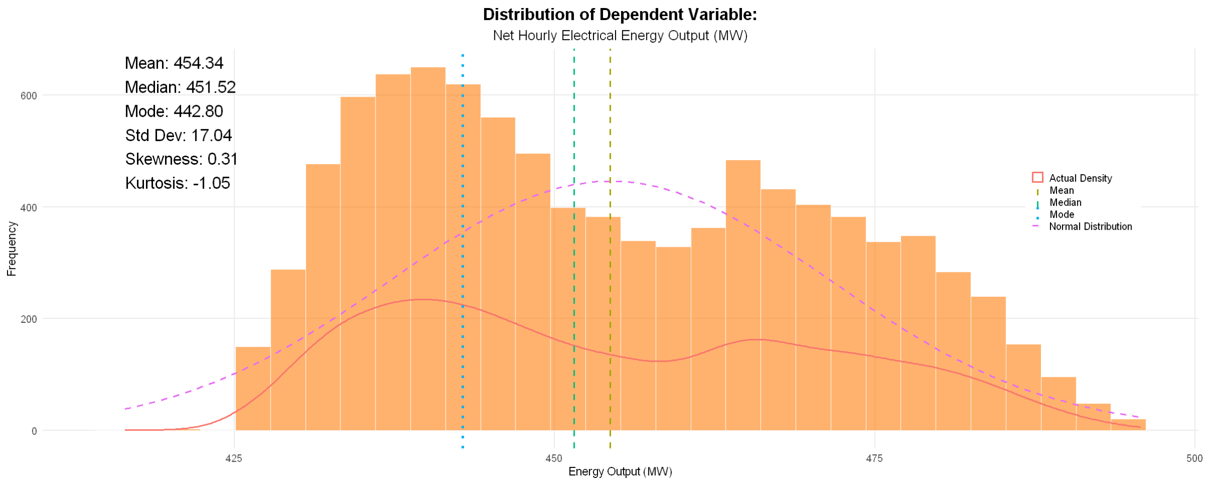


Figure 5: (Distribution energy output)

Figure- 5 shows the distribution of the dependent variable, net energy output ($x2$) ‘MW‘. The distribution for energy output is slightly right-skewed, where most values are in between $430MW$ and $470MW$. The mean energy output is about $454MW$ whereas median is sligtly lower at around $451MW$ which is an indication of slight right skewness. The standard deviation of about $17MW$ also indicates moderate variability in the enorgy output.

Task 1.4: Normality of Data

Q-Q plot, or Quantile-Quantile plot, is a plot used to check whether a dataset follows a particular theoretical distribution; most commonly, the normal distribution. It works by comparing the quantiles of dataset we have to the quantiles of the normal distribution.

To interpret a Q-Q plot, we simply examine how closely the plotted points follow the reference line. If the points lie on the reference line or form clusters around the line then generally, our data follows normal distribution. If the points curve above or below the line then it indictes left or right skewness and so the data is not symmeteically distributed. If the points form a s-shaped curve around the line it is an indication of high kurtosis.

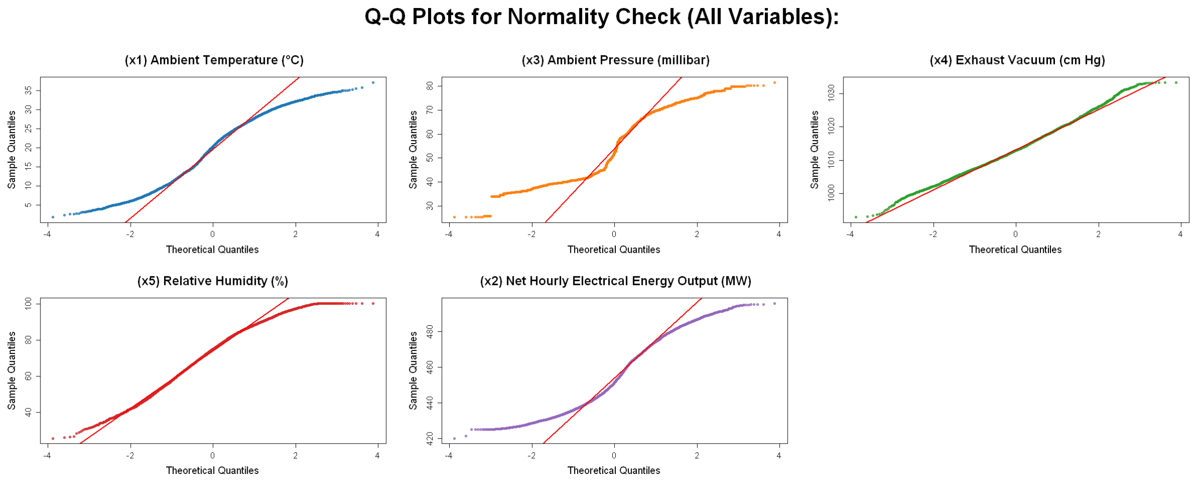


Figure 6: (Normality test for variables)

Figure-6 shows, ambient temprature ($x1$) and relative humidity ($x5$) shows deviations at both tails (clearly visible *S-shape* curves) results in slight deviation from normal distribution. Energy output ($x2$) shows a noticeable long tails (clear *s-shape* curve) on both ends. The ambient pressure ($x3$) is clearly not normally distributed as it shows very high deviations from the reference line. Exhaust vaccum ($x4$) aligns very closely with the red line, it is mostly normally distributed.

Task 1.5: Outliers Detection

The box plot (sometimes also called box-and-whisker plots) shows the spread and summary of data using five summary statistics: maximum, first quantile (Q1), median (Q2), third quantile (Q3) and maximum value.

The observstions that fall outside the range (whisker tips on both sides) are the potential outliers. A large box generally indicates high variability in the data values, one long whisker is an indication of skewness (right skewness or the right skewness) and a symmetrica box shape represents normality.

A violin plot extends the box plot by showing the distribution's shape using kernal density estimation plot. It displays the central tendency, shape and variability of data in a single plot. More wider sections in the violin plot indicates more concentration of observations whereas irregular and multiple dips and valleys indicate complex distribution structure.

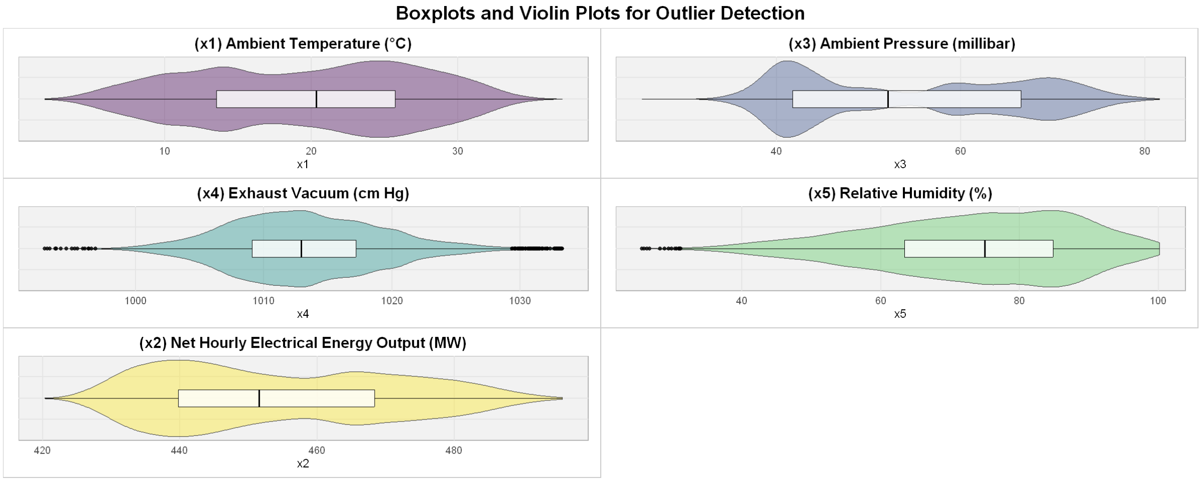


Figure 7: (Box-Violin plots)

From Figure 7; relative humidity ($x5$) have outliers on both tails because dots are present on both ends. Exhaust vacuum ($x4$) also shows several outliers along the left-tail which makes it left-skewed. Ambient pressure ($x3$) also shows potential outliers on both ends. In contrast, ambient temperature ($x1$) and energy output ($x2$) shows relatively fewer extreme values. The shape and distribution of ambient pressure ($x3$) is complex as the distribution is almost divided into two halfs around median (multimodal distribution).

Task 1.6: Scatter Plots and Correlation Analysis

Scatter plot shows individual data points using dots in a 2D plot. Each dot represent a pair of values for two variables (x, y). To interpret the relation, we look for the trend of points like; dots go up from left to right (up-trend), dots go down from right to left (down-trend), dots scattered randomly indictes weak or no relationship between variables. If the points shape into some curve like patterns then it is an indication of non-linear relation or higher-order relation.

Correlation is a way to measure how two variables are related considering all other variables constant. It shows whether an increase in value of one variable is associated with the change in values of other variable.

For correlations between independent and dependent variables (like temperature ($x1$) and energy output ($x2$)), stronger correlations are beneficial. However, strong correlations between independent variables can create a challenge called multicollinearity (when predictor variables are highly correlated with each other).

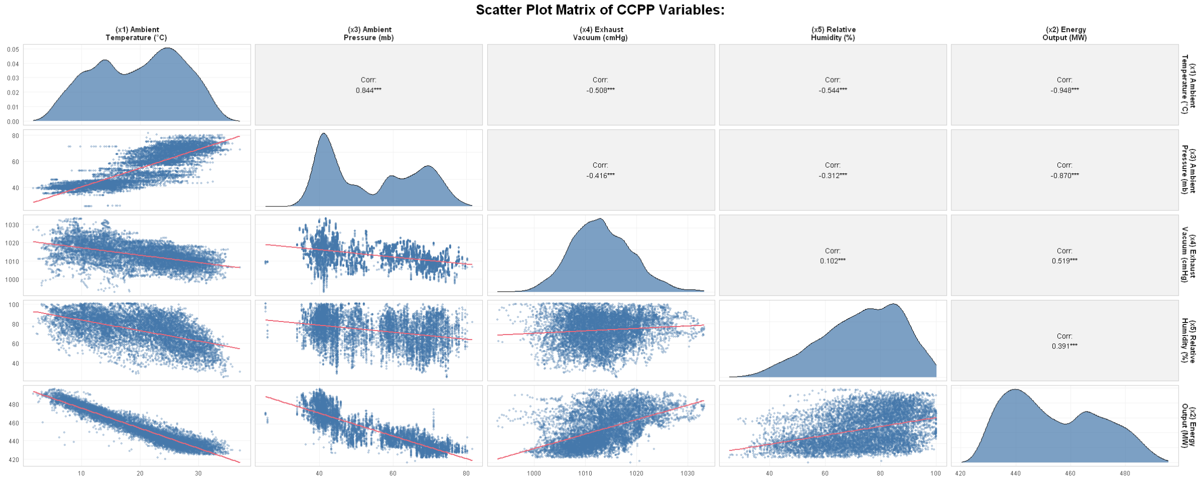


Figure 8: (Scatter plots of all variable combination where, diagonal cells represent density plot for each variable)

In above figure, some environmental factors shows a strong link with energy output. Temperature ($x1$) have a strong negative relationship with energy output (monotonic relation) meaning, increase in one values (temperature) results in decrease in the value of other (energy). Ambient pressure ($x3$) also have a strong relationship with energy ($x2$), and high vacuum ($x4$) inside the chamber generally leads to more energy output (a rule from Thermodynamics). The relation between humidity ($x5$) and energy output ($x2$) is moderate so the influence of humidity is not that noticable compared to effects of temperature and vaccum. The diagonal cells represent the distribution of each variable using kernal density estimation (kde). Temprature ($x1$) and ambient pressure ($x3$) may be multimodal as indicated by multiple peaks.

Pearson's Correlation analysis was performed to further justify the relation between variables and resulting correlations are recorded on the following heatmap:

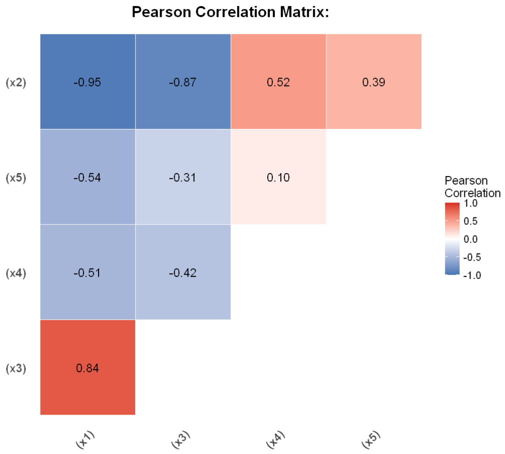


Figure 9: (Pearson's Correlation heatmap between different variables)

To better capture the monotonic relation between variables, Spearman's Rank Correlation analysis was performed and the resulting coefficients are recorded in the following heatmap:

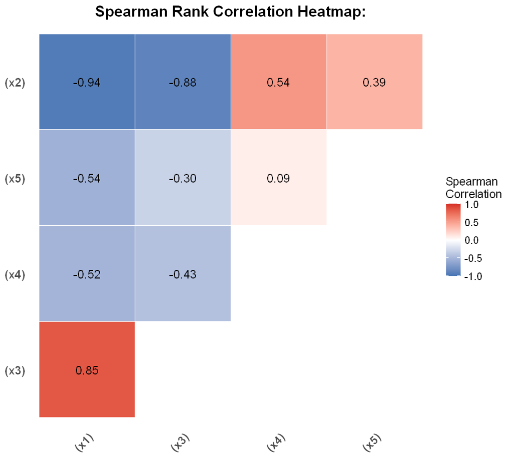


Figure 10: (Spearman's Rank Correlation heatmap between different variables)

A very correlation (Corr $= -0.94$) is observed between temperature (x_1) and output (x_2) , so generally as one value (like (x_1) increases), the other variable value x_2 consistently decreases in a monotonic fashion. So, temperature becomes one of the most influential factors affecting energy output. Similarly, pressure (x_3) and energy (x_2), also have high negative correlation (-0.88), which means as the value of x_3 increases the value in x_2 decreases in a monotonic ways.

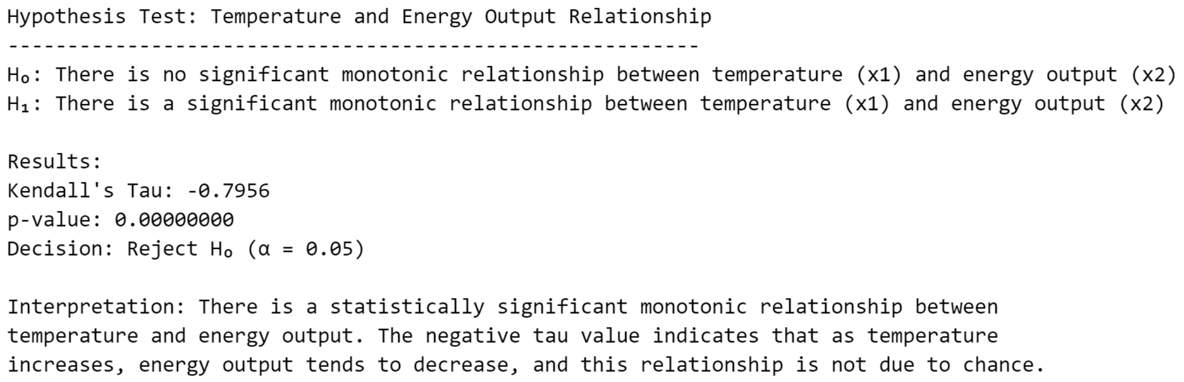
A strong positive correlation (Corr $= 0.85$) is observed between predictor variables temperature (x_1) and humidity (x_3), which indicates as the temperature increases the water (moisture) in the atmosphere also increases in a monotonic fashion (potential multicollinearity case).

Task 1.7: Hypothesis Testing

Hypothesis testing is a way to make decisions using available data instead of guess or relying on intuitions.

We contruct null hypothesis considering ”no-effect” or ”no change” in the outcome and, construct the alternative hypothesis by taking the ”effect of change” or ”what we want to prove”. Then we perform test utilizing the data available. If we found enough evidence to support the default assumption of no changes then, we conclude - ”no evidence to reject the null hypothesis” otherwise we conclude by- ”enough evidence to reject null hypothesis” ie. ”accept alternative.”

A test between temperature ($x1$) and the energy output ($x2$) was conducted using Kendall’s Tau test and the result with its interpretations are presented below:

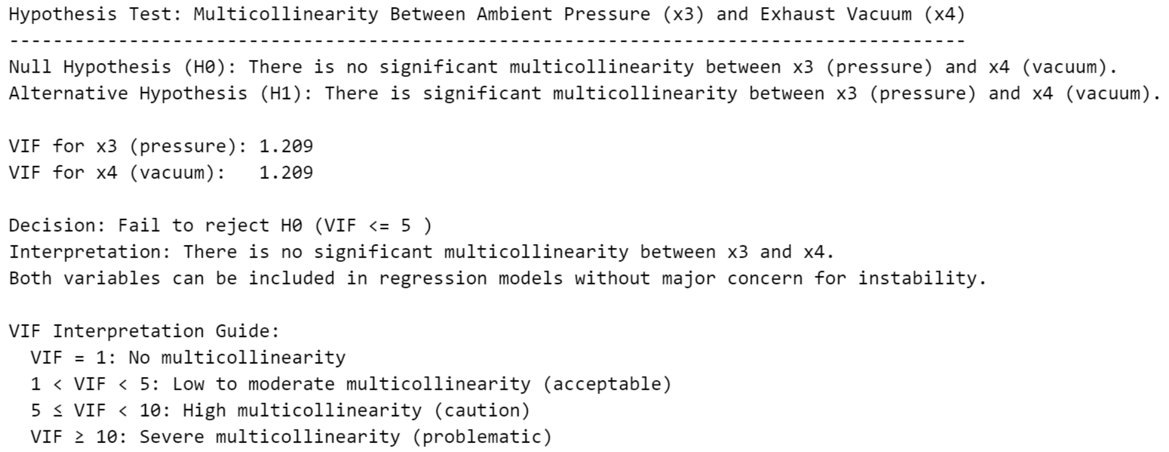


```
Hypothesis Test: Temperature and Energy Output Relationship
-----------------------------------------------------------
H₀: There is no significant monotonic relationship between temperature (x1) and energy output (x2)
H₁: There is a significant monotonic relationship between temperature (x1) and energy output (x2)

Results:
Kendall's Tau: -0.7956
p-value: 0.00000000
Decision: Reject H₀ (α = 0.05)

Interpretation: There is a statistically significant monotonic relationship between
temperature and energy output. The negative tau value indicates that as temperature
increases, energy output tends to decrease, and this relationship is not due to chance.
```

Also another test for the multicollinearity between pressure (x_3) and vaccum (x_4) was conducted using VIF and the result obtained is shown below:



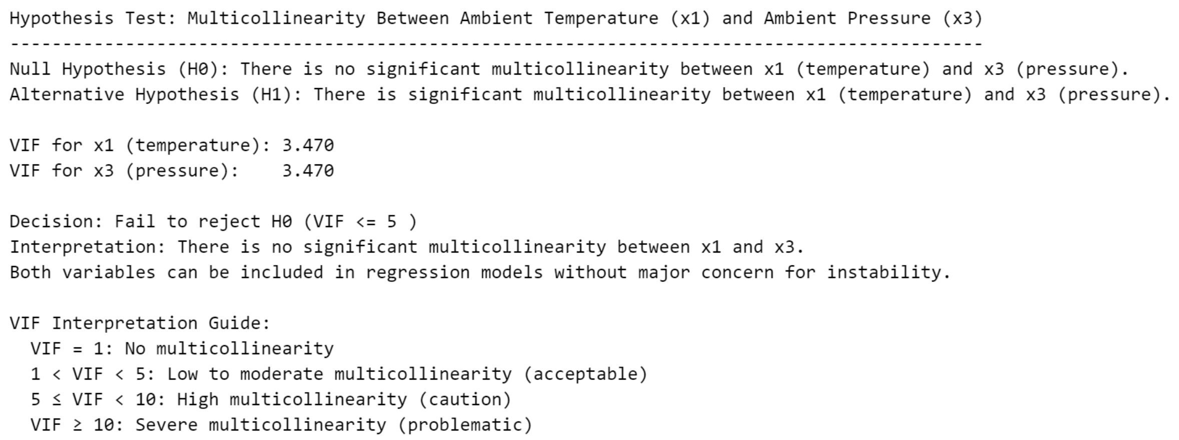
```
Hypothesis Test: Multicollinearity Between Ambient Pressure (x3) and Exhaust Vacuum (x4)
-----------------------------------------------------------------------------------
Null Hypothesis (H0): There is no significant multicollinearity between x3 (pressure) and x4 (vacuum).
Alternative Hypothesis (H1): There is significant multicollinearity between x3 (pressure) and x4 (vacuum).

VIF for x3 (pressure): 1.209
VIF for x4 (vacuum):   1.209

Decision: Fail to reject H0 (VIF <= 5 )
Interpretation: There is no significant multicollinearity between x3 and x4.
Both variables can be included in regression models without major concern for instability.

VIF Interpretation Guide:
  VIF = 1: No multicollinearity
  1 < VIF < 5: Low to moderate multicollinearity (acceptable)
  5 ≤ VIF < 10: High multicollinearity (caution)
  VIF ≥ 10: Severe multicollinearity (problematic)
```

And, another test using VIF for multicolinearity between temprature ($x1$) and pressure ($x3$) was conducted and the result is shown below:



```
Hypothesis Test: Multicollinearity Between Ambient Temperature (x1) and Ambient Pressure (x3)
-------------------------------------------------------------------------------------------
Null Hypothesis (H0): There is no significant multicollinearity between x1 (temperature) and x3 (pressure).
Alternative Hypothesis (H1): There is significant multicollinearity between x1 (temperature) and x3 (pressure).

VIF for x1 (temperature): 3.470
VIF for x3 (pressure):    3.470

Decision: Fail to reject H0 (VIF <= 5 )
Interpretation: There is no significant multicollinearity between x1 and x3.
Both variables can be included in regression models without major concern for instability.

VIF Interpretation Guide:
  VIF = 1: No multicollinearity
  1 < VIF < 5: Low to moderate multicollinearity (acceptable)
  5 ≤ VIF < 10: High multicollinearity (caution)
  VIF ≥ 10: Severe multicollinearity (problematic)
```

Task 2: Model Development and Evaluation

The following models were evaluated and estimated the parameters for:

- Model 1: $y = \theta_1 x_4 + \theta_2 x_3^2 + \theta_{\text{bias}}$
- Model 2: $y = \theta_1 x_4 + \theta_2 x_3^2 + \theta_3 x_5 + \theta_{\text{bias}}$
- Model 3: $y = \theta_1 x_3 + \theta_2 x_4 + \theta_3 x_5^3$
- Model 4: $y = \theta_1 x_4 + \theta_2 x_3^2 + \theta_3 x_5^3 + \theta_{\text{bias}}$
- Model 5: $y = \theta_1 x_4 + \theta_2 x_1^2 + \theta_3 x_3^2 + \theta_{\text{bias}}$

Task 2.1: Parameters Estimation

To find optimal model parameters, we were requested to use a direct relation called normal equation. This method helps to calculate optimal model parameters without running loops or without using trial-and-error methods. We organize essential input data into a metrix form called "design metrix". In design matrix, each row represents one observation or one data point. Each column represents a feature including a column of 1's for intercept or bias.

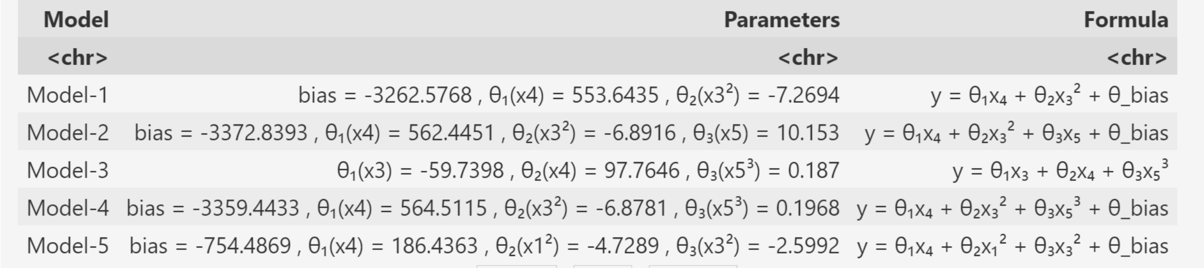
The mathemetical expression for Ordinary Least Squares is:

$$\boldsymbol{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Where, $\boldsymbol{\theta}$ is the vector of estimated model parameters, $\mathbf{X}$ is the design matrix containing the input features or independent variables, and $\mathbf{y}$ is the vector of observed target values.

In R, a function estimate_parameters() was defined to apply OLS, where operations like matrix multiplications, matrix transpose and matrix inverions were performed. The solve() method was used to perform inverse matrix operations.

`estimate_parameters()` was called for each model using their respective design matrix to estimate the model parameters. The `estimate_parameters()` function actually performs the model training behind the scenes and finds the optimal model parameters for each model. The calculated model parameters are listed in a tabular structure for easy reference:



Model	Parameters	Formula
<chr>	<chr>	<chr>
Model-1	bias = -3262.5768 , θ_1(x4) = 553.6435 , θ_2(x3^2) = -7.2694	$y = \theta_1 x_4 + \theta_2 x_3^2 + \theta\_bias$
Model-2	bias = -3372.8393 , θ_1(x4) = 562.4451 , θ_2(x3^2) = -6.8916 , θ_3(x5) = 10.153	$y = \theta_1 x_4 + \theta_2 x_3^2 + \theta_3 x_5 + \theta\_bias$
Model-3	θ_1(x3) = -59.7398 , θ_2(x4) = 97.7646 , θ_3(x5^3) = 0.187	$y = \theta_1 x_3 + \theta_2 x_4 + \theta_3 x_5^3$
Model-4	bias = -3359.4433 , θ_1(x4) = 564.5115 , θ_2(x3^2) = -6.8781 , θ_3(x5^3) = 0.1968	$y = \theta_1 x_4 + \theta_2 x_3^2 + \theta_3 x_5^3 + \theta\_bias$
Model-5	bias = -754.4869 , θ_1(x4) = 186.4363 , θ_2(x1^2) = -4.7289 , θ_3(x3^2) = -2.5992	$y = \theta_1 x_4 + \theta_2 x_1^2 + \theta_3 x_3^2 + \theta\_bias$

Figure 11: (Optimed Model Parameters Calculated Using OLS)

The estimated optimum model parameters fits into the model equation as follows (used model-1 as an example):

> $EnergyOutput = -3262.576790 + 553.643467 \times (vaccum) - 7.269425 \times (pressure^2)$

The *bias* (-3262.576790) is the intercept term and it represents the baseline energy output when all other variables are zero. Theoretically this is the energy output when all other parameters set to constant or zero. But, practically, this will never happen because $Exhaust\ vacuum\ (cmHg)$: $always > 0$ (total vacuum is impossible) and; $Pressure\ (millibars)$: $always > 0$ (vacuum would imply system failure).

The coefficient for relative humidity ($x4 = 8.79$) indicates assuming all other variables constant, for 1 unit increase in humidity results in 8.79 units increase in energy.

The coefficient for square of pressure ie. $x_3^2 = -0.35$ means as the square of pressure increases, the energy output decreases. This is a non-linear relationship.

Task 2.2: Residual Sum of Squares (RSS)

In regression, a residual is the difference between the model prediction and give actual value. This difference tells us how far our model prediction is for each point. The residual for $j - th$th point is calculated as:

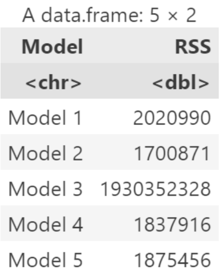
$$e_j = y_j - \hat{y}_j$$

Where; (y_j) is the actual value for $j - th$ point and, ($\hat{y}_j$ is the predicted value for $j - th$ point.

We take the sum of the squares of all the residuals to measure the overall error in the prediction. A smaller value means the predictions closer to actual values.

$$RSS = \sum_{j=1}^{n} (y_j - \hat{y}_j)^2$$

Here, residuals are squared so they do not cancel out each other, and to give more weight to larger errors.



A data.frame: 5 × 2

Model	RSS
<chr>	<dbl>
Model 1	2020990
Model 2	1700871
Model 3	1930352328
Model 4	1837916
Model 5	1875456

Smaller residual sum of squares means the model is making small errors. Here from above comparision table, model 5 has the smallest RSS value ($RSS = 198462.5$) followed by model 4 ($RSS = 515095.0$) and model 2 ($RSS = 516315.9$). Model 3 ($RSS = 568957.4$) has the higest value among all the given models. Considering RSS values, model 5 is performing well while model 3 is lagging well behind possibly due to the abscence of bias term.

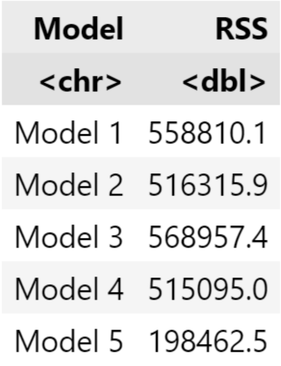
Task 2.3: Log-Likelihood

Log-Likelihood tells- how good is the model at explaining the data we have observed. It assumes the residuals are normally distributed. LL measures the closeness of predicted values to the actual values by taking into account how spread out the residuals are (the variance).

The mathemetical expression to calculate log-likelihood is:

$$\log L = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where: n = number of observations, σ^2 = variance of residuals, y_i = actual value, $\hat{y}$ = predicted value.



Model	RSS
<chr>	<dbl>
Model 1	558810.1
Model 2	516315.9
Model 3	568957.4
Model 4	515095.0
Model 5	198462.5

High log-likelihood value means the model fits the data better. Model 5 has comperatively largest value of ($LL = -27982.45$) followed by model 4 (-32525.65). The model 3 has shows smallest log-likelihood value (-32999.4). Based on the LL-values, model 5 is again performing well. And, model 3 is not of the great choice again possible because this model is not capturing the underlying relation between the environmental variables and the output.

Task 2.4: AIC and BIC

AIC is a numerical value that helps us decide which model fits the data best. It takes into consideration two things: how well the model fits the data and how simple the model is. BIC and AICs are similar concepts but they differ in one thing; BIC gives more penalty for complex models and prefers simpler models even more.

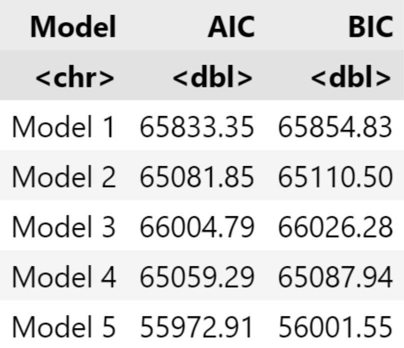
Interpretation is similar: low is better. Both are used to compare different models; and they are not to judge single model alone.

The formulas to calculate these Valuesare as follows:

$$\text{AIC} = 2k - 2\log L$$

$$\text{BIC} = k\log n - 2\log L$$

where, k: number of model parameters, n: number of observations, and $\log L$: log-likelihood calculated earlier.



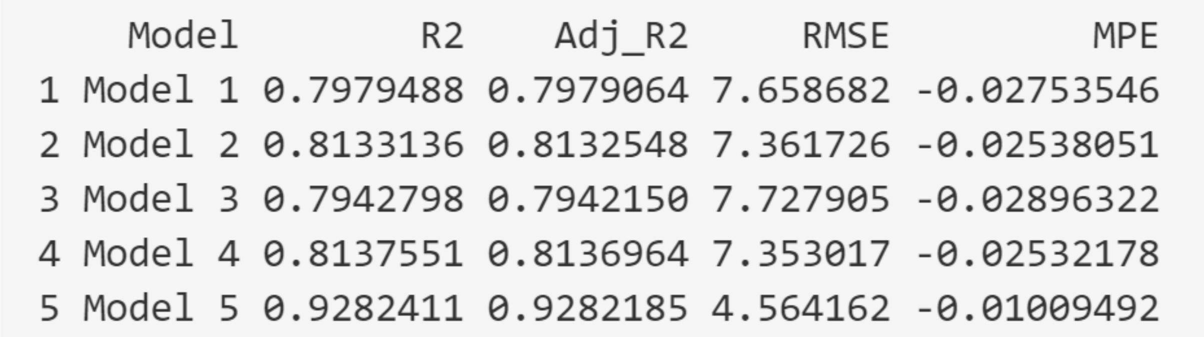
Model	AIC	BIC
<chr>	<dbl>	<dbl>
Model 1	65833.35	65854.83
Model 2	65081.85	65110.50
Model 3	66004.79	66026.28
Model 4	65059.29	65087.94
Model 5	55972.91	56001.55

Model 5 is clearly better with $AIC = 55972.91$ and $BIC = 56001.55$. Model 4 is just behind model 5 with $AIC = 65059.29$ and $BIC = 65087.94$. Model 3 is still lagging and reserves the fifth position with $AIC = 66004.79$ and $BIC = 66026.28$. All the performance metrices indicates model 5 is performing comparatively well whereas model 3 is performing poorly.

Task 2.5: R-squared(R^2), Adjusted R-squared (R^2_{adj}), Root mean squared error ($RMSE$) and Mean percentage error (MPE)

Some other regression metrics like R-squared(R^2), adjusted R-squared (R^2_{adj}), root mean squared error ($RMSE$) and mean percentage error (MPE) were also calculate to support the regression metrics we have calculated above (RSS, LL, AIC and BIC). Detailed explaination along with how to calculate each metric is given in the appendix.

Below table represents a summary of the statistics that we have calculated which acts as the additional performance comparision metrics:



```
    Model        R2    Adj_R2     RMSE         MPE
1 Model 1 0.7979488 0.7979064 7.658682 -0.02753546
2 Model 2 0.8133136 0.8132548 7.361726 -0.02538051
3 Model 3 0.7942798 0.7942150 7.727905 -0.02896322
4 Model 4 0.8137551 0.8136964 7.353017 -0.02532178
5 Model 5 0.9282411 0.9282185 4.564162 -0.01009492
```

Figure 12: (R-Squared, Ajusted R-Squared, RMSE, MPE)

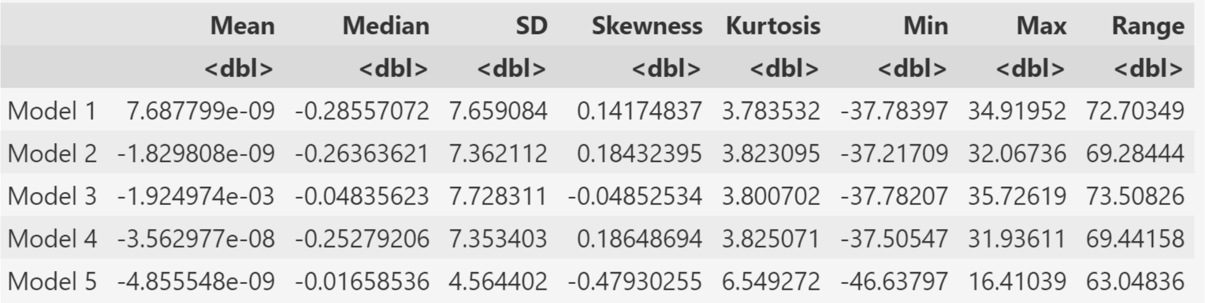
Model 5 again is better-performing model, with the highest R-squared (= 0.9282) and adjusted R-squared (= 0.9282), indicating it explains the largest proportion of variance in the target variable. It also has the lowest RMSE (= 4.5642) and smallest MPE (= −0.0101), meaning it's the most accurate and least biased in predictions.

Model 3 is again the worst-performing model, with the lowest R-squared (= 0.7943), lowest adjusted R-squared (= 0.7942), highest RMSE (= 7.7279), and most negative MPE (= −0.0290). This indicates that it explains the least variance, makes the largest errors, and has the greatest bias among all models. Model 3 is likely missing important nonlinear terms or interactions that other models (especially Model 5) capture.

Task 2.6: Distribution of Residuals

In regression, the residuals should be normally distributed with zero mean and a constant spread or variance across all values, this phenomenon is called homoscedasticity. If the residuals shows some pattern then it generally indicates that the model is missing something like missing important variables, very complex relationship between variable which is not being captured by the current modeling approach or the situation where predictor variables are highly correlated with each other.

We calculated some basic statistics on residuals like mean, median, standard devition, skewness etc. and their summary is presented in the table below:



	Mean	Median	SD	Skewness	Kurtosis	Min	Max	Range
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Model 1	7.687799e-09	-0.28557072	7.659084	0.14174837	3.783532	-37.78397	34.91952	72.70349
Model 2	-1.829808e-09	-0.26363621	7.362112	0.18432395	3.823095	-37.21709	32.06736	69.28444
Model 3	-1.924974e-03	-0.04835623	7.728311	-0.04852534	3.800702	-37.78207	35.72619	73.50826
Model 4	-3.562977e-08	-0.25279206	7.353403	0.18648694	3.825071	-37.50547	31.93611	69.44158
Model 5	-4.855548e-09	-0.01658536	4.564402	-0.47930255	6.549272	-46.63797	16.41039	63.04836

Figure 13: (Residual Statistics Comparision Table)

To get an idea of how the the residuals are distributed, histogram along with density and rug plots were constructed and present in the following plot:

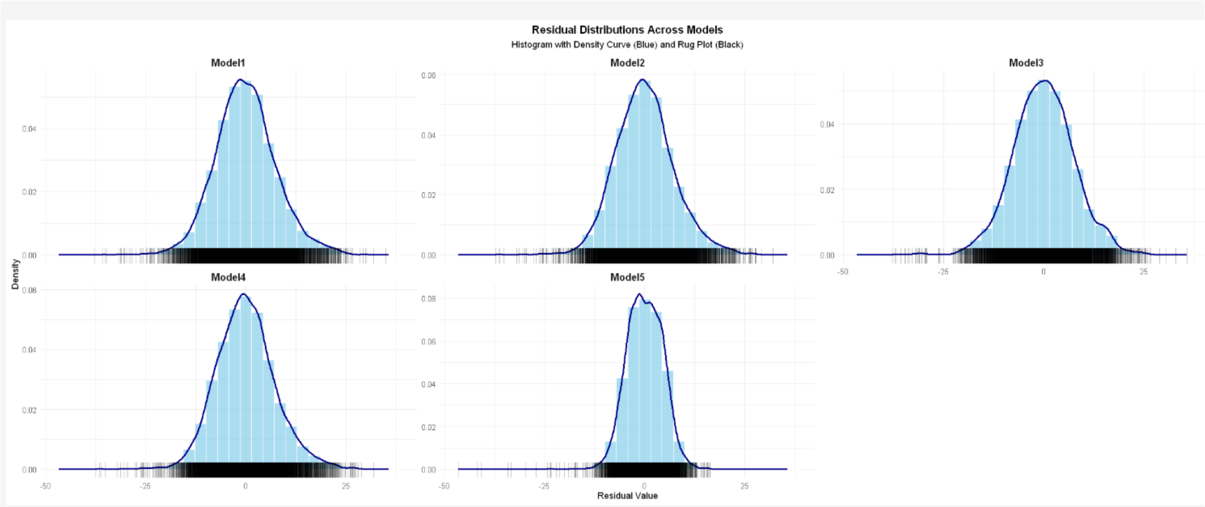


Figure 14: (Residual Statistics Comparision Table)

Residuals were standerzed using z-score standerization and QQ-plot for each of the models were drawn, where, straight diagonal line represents the theoretical quantiles and colored points represents the actual quantiles of residuals for each model:

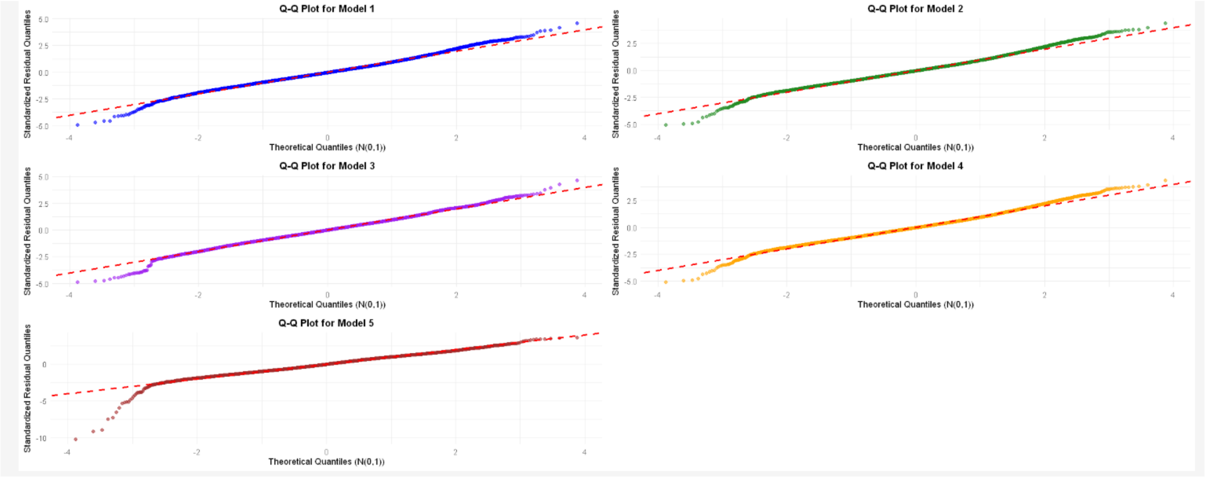


Figure 15: (QQ-Plots of Standerized Residuals)

Almost all models show slight deviations at the lower left tails, suggesting mild to moderate tails and potential outliers. The residuals for model 1 and 2 are closely alligned with reference line with very little deviations at the tail ends. Models 3 and 4 also have moderate deviations at the ends. Model 5 has higest deviations at the left-tail end. It is also notable that residuals distribution for each model deviates from normality only at left tail, otherwise there is almost a perfect normality.

From normality of residuals, model 1 and 2 are performing well whereas model 5 shows the higest deviations at the left tail.

Residuals vs. Fitted Values Plot

Residual vs. fitted values curve is a scatter plot where predicted values are on the x-axis and residuals on the y-axis. Each point on the curve represents how far the model prediction is from the actual value.

If a model is a good fit then; residuals should be randomly scattered around zero, there should be no clear pattern or shape and, the spread of the residuals should be roughly constant across all the fitted values.

Relationship between residuals and fitted values were throughly studied, and the plots are presented below:

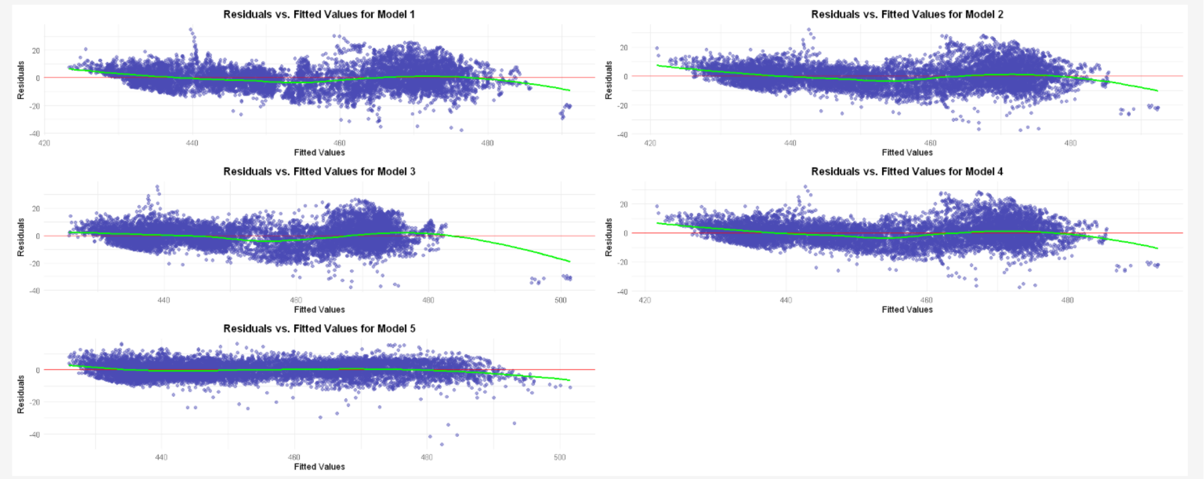


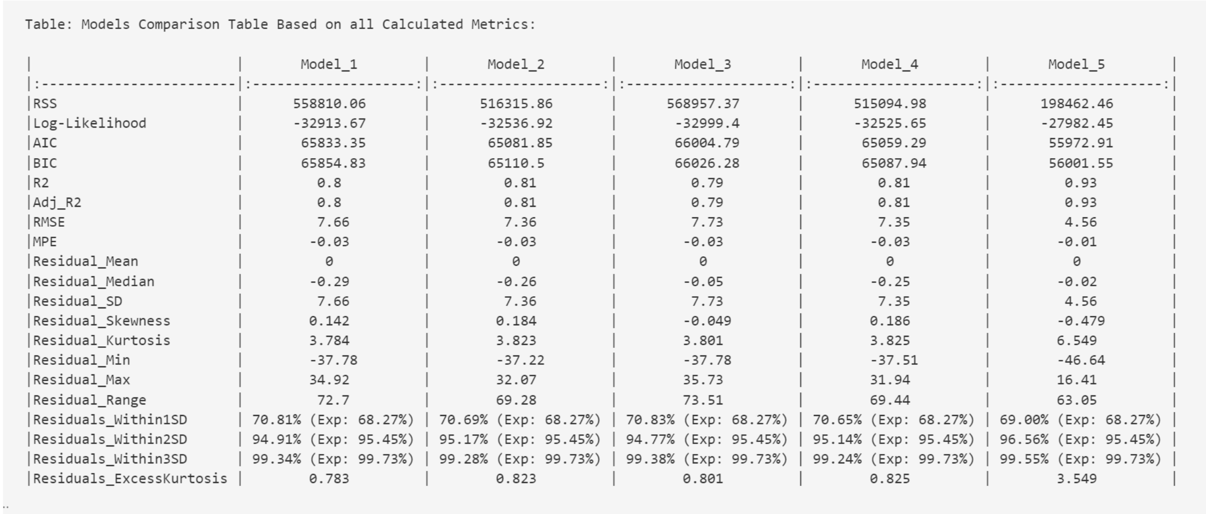
Figure 16: (Comparision: (Residuals vs. Fitted Values))

For model 5, except some outliers, majority of points are grouped arround the reference line (red line) with constant variance, no curve or funnel shape is seen around the reference line. The LOESS (green line) is almost parallel and overlapping with the reference (red line); indicating model 5 is performing well compared to other models.

By analyzing the residual vs. fitted values curves; model 5 emerges out as the better performing model among the given models.

Task 2.7: Selecting Better Performing Model

The following table provides comparison for all candidate models. It includes key performance metrics such as RSS, AIC, BIC, log-likelihood values (LL), residual statistics along with MSE, $R - squared$:



```
Table: Models Comparison Table Based on all Calculated Metrics:

|                         |       Model_1        |       Model_2        |       Model_3        |       Model_4        |       Model_5        |
|:------------------------|:--------------------:|:--------------------:|:--------------------:|:--------------------:|:--------------------:|
|RSS                      |      558810.06       |      516315.86       |      568957.37       |      515094.98       |      198462.46       |
|Log-Likelihood           |      -32913.67       |      -32536.92       |       -32999.4       |      -32525.65       |      -27982.45       |
|AIC                      |       65833.35       |       65081.85       |       66004.79       |       65059.29       |       55972.91       |
|BIC                      |       65854.83       |       65110.5        |       66026.28       |       65087.94       |       56001.55       |
|R2                       |         0.8          |         0.81         |         0.79         |         0.81         |         0.93         |
|Adj_R2                   |         0.8          |         0.81         |         0.79         |         0.81         |         0.93         |
|RMSE                     |         7.66         |         7.36         |         7.73         |         7.35         |         4.56         |
|MPE                      |        -0.03         |        -0.03         |        -0.03         |        -0.03         |        -0.01         |
|Residual_Mean            |          0           |          0           |          0           |          0           |          0           |
|Residual_Median          |        -0.29         |        -0.26         |        -0.05         |        -0.25         |        -0.02         |
|Residual_SD              |         7.66         |         7.36         |         7.73         |         7.35         |         4.56         |
|Residual_Skewness        |        0.142         |        0.184         |        -0.049        |        0.186         |        -0.479        |
|Residual_Kurtosis        |        3.784         |        3.823         |        3.801         |        3.825         |        6.549         |
|Residual_Min             |        -37.78        |        -37.22        |        -37.78        |        -37.51        |        -46.64        |
|Residual_Max             |        34.92         |        32.07         |        35.73         |        31.94         |        16.41         |
|Residual_Range           |         72.7         |        69.28         |        73.51         |        69.44         |        63.05         |
|Residuals_Within1SD      | 70.81% (Exp: 68.27%) | 70.69% (Exp: 68.27%) | 70.83% (Exp: 68.27%) | 70.65% (Exp: 68.27%) | 69.00% (Exp: 68.27%) |
|Residuals_Within2SD      | 94.91% (Exp: 95.45%) | 95.17% (Exp: 95.45%) | 94.77% (Exp: 95.45%) | 95.14% (Exp: 95.45%) | 96.56% (Exp: 95.45%) |
|Residuals_Within3SD      | 99.34% (Exp: 99.73%) | 99.28% (Exp: 99.73%) | 99.38% (Exp: 99.73%) | 99.24% (Exp: 99.73%) | 99.55% (Exp: 99.73%) |
|Residuals_ExcessKurtosis |        0.783         |        0.823         |        0.801         |        0.825         |        3.549         |
..
```

Figure 17: (Comparision of Models based on their performance Statistics)

For model 5, we have the higest AIC ($= 55,972.91$) and BIC ($= 56,001.55$) values, have lowest RSS ($= 198,462.46$) value and higest LL value ($= -27,982.45$) among all the contestent models. These values indicates a strong balance between model fit and simplicity. Model 5 also have higest adjusted R-squared ($= 0.93$) R-squared ($= 0.93$) values. Model 5 is also best in prediction which have lowest RMSE ($= 4.56$) and RSD ($= 4.56$) which indicates very small average errors. MPE is ($= -0.01$) which is very close to zero so the model has very little bias in predictions.

QQ-plot for model 5 is also acceptable but is not best among other models, model-1 and model-2 have more normal-like curves. Model 5 show deviations from normal look particularly at the lower left tail but still it is good overall. Plotted histograms and density curves are also acceptably normal which otherwise shows mild skewness.

From residuals vs. fitted values curve; for model 5, except some outliers, majority of points are clustered around the reference line (red line) with constant variance, no curve or funnel shape of the points around the line.

Based on these evidence - metrics, plots and assumptions; model 5 is better performing model among all. It fits the data well and, is statistically sound and practically useful.

Task 2.8: Train-Test Validation

The data was divided into train and test splits where 70% of data was used to estimate the model parameters (for the selected model-2) and remaining 30% was used to evaluate the model.

Task 2.8.1: Evaluation on Training Data

The selected model 5 was then trained on training dataset (obtained from the splitting) using OLS method and then predictions were made using the estimated parameters. A analysis on the residuals was conducted by plotting histograms with density and rug plots:

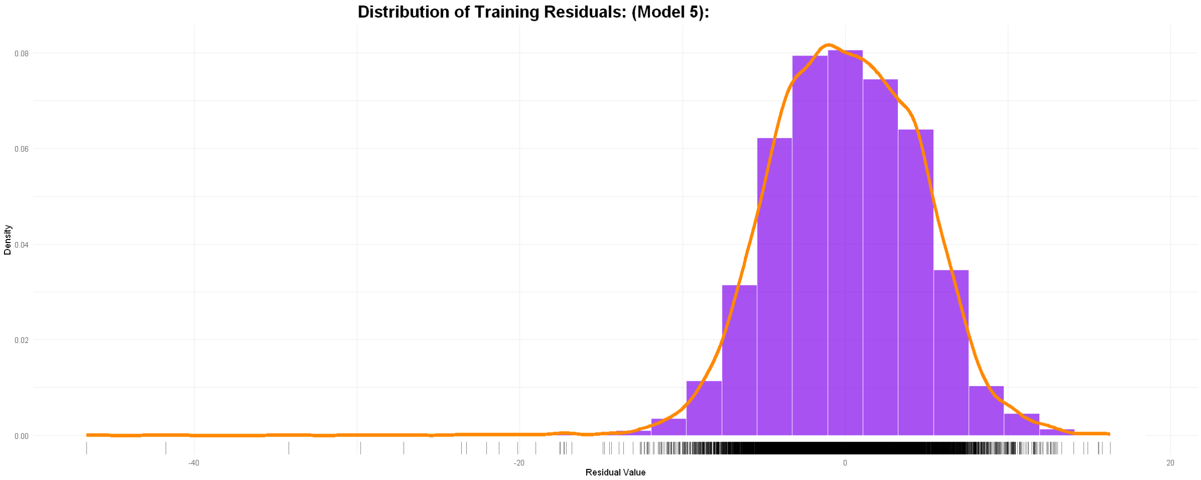


Figure 18: (Distribution of training residuals)

Due to a few outliers, the distribution appears left-skewed but if we omit about 5-6 residuals then it is very close to normal distribution. The residuals form roughly symmetrical (balenced on both sides of zero) curve (omitting few extreme residuals on left side of the curve). It has only one peak (which means it is unimodal), and the plot is almost centered around zero ($mean = 0$). The distribution is comparable to the normal distribution.

Scatter plot was created which compares the actual and predicted values for the training dataset. This get an idea of how well model is performing on the training data.

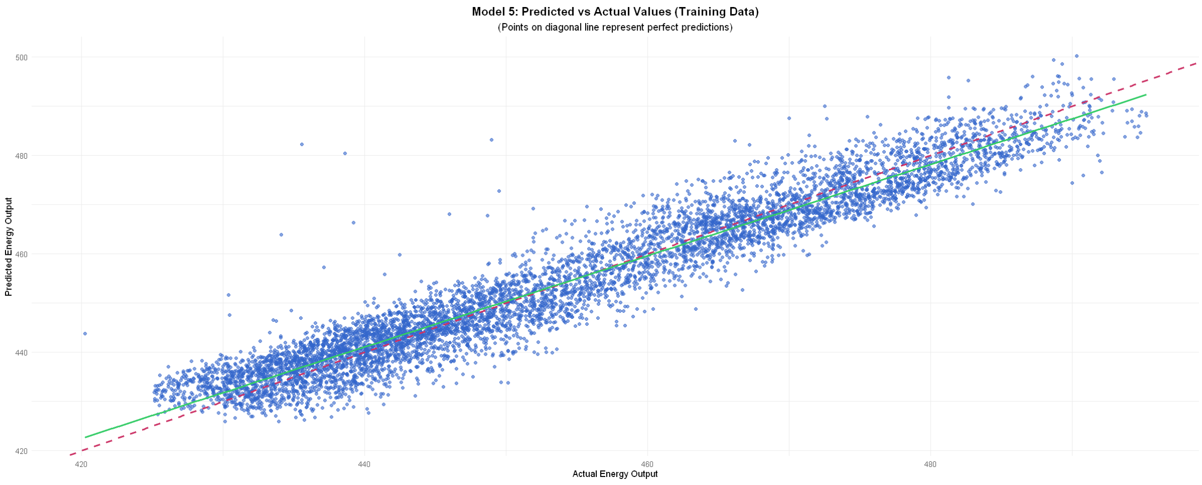


Figure 19: (Scatter Plot: Actual vs. Predicted values)

Almost all points are clustered around the diagonal line with constant variance, which means model's predictions are highly accurate. Very few points are scattered randomly which is normal because it is imposible to have perfect prediction on each and every observation. The reference line (red) and the overall trend line are parallel and almost overlap further indicating high prediction accuracy.

Next, the 95% confidence intervals for the predictions were computed and plotted alongside actual and predicted values. Also, to enhance visualization, a zoomed-in view was generated by displaying the first 100 observations:

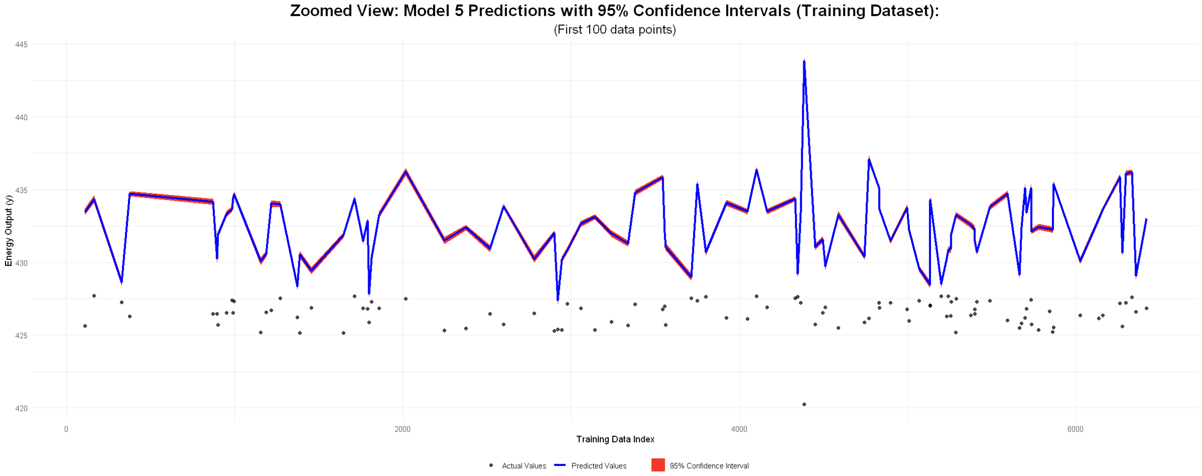
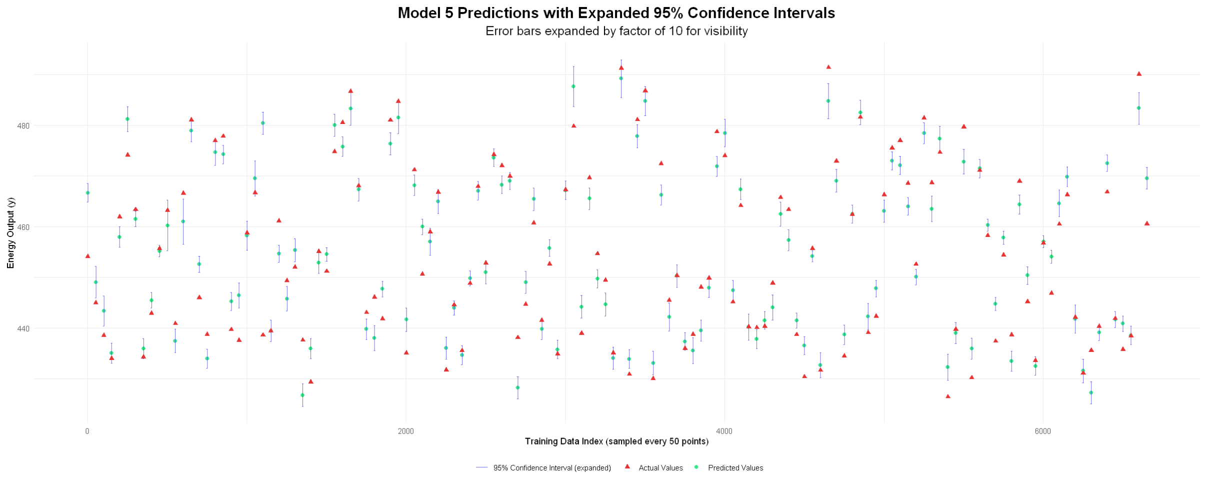


Figure 20: (Model-5 predictions with 95% Confidence Intervals)

The blue line shows the predicted energy output, the red ribbon indicates the 95% confidence interval, and the black dots represent the observed values. Most of the black dots fall below the blue line, so the model slightly over predicts the energy outputs. But, since this segment is a small portion of the larger training dataset, such over-predictions can be ignored.

Even when zoomed in, the confidence intervals remain very narrow, indicating low uncertainty in the predictions. This suggests the model performs well on the training data.

For the 95% confidence intervals, error bars were plotted and expanded by a factor of 10 to make them visible on the plot. These enlarged intervals were intended solely for better visualization and do not represent the true statistical width.



Green dots represent predicted values and the red triangles show the actual observed energy output. The blue error bars, scaled up ten times, highlights 95% confidence intervals for predictions. Most actual values are close to the predictions, suggesting good model accuracy. Few red triangles lie outside the blue bars, indicating potential outliers or regions where the model's performance is less reliable.

Task 2.8.2: Evaluation on Testing Data

The optimum model parameters from Task 2.7.1 were used to make predictions on the testing data. Distribution plots for actual and predicted values were then studied to analyze whether the model captures the overall structure of the data.

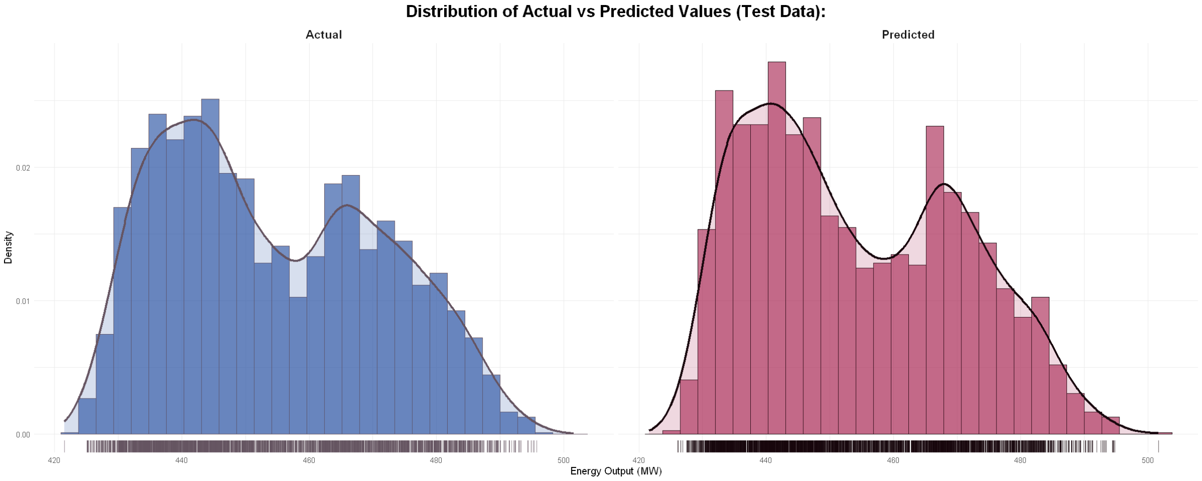


Figure 21: (Distribution of Actual vs. Predicted Values for test data)

Left curve shows the histogram (with density curve overlayed over it) for the actual values. The distribution is not symmetrical, it possess characterstics of bimodal distribution. The rugplot at the buttom indicates the values are somehow distributed widely (between 430 to 490) along x-axis.

The distribution of predicted values for the testing dataset almost resembles with left curve (curve of actual values). The shape is not normal, this curve also possess the properties of bimodal distribution. Rug plot at bottom shows, the predictions are more centered around 430 − 490 range. Both the curves are similar in shape and appearance indicating model is making highly accurate predictions and because of high accuracy curves are almost identical.

Also, the residual were calculated and their distribution was plotted to check for their normality (a key assumption in Regression):

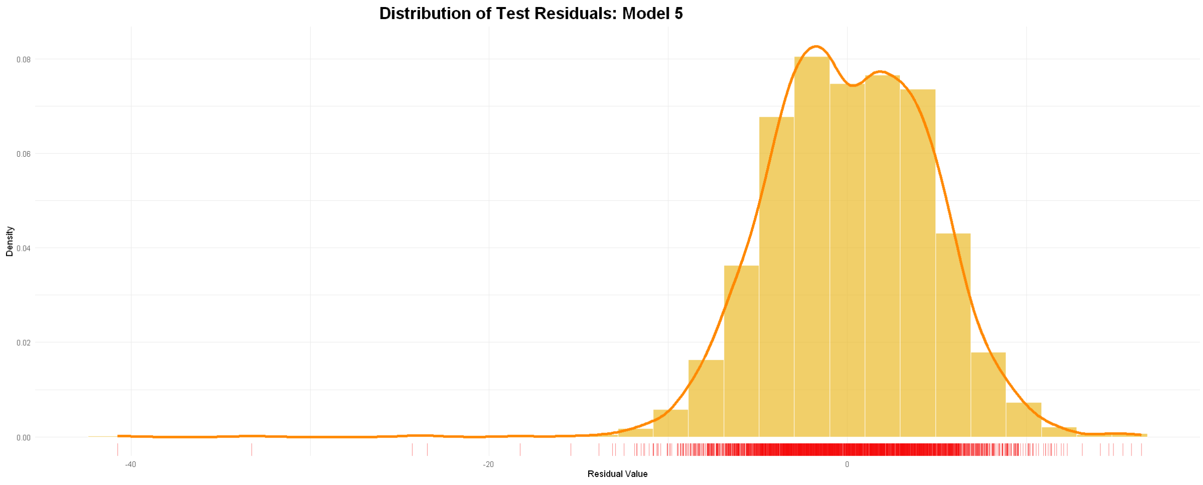


Figure 22: (Normality of residuals for test data)

Residuals are distributed around center zero ($mean = 0$) which indicates the model is unbiased. The distribution is almost symmetrical around zero with approximate bell-shape; excluding those five extreme residual values at the left tail. The distribution is acceptably normal which is the assumption in regression analysis. Model is performing well in the training dataset.

A comparison between predicted and actual values for test dataset was conducted using scatter plot to check if they are close enough:

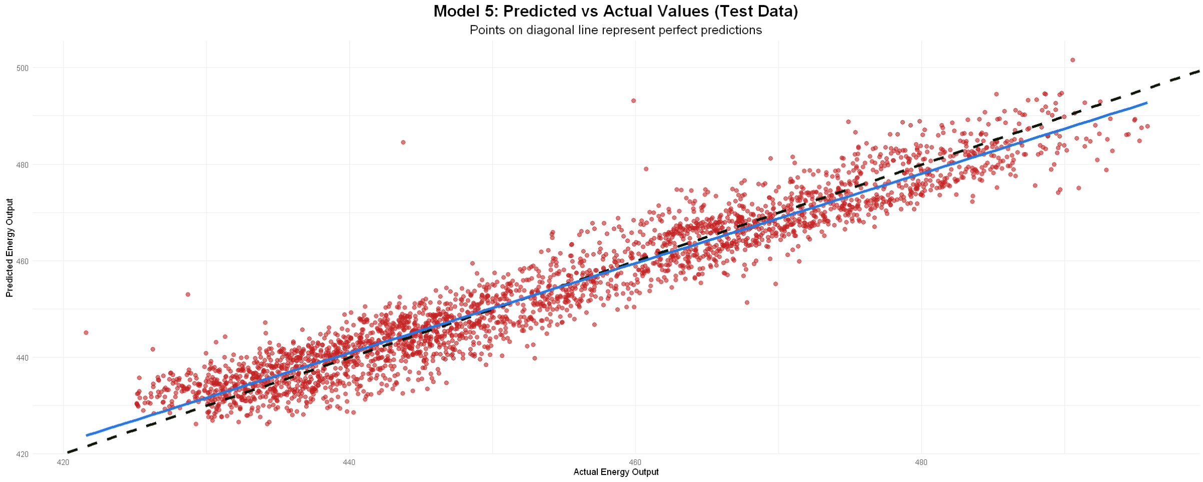


Figure 23: (Predicted vs. Actual values for test data)

The general trend line (blue line) and the reference line (black dashed line) are parallel and almost overlap on each other. Also, the actual values (red dots) are linearly clustered around the reference line with constant variance. There are very few points on the curve which are randomly scattered. This is very accurate model with high percision.

A through study on 95% confidence interval were performed and found that the intervals are quite narrow with mean width of about (0.4207347) , maximum width of about (1.301707) , and minimum width of (0.2222644). Narrow confidence intervals means model is making predictions with high level of ceretainty. To make error bars and CI visible, I zoomed-in the y-axis by a certain factor. Zooming y-axis may results into cutting-off the values from the screen frame but this is only for visualization purpose:

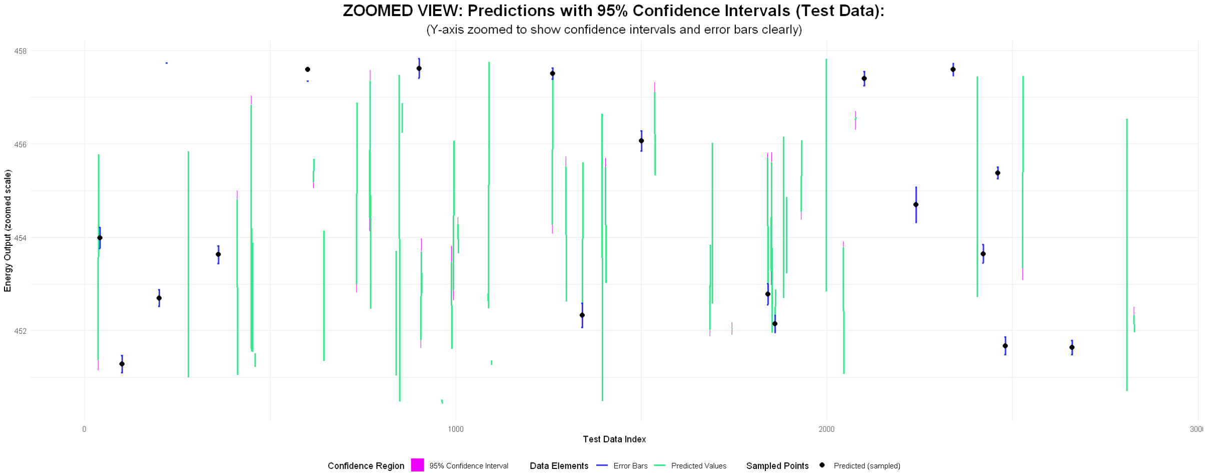


Figure 24: (95% Confidence Interval with Error Bars (Zoomed y-axis))

Even after zooming the y-axis, the confidence intervals and error bars are still very narrow and requirs through investigation to see them properly.

A small sample of 16 data points $(100 - 115)$ was choosen and plotted to visualize how narrow our Confidence Intervals for the predictions were:

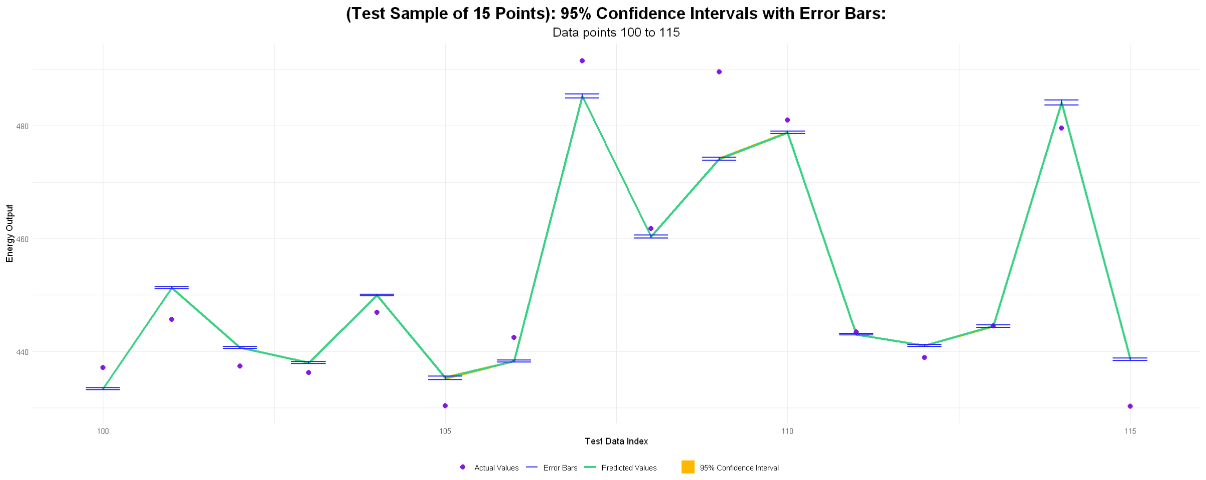


Figure 25: (95% Confidence Interval with Error Bars for 16 data points)

Error bars are denoted by blue marks which shows how narrow they are compare to the full range of energy output values. The intervals appear very thin and only cover less then 1% of the total range.

So, our model 5 is performing well on both training data as well as on the testing datasets. Predictions are very accurate with very narrow confidence intervals where predictions almost resembles the true values.

Task- 3: Approximate Bayesian Computation

ABC is used when we want to do basian inference; instead of calculating the likelihood, we simulate data from our model and compare with the real data. In ABC, we first guess a value for the parameter, then, we use that value to simulate the data. Then, we compare the simulated data with real data. If the simulated data is close enough, we accept the parameter. We repeate this process many times to build the posterior distribtion.

Our chosen model was Model-5:

$$y = \theta_1 \cdot x_4 + \theta_2 \cdot x_1^2 + \theta_3 \cdot x_3^2 + \theta_{\text{bias}}$$

where the optimized model parameters we obtained from Task 2.1 are:

$$\theta_{\text{bias}} = -754.486865, \quad \theta_1\ (x_4) = 186.436251, \quad \theta_2\ (x_1^2) = -4.728874, \quad \theta_3\ (x_3^2) = -2.599210$$

Varying θ_1 and θ_2

Two parameters having higest absolute values were selected for constructing the posterior distribution where, other parameters were kept fixed. Uniform distribution was used to construct the prior distribution for the selected parameters θ_1 and θ_2 within 20% around those values. Appropriate sample size and acceptance thresold were choosen to perform ABC rejection sampling to simulate the real data for θ_1 and θ_2.

Posterior distribtion for parameter θ_1 was plotted and graph obtained is shown in the following figure:

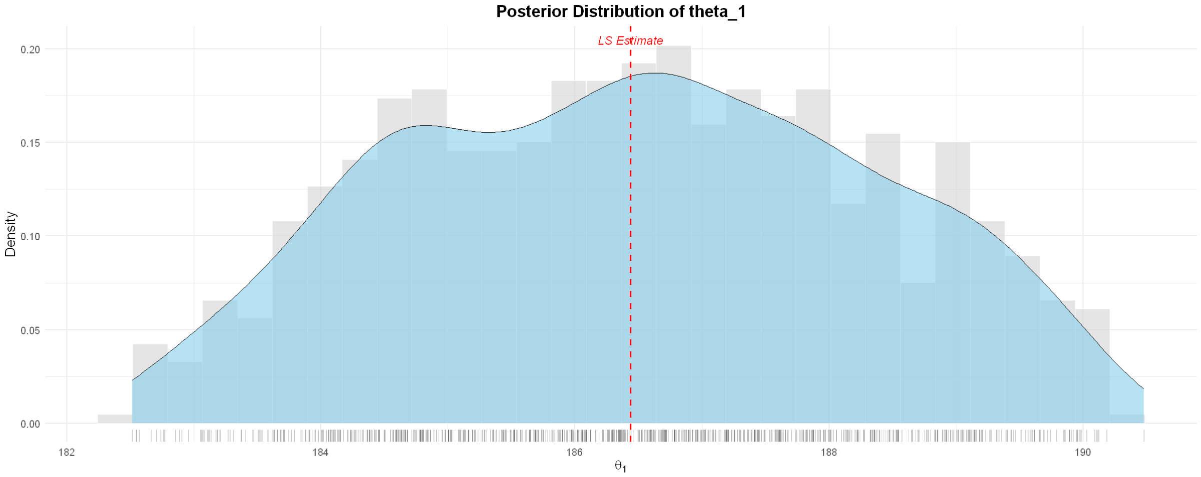


Figure 26: (Posterior Distribution of θ_1)

The distribution is almost bell-shaped and symmetrical, having the Least Squares ($LS =$ 186.43) estimate value near its center. The majority of the backbone's mass is located in a small range, indicating a high level of parameter estimation confidence. The ABC posterior supports the LS estimate, with little evidence for values far from the LS solution.

Plotting of the posterior distribution for parameter θ_2 produced the graph that is displayed in the following figure:

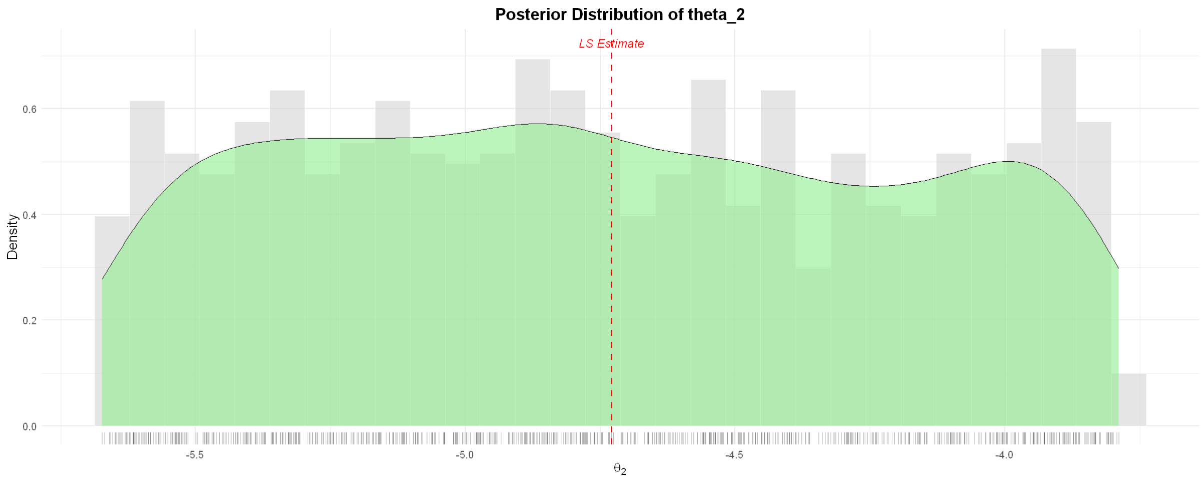


Figure 27: (Posterior Distribution of θ_2)

This distribution is neary flat almost like a uniform type of curve as (θ_2 values ranges for a wide range of values), an indication of high uncertainty for θ_2. The posterior is spread widely around the least square estimate value for θ_2. The distribution suggests many values for θ_2 are acceptable. The data provides very little information to preciously estimate θ_2.

A joint posterior distribution for θ_1 and θ_2 is plotted for both accepted and rejected posterior samples:

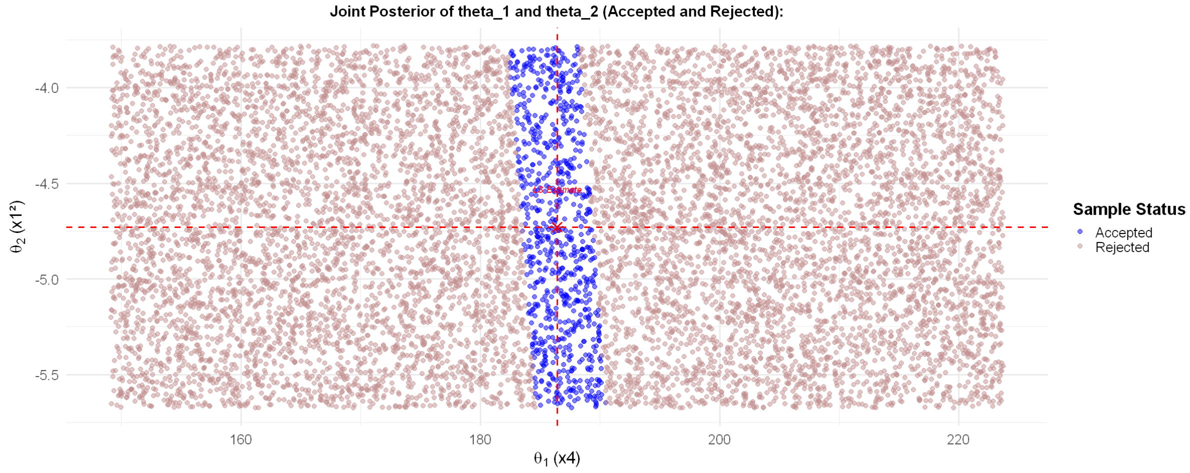


Figure 28: (Joint Posterior of θ_1 and θ_2 (accepted + rejected))

The accepted samples form a narrow vertical band ($\theta_1 = (183 - 190)$) arond the Least Squares line indicating only a limited range of θ_1 values produce good model fit. The effect of θ_2 is much less certain and many vales (the limit for θ_2 values is open as seen from the joint distribution indicating an uncertain range of values) of θ_2 can fit the observed data.

Fixed θ_1 and θ_2

Now, ABC estimation is performed by varying *bias* and θ_3, keeping θ_1 and θ_2 fixed. Joint posterior distribution obtained for accepted samples is shown below:

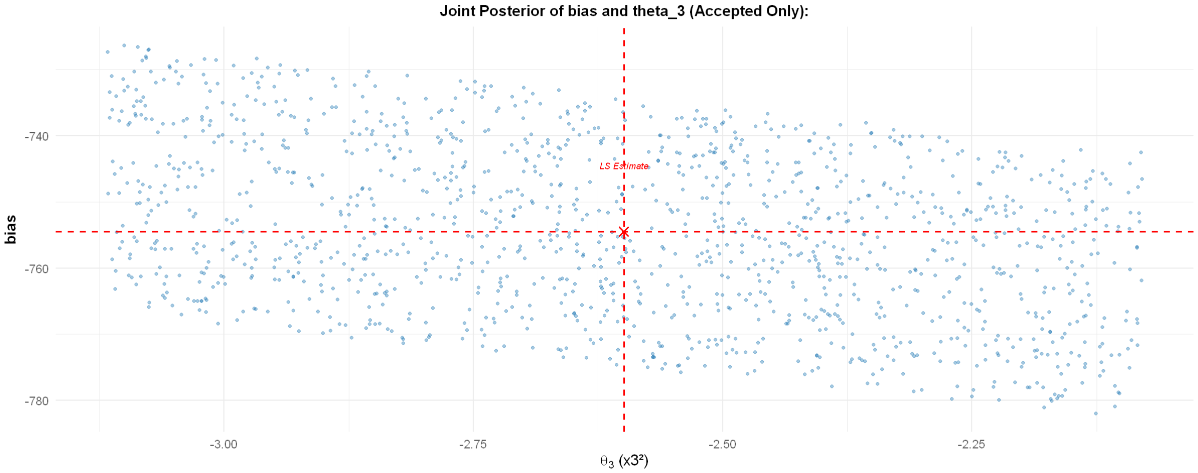


Figure 29: (Joint Posterior of *bias* and θ_3 (accepted only))

There is no clear relationship or pattern between *bias* and θ_3; accepted values are scattered in a rectangle. The LS (*bias* = -754.49, θ_3 =-2.60) estimates (red lines) are in the center of the accepted region, showing that the LS solution is supported by the Bayesian approach.

Many combinations of *bias* and θ_3 fit the data well, and the model is not very sensitive to small changes in these parameters within the accepted range.

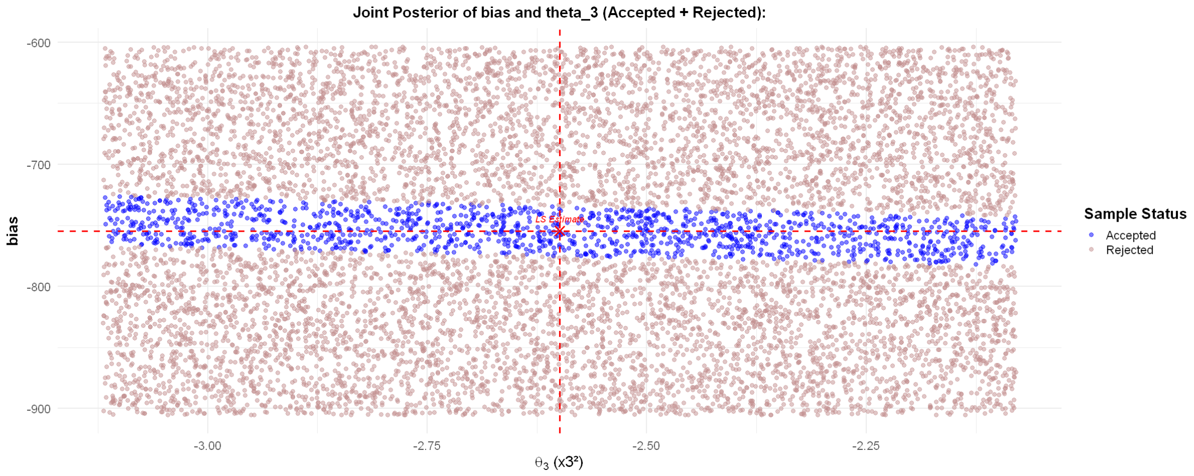


Figure 30: (Joint Posterior of *bias* and θ_3 (accepted + rejected))

Accepted samples (blues) are tightly clustered horizontally from about $\theta_3 = -3.12$ to -2.08 and vertically from about $bias = -905$ to -603. Rejected samples (brown) fill the rest of the rectangle, showing that most combinations do not fit the data well.

There is no visible relationship between *bias* and θ_3; accepted values form a horizontal band. The LS estimates are in the center of the accepted region, confirming the LS solution is supported by the Bayesian approach.

Only a narrow range of *bias* values (around -754.49) with a wide range of θ_3 values fit the data well, and the model is not sensitive to $theta_3$ within this accepted range.

Limitations

While this study provides a well performing regression model for predicting power plant energy output, some limitations of the study should not be ignored. First, this analysis is based on historical data generated from a powerplant, which certainly limits generalizability to other plants or regions. Second, the model assumes that relationships between variables remain stable over time; any structural changes or unmeasured factors could affect predictive accuracy. Third, although the model performs well on available data, it may not capture rare or extreme operational conditions not present in the dataset. Finally, the uncertainty present in some parameter estimates (e.g., squared temperature effect) suggests that further data or alternative modeling approaches could improve reliability.

Conclusion

Model 5 emerges out as the better regression model because it consistently outperformed all other candidate models across multiple evaluation metrics. It achieved the lowest Residual Sum of Squares ($RSS = 198,462$), the highest $R - squared$ and $Adjusted - R - squared$ (both = 0.93), and the lowest AIC and BIC values ($AIC = 55,973$, $BIC = 56,002$). It explains about 93% of the variance in energy output, with minimal prediction error and optimal balance between fit and complexity.

Confidence intervals for predictions were very narrow, reflecting high certainty in model estimates. The model also shows excellent generalization, performing equally well on both training and testing datasets. The predicted vs. actual plots showed points tightly clustered around the diagonal, confirming high predictive accuracy and no signs of overfitting.

Approximate Bayesian Computation (ABC) analysis confirmed the reliability of the least squares (LS) parameter estimates. The posterior distributions for the main parameters were centered around the LS solutions, and the LS point always lay within the high-density region of the posterior. This agreement between Bayesian and LS approaches increases confidence in the model's parameter estimates.

When varying θ_1 (exhaust vacuum) and θ_2 (squared temperature), ABC showed θ_1 is tightly constrained by the data (narrow posterior), while θ_2 is weakly identified (wide, flat posterior), indicating high uncertainty. When θ_1 and θ_2 were fixed and $bias$ and θ_3 were varied, the accepted region was a horizontal band: $bias$ was tightly constrained, but θ_3 could vary widely. There was no strong dependency between parameter pairs in either case, and many combinations fit the data well within the accepted region.

The model's predictive power is driven mainly by exhaust vacuum (x_4), while the effect of squared temperature ($x_1 - squared$) is less certain and should be interpreted with caution.

Appendix:

All R scripts and supplimentary materials required for data processing, modeling, plotting and pictures used in the LaTex file were provided in the following github repository:

https://github.com/rajesh-coventry/STW7089CME-Modeling-Power-Plant-Energy-Output-Using-Nonlinear-Regression

References

1. Tufekci, S. (2014). *Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. International Journal of Electrical Power & Energy Systems*, **60**, 126–140. https://doi.org/10.1016/j.ijepes.2014.02.027

2. Kaya, A., & Tüfekci, S. (2015). *Performance evaluation of regression models for prediction of power output of combined cycle power plants. Energy*, **88**, 417–425. https://doi.org/10.1016/j.energy.2015.05.098

3. Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis* (3rd ed.). Wiley.

4. Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models* (5th ed.). McGraw-Hill.

5. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning* (2nd ed.). Springer. https://www.statlearning.com/

6. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5th ed.). Wiley.

7. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). CRC Press.

8. Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). Springer.

9. R Core Team (2024). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

10. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). *Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research*, **12**, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html