Introduction

This report is dedicated to assist the Australian federal government in developing a website with a tool of determining the feasibility of installing the modern (installed between 2019-2021) solar panel system. In the report, three predictive models are constructed and rigorously, each varying in complexity and style. The "SolarSurvey" dataset with 3000 observations comprises of 11 variables (excluding generation and household ID), each of which represents different facets potentially impacting power generation prediction. Finally, the report concludes that using a multivariate linear regression model with 6 predictor variables could optimally predict the power generation. This conclusion is based on the test Mean Squared Error (MSE), where our chosen model outperforms the two comparative benchmark models.

Candidate Models

The three models in this report are designed solely on the modern system, which is installed between 2019 and 2021.

1. Model 1 (M1)

$$y = \beta_0 + \beta_1 x_i(Panel\ Capacity) + \beta_2(Roof\ Azimuth) + \epsilon_i$$

Model 1 (M1) is a **bivariate linear regression model**, comprising 2 parameters, which are Panel_Capacity and Roof_Azimuth. Both parameters are the top two variables having the highest correlation with total generation, as shown in *Figure 2-1*.

M1 is a **simple linear regression model**, consisting of two key parameters, Panel_Capacity and Roof_Azimuth (assuming both variables are independent). These parameters are the top two variables with the most significant correlation with total power generation, as illustrated in *Figure 2-1*. Notably, Panel_Capacity has a correlation coefficient of 0.6, suggesting a strong positive relationship with generation; as the panel capacity increases, the total power generated typically rises accordingly. On the other hand, Roof_Azimuth has a correlation of -0.47, indicating a moderate negative relationship. In the context of Australia, situated in the southern hemisphere, an increase in Roof_Azimuth, implying a deviation from the optimal solar exposure direction, usually leads to a reduction in power generation. By incorporating the two dominant predictive factors into the model, this simple linear regression model can easily forecast solar power generation. Achieved through the fitting of a line to the data by minimizing least square error, this linear regression model illustrates the linear correlation between the input independent variables (x) and the outcome dependent variable (y).

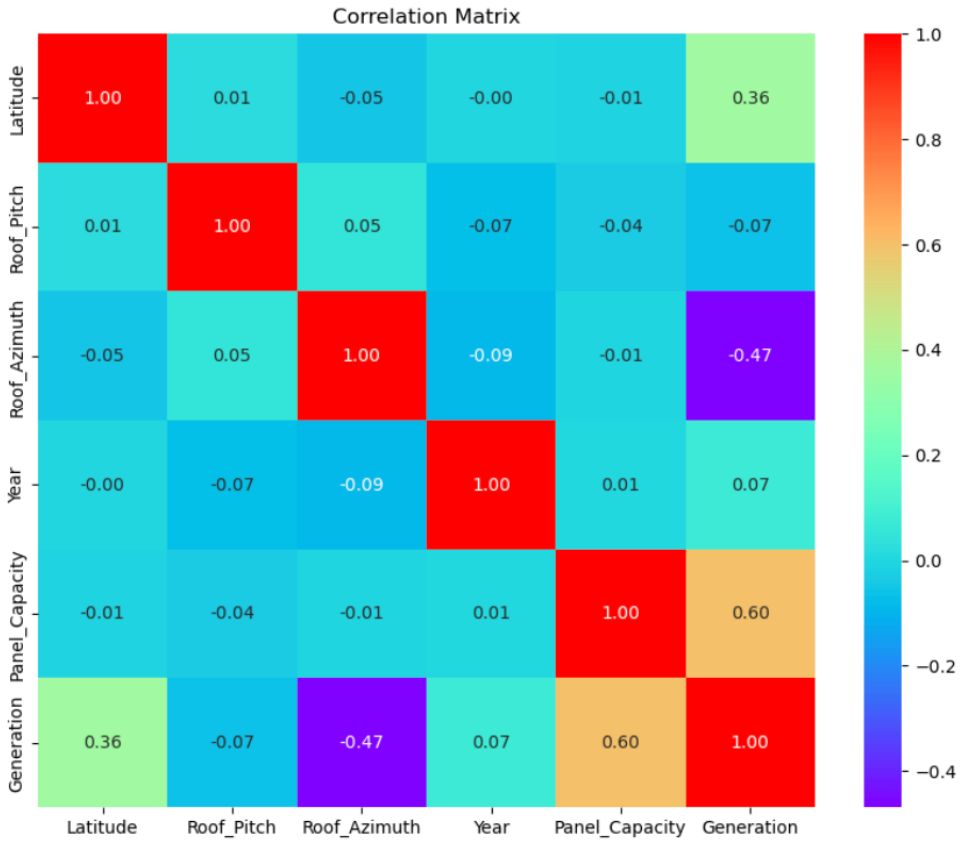

Figure 2-1

2. Model 2 (M2)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \cdots + \beta_p x_2^p + \varepsilon$$

Consider that the relationship between Generation and Panel_Capacity cannot entirely be explained by linear regression line, as concluded from *Figure 2-2* in the first assignment, I utilise **polynomial regression** in Model 2 (M2) to better illustrate the non-linear relationship. Using the built pandas dataframe for degrees 1 to 20 of Panel_Capacity variable, I construct a holdout diagram using the mean squared error (MSE) of both the training set and validation set to closely examine the optimal complexity and degree that corresponds to the minimum validation MSE and the optimum bias-variance trade-off point. After scrutinising *Figure 2-3*, the ideal polynomial degree identified is 4.

Furthermore, as outlined in M1, Panel_Capacity (x_2) and Roof_Azimuth (x_1) stand out as the key variables among the variables offered in the dataset. Therefore, I introduce both the polynomial relationship of Panel_Capacity and the linear relationship of Roof_Azimuth. Under this framework, Roof_Azimuth is considered a polynomial of degree one, essentially implying a linear relationship.

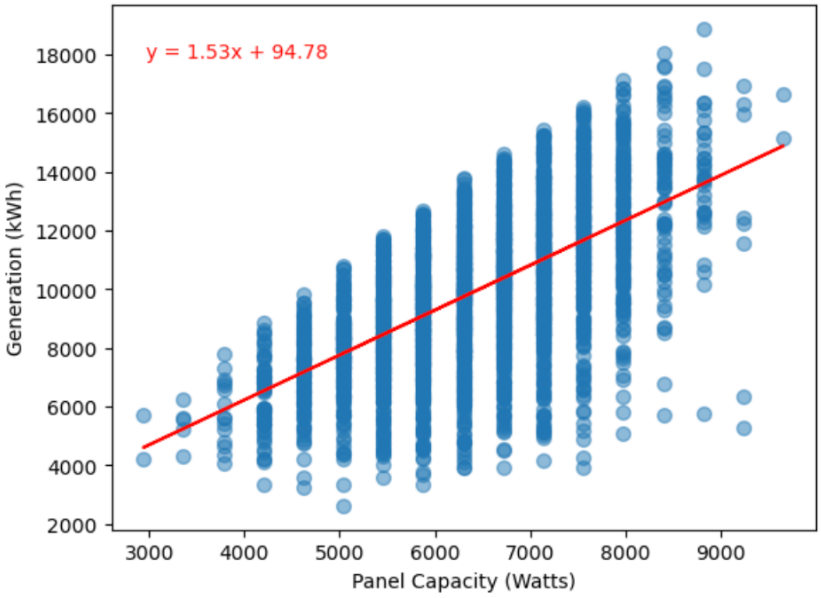

Figure 2-2

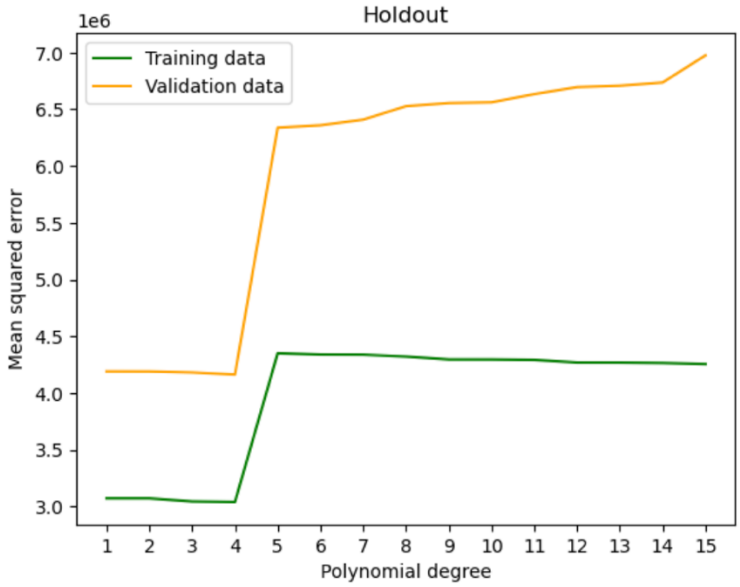

Figure 2-3

3. Model 3 (M3)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_9 (x_5 \times x_6) + \varepsilon$$

Model 3 (M3) is the revised and more advanced version of the previous two models. Given that the polynomial regression performance of M2 exhibits minimal difference from M1, I persist with the initial linear regression model and modify it into a **multivariate linear regression model**. In terms of predictors, I incorporate 7 explanatory parameters into the model. These include Panel_Capacity (x_1), Shading_Partial (x_2), Shading_Significant (x_3), Latitude (x_4), Roof_Azimuth (x_5), Roof_Pitch

(x_6), along with the interaction variables of Roof_Pitch with Roof_Azimuth (x_7). For Shading (three forms: None, Partial, and Significant), 2 dummy variables are applied to help deal with the categorical data type, as outlined below.

$$Partial = \begin{Bmatrix} 1, if\ Shading='Partial' \\ 0, if\ Shading='None' \end{Bmatrix}$$

$$Significant = \begin{cases} 1, if\ Shading='Significant \\ 0, if\ Shading='None' \end{cases}$$

Moreover, considering the interaction effects between the variables as shown in *Figure 2-4*, the interaction term has a moderately negative relationship with solar panel generation, with the correlation of **-0.4811**. This highlights the combined effect of Roof_Pitch and Roof_Azimuth on predicting power generation. Considering the three-dimensional nature of sunlight angles in the real world, the combined variable makes more practical sense since Roof_Azimuth determines the degree to which the panel is facing from North, while the Roof_Pitch influences the tilt angle of the solar panel mounting.

Additionally, as show in *Figure 2-5*, the P-value of coefficient × 7 ("Interaction") is statistically significant (less than the chosen significance level 0.05), proving the interaction between the two variables. Finally, to confirm the independence between of each variable, multicollinearity is checked through the **correlation matrix** as depicted below in *Figure 2-6*. A high correlation observed between the interaction term and roof azimuth suggests potential multicollinearity within the model.

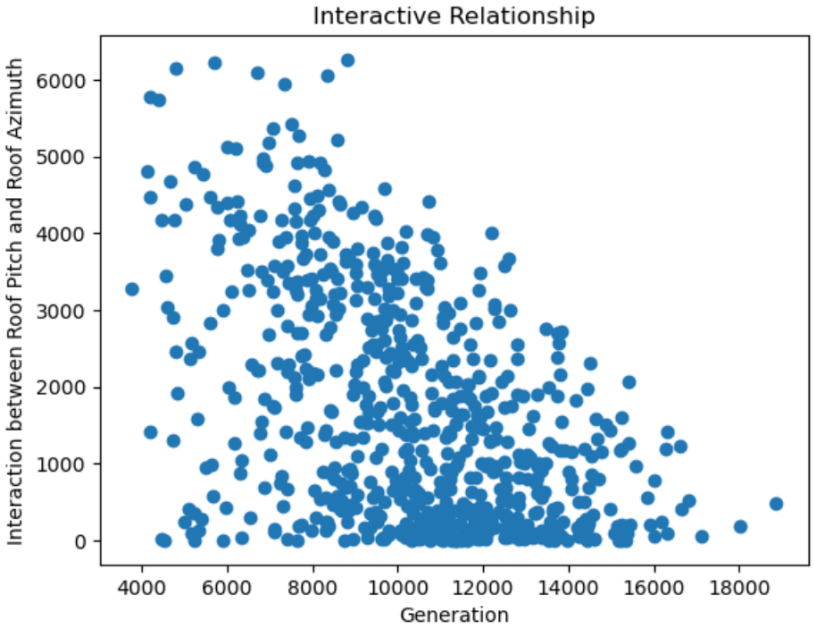

Figure 2-4

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.978
Model:                            OLS   Adj. R-squared:                  0.977
Method:                 Least Squares   F-statistic:                     2090.
Date:                Sat, 20 May 2023   Prob (F-statistic):          1.10e-272
Time:                        22:56:42   Log-Likelihood:                -2555.7
No. Observations:                 344   AIC:                             5127.
Df Residuals:                     336   BIC:                             5158.
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       8455.4991    245.265     34.475      0.000    7973.051    8937.948
x1             1.6183      0.023     70.671      0.000       1.573       1.663
x2         -2098.2317     57.782    -36.313      0.000   -2211.892   -1984.572
x3         -5541.0566    101.375    -54.659      0.000   -5740.466   -5341.647
x4           219.8213      5.213     42.169      0.000     209.567     230.075
x5            -4.3696      1.387     -3.151      0.002      -7.098      -1.642
x6            40.9349      4.906      8.344      0.000      31.285      50.585
x7            -0.6969      0.056    -12.436      0.000      -0.807      -0.587
==============================================================================
Omnibus:                       13.300   Durbin-Watson:                   2.060
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               29.697
Skew:                           0.028   Prob(JB):                     3.56e-07
Kurtosis:                       4.438   Cond. No.                     7.43e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.43e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 2-5

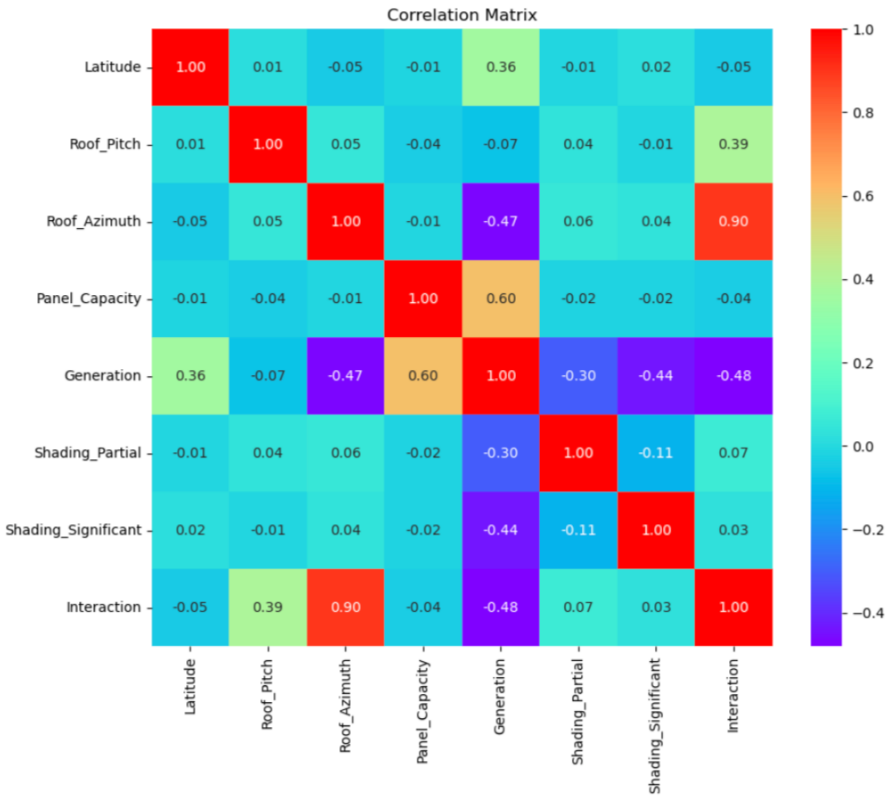

Figure 2-6

The error term representing the difference between the observed and the true value for each observation captures the unexplainable, unobservable randomness that cannot be explained by the model. The three models presented in this report all follows the common assumptions of $\mathbb{E}(\varepsilon) = 0$, $var(\varepsilon) = \sigma^2$, and ε is independent of the variables.

Model Estimation and Selection

Firstly, the three models are estimated based on the training set. **Second**, each model makes prediction for each observation in the validation set (25%, as shown in Table 3-1). **Third**, to estimate the best performing model, I compute the mean squared error (MSE) based on the validation set for each model to quantify the deviations between the observed and the model-predicted data. This validation MSE serves as an approximation of the unobserved expected prediction error (ERE). Before examining the MSE, the estimation results of each model are presented beforehand.

Data Size	Percentage
Training set	50%
Validation set	25%
Test set	25%

Table 3-1

1. Model 1 (M1)

$$y = 389.1268 + 1.7832x_1 - 21.1411x_2$$

M1 is represented by the above equation under the linear regression model, indicating the relationship between a dependent variable and 2 independent variables, x_1 and x_2. The combination of y intercept of 389.1268 and the coefficients of the two variables Panel_Capacity (x_1) and Roof_Azimuth (x_2) suggests that the estimated increase in Panel_Capacity (kWh) is associated with a 1.7832 (Watts) in the expected value of power generation.

Based on the residual scatterplot in *Figure 3-1*, we deduce that the assumption of homoskedasticity in the residual, retrieved by fitting the model to the training set, is held. Moreover, the residuals appear to roughly follow a normal distribution, confirming the model's assumption of normality (*Figure 3-2*). Lastly, since M1 does not include time-series data, we do not expect to observe any significant autocorrelation.

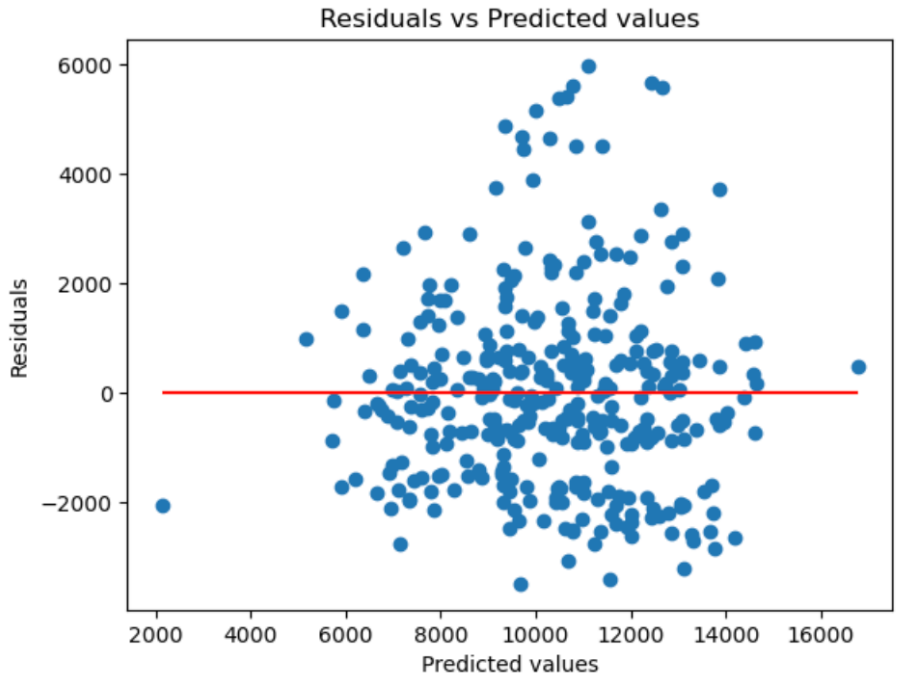

Figure 3-1

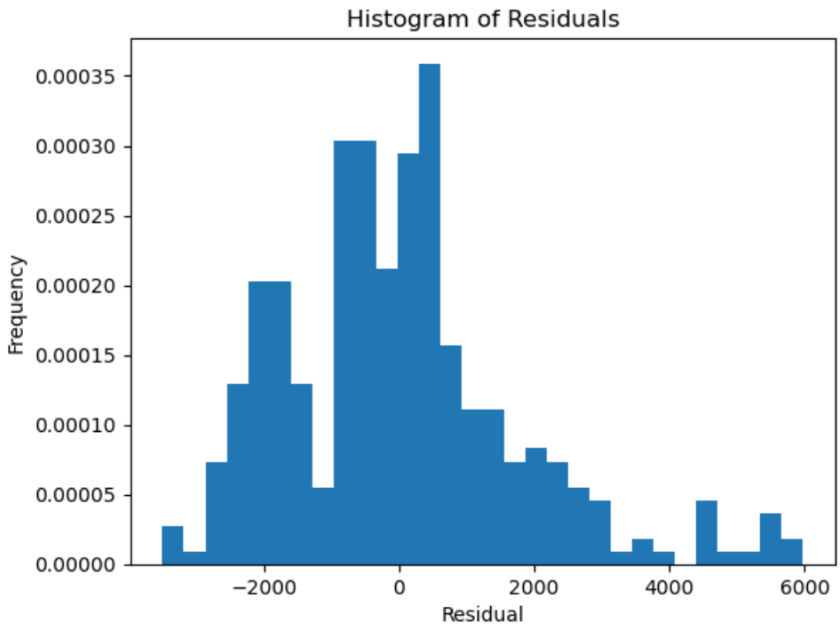

Figure 3-2

2. Model 2 (M2)

$$y = -2976.3929 - 21.1239x_1 + 7.1659x_2 - 0.0021x_2^2 + 3.0107 \times 10^{-7}x_2^3 - 1.4817 \times 10^{-13}x_2^4$$

In M2, the intercept of -2976.3929 serves as the baseline value, when all degrees of the independent variable are zero. The coefficients signify the increasing order of power of the independent variable. For every unit increase in the dependent variable, we expect a 21.1239 unit decrease in roof azimuth; correspondingly a 7.1659 unit increase in the linear term, a 0.0021 unit decrease in the square term, a marginal 3.0107×10^{-7} unit increase in the cubic term, and a negligible 1.4917×10^{-13} unit decrease in the quartic term of the panel capacity variable.

From the residual scatterplot shown in *Figure 3-5*, we can infer the consistency of the variance across the model's predictions, validating the homoskedasticity assumption. Similarly, in *Figure 3-6* the close resemblance of the residuals to a normal distribution can be observed, substantiating the normality assumption of the model. In addition, since M1 does not encompass time-series data, we do not foresee any significant autocorrelation.

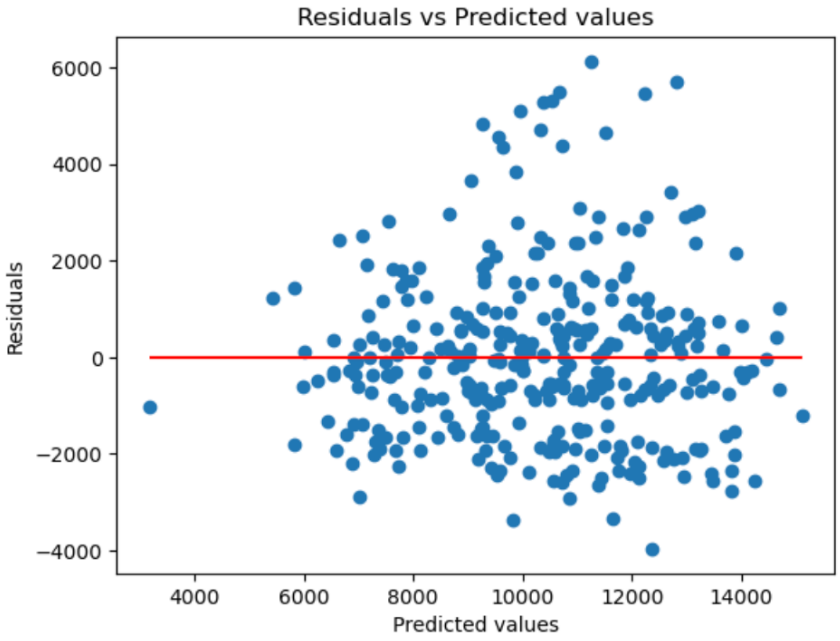

Figure 3-3

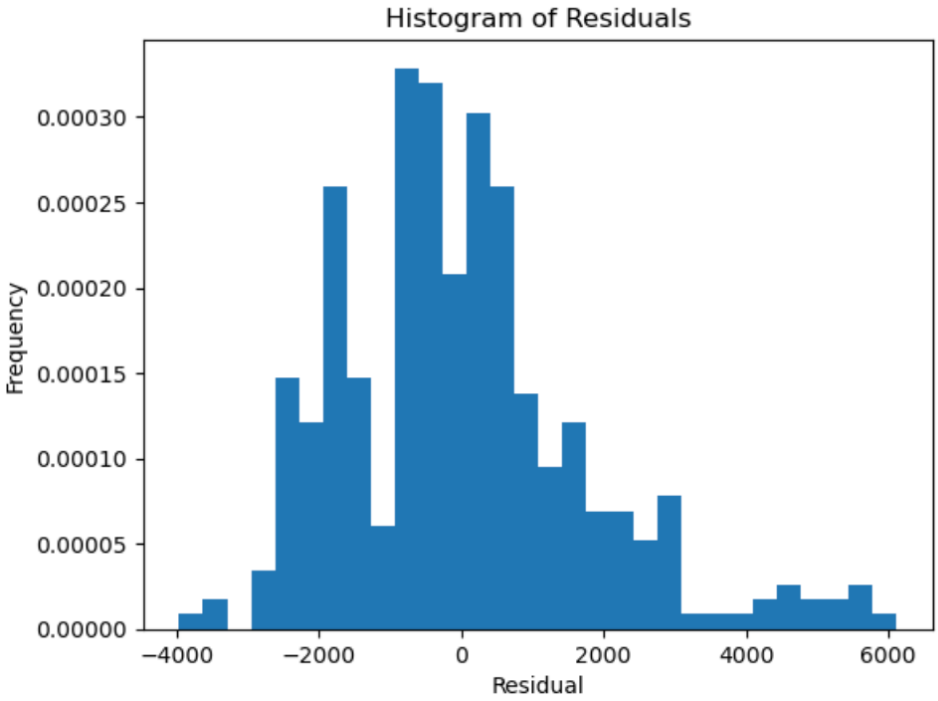

Figure 3-4

3. Model 3 (M3)

$$y = 8455.4991 + 1.6183x_1 - 2098.2317x_2 - 5541.0566x_3 + 219.8213x_4 - 4.3696x_5 + 40.9349x_6 - 0.697(x_5 \times x_6)$$

Model 3 portraits the linear relationship between 5 distinct variables, including one pair of dummy variables and an interaction term, as the formula stated above. From the abovementioned Figure 2-

5, all the coefficients are statistically significant, with P-value less than 0.5. Among all the predictors, a unit increase in shading (partial) and shading (significant) accounts for a significant decrease in energy generation, whereas panel capacity and roof pitch show a positive relationship with the output generation. Furthermore, the interaction terms can be interpreted by an increase in one degree of roof azimuth, the average power generation is increased by $-4.3696 - 0.6969 \times x_6(Roof\ Pitch)$.

Since there is no significant trend in the distribution of residual scatterplot in *Figure 3-5*, the model assumption of homoskedasticity is valid. Furthermore, the residuals approximately follow a normal distribution, supporting the normality assumption (*Figure 3-6)*. Finally, since M1 does not incorporate time-series data, significant autocorrelation is not anticipated to be observed.

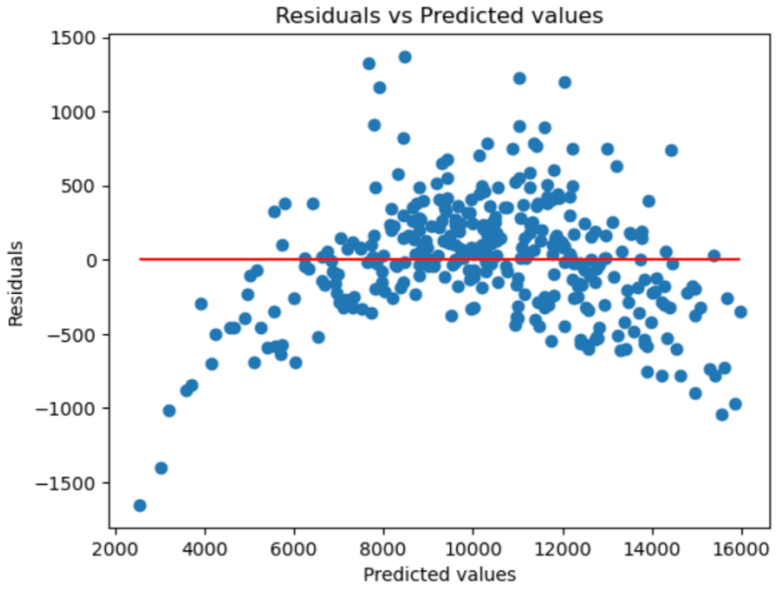

Figure 3-5

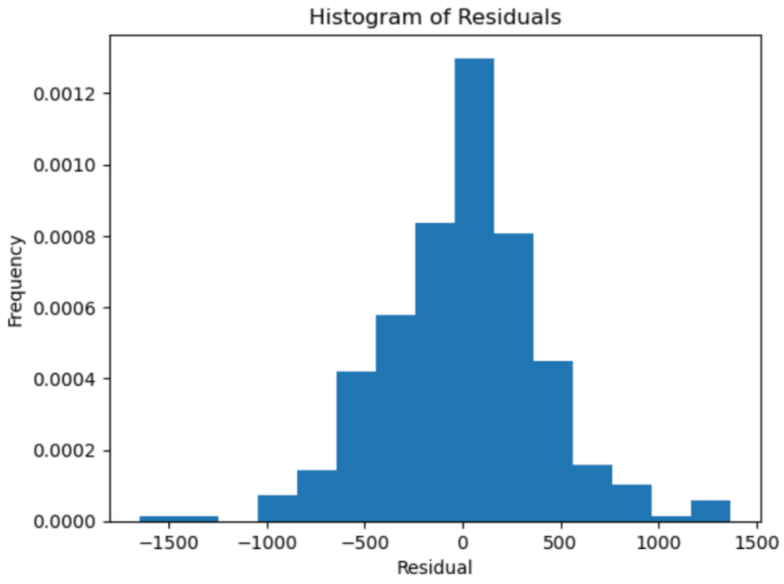

Figure 3-6

The validation MSE of the three models is as follows:

Model Type	Validation MSE Values
M1	4190725.9271
M2	4162618.3525
M3	170474.8097

Table 3-2

As shown in *Table 3-2*, M3 has the lowest the validation MSE, indicating that M3 has the highest accuracy in predicting data, and thus is the most ideal among the three model.

The chosen model M3, a multivariate linear regression model, strikes a balance between the **bias and variance**. Although it is relatively simple as compared to more complicated models, M3 demonstrates great performance in managing variances, since when cross validating the validation and test MSE, the values show little difference. In respect of the bias, M3 has integrated a good number of predictors to address the potential underfitting problem. The model's performance on both the validation and test sets manifests a low Expected Prediction Error (EPE), showcasing its robustness and efficacy in managing bias and variance. Consequently, M3 outstand as a reliable model for forecasting in real-world scenarios.

Model Evaluation

To access the general performance of the M3 against the two benchmark models, I **first** apply the combination set of training and validation sets to the three comparing models. **Second**, the re-estimated selected model, M3, is employed to predict observations in the test set. **Third**, the selected model's performance is evaluated against the two benchmarks using the test MSE.

1. Benchmark Model 1 (BM1)

BM1 is a simple model that predicts the power generation in the most fundamental fashion. It utilises the **arithmetic average** of each city's historical generation data in subset $C(x)$ to estimate the generation data from 2019 to 2021 in city x . The accuracy of BM1 prediction is measured using MSE on a test set. The test MSE is shown in *Table 3-2* below.

$$\hat{y} = \frac{1}{m(x)} \sum_{y \in C(x)} y$$

2. Benchmark Model 2 (BM2)

Compared with BM1, BM2 is more narrowly defined. $C(x_1, x_2)$ includes only the solar generation data from the city X_1, with panel capacity X_2 between the year 2019 to 2021. Similarly, BM2 computes the arithmetic average for each subset as a prediction of for the corresponding observation in the test set. In BM2, regarding to any observations in the test set without a matching

group in the training and validation data, the overall average of the generation is used to fill the missing values (NAN).

In conclusion, test MSE is calculated to evaluate the accuracy of the benchmark. The result in *Table 3-2* below clearly demonstrates that, accuracy compared to the benchmark models, M3 exhibits superior accuracy across test data.

Model Type	Test MSE Values
Benchmark 1	5219490.029
Benchmark 2	3162064.6951
Model 3 (M3)	176382.6604

Table 3-2

Conclusion

Among the three models constructed, M3 performs the best in terms of accuracy and the balance of bias-variance trade-off. This can be seen by comparing the validation and test MSE to the rest of the models as well as two benchmark models. However, M3 has some disadvantages and limitations when dealing with multicollinearity. This issue could cause the model to overfit the training set and compromise the interpretability and significance of the predictors.

Furthermore, since M3 depends mainly on linear relationship, non-linear relationship could be discovered using machine learning algorithm in future work. Lastly, expanding the dataset could also significantly improve model performance and predictive power. Methods include obtaining more recent data or incorporating more variables. These adjustments could enhance our model, fostering increased accuracy and robustness.