Optimizing Public Bike-Sharing Systems:

The Role of Temporal and Weather Factors in Demand Forecasting

（Statistical Linear Regression Analysis)

Maojia Wang, Sophia Cao

Department of Statistics, University of Toronto

STA302

Dr. Austin Brown

December 6, 2024

Contribution:

Maojia Wang: Responsible for writing Introduction , Method, and limitation

Sophia Cao: Responsible for writing Result, conclusion and coding in R studio

1. Introduction (236 words)

Public bike-sharing systems offer eco-friendly, convenient urban transportation, addressing environmental and traffic challenges. This study investigates how temporal (holidays, weekends, time of day) and weather factors (rainfall, temperature, wind speed) affect bike-sharing demand. Using data from Seoul's bike-sharing program, the analysis aim to provide insights for optimizing bike-sharing operations by improving resource allocation, ensuring better bike availability, and enhancing station capacity to meet demand efficiently.

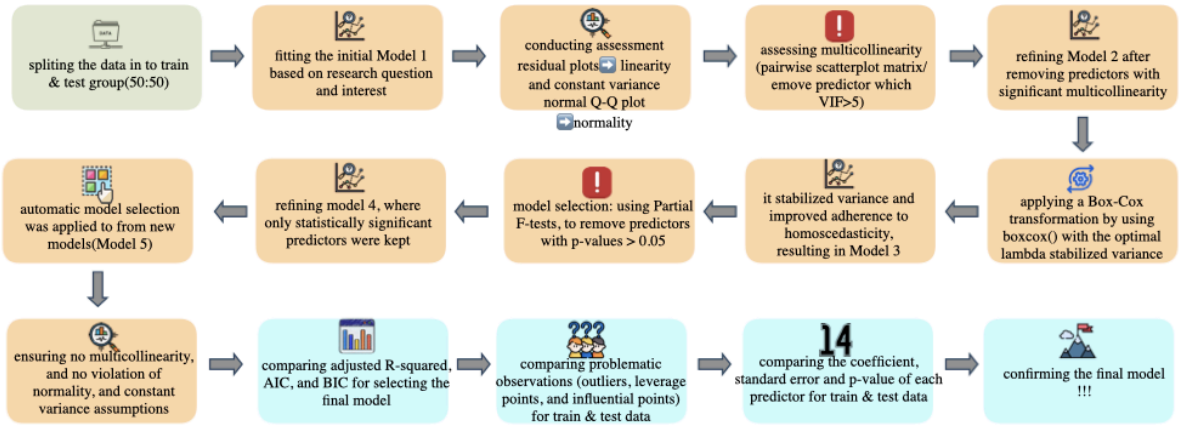
Previous studies have analyzed factors affecting bike-sharing demand that provided us insights on making better predictions on shared-bike demand. VE and Cho (2020) highlighted the importance of time-of-day variations as a primary determinant. Gao and Chen (2022) demonstrated that temperature and rainfall significantly affect demand patterns. Similarly, El-Assi, Salah Mahmoud, and Nurul Habib (2017) examined Toronto's bike-sharing system, underscoring the impact of temperature and precipitation on usage levels. While these studies confirm the relevance of temporal and weather factors, the relative importance of these predictors remains uncertain. This study addresses this gap by quantifying their influence through a linear regression framework.

Linear regression is an effective tool for analyzing the relationship between bike rental demand and its predictors. It allows for the estimation of each variable's contribution to demand, providing interpretable coefficients that quantify the influence of weather and

time-related factors. This clarity makes linear regression particularly valuable for policymakers. Assuming a primarily linear relationship between predictors and demand, this study leverages the strengths of linear regression to provide actionable insights.

2. Method (515 words)

2.1 A simplified flowchart outlines our method:



2.2 Data Paritioning

A 50:50 split is used when subdividing our dataset into training and testing datasets for model development purposes. The training dataset is used for fitting model, hypothesis testing, and model selection. While the testing dataset will serve as an independent set to assess predictivity, validity and overall interpretation ability of model results - this unbiased method ensures fair evaluation.

2.3 Model construction

Model 1: We began our analysis by identifying the variables of interest based on our research question and fitting our initial Model 1. This model included a combination of categorical and numerical predictors hypothesized to influence the bikes demand. After fitting the model, we assessed linear regression assumptions to ensure validity.

Model 2: We evaluated the linearity assumption using a residual vs. fitted plot to see if the residuals were randomly scattered without patterns. Next, we examined the homoscedasticity assumption by observing the same residual plot for constant variance in the residuals across fitted values. To check for the normality of residuals, we plotted a normal Q-Q plot and verified if the residuals aligned closely with the 45-degree reference line. Finally, we assessed multicollinearity among predictors by creating a pairwise scatterplot and calculating VIF. A VIF > 5 indicates significant multicollinearity, it should be removed to refine Model 2.

Model 3: Next, we revisited the assumptions with Model 2. While multicollinearity was resolved, the residual vs. fitted plot suggested non-constant variance. Applying a Box-Cox transformation by using boxcox() function with the optimal lambda stabilized variance, and the response variable was transformed accordingly. This adjustment stabilized variance and improved adherence to homoscedasticity, resulting in Model 3.

Model 4: Building on Model 3, we refined the model further by conducting subset selection based on statistical significance. Using Partial F-tests, to remove predictors with p-values > 0.05 This process produced a simplified and effective Model 4, where only statistically significant predictors were kept.

Model 5: To explore additional optimizations, we apply automatic model selection, a stepwise selection method is applied to evaluate predictors iteratively, based on AIC. This leads to Model 5, which demonstrates improved performance while maintaining simplicity.

2.4 Model Diagnostics and Validation

We check the linear regression assumptions for our final model methods from model construction. We observe each of the residual plots to see if residual plots show no pattern. Normality is checked using Q-Q plot as lack of normality will affect accuracy of p-value and confidence interval. We also check VIF again to ensure no multicollinearity among predictors.

Problematic observations were evaluated using Cook's Distance, DFFITS, and DFBETAS. Cook's Distance highlights overall impact on model stability, DFFITS assesses influence on fitted values, and DFBETAS evaluates slope influence. Observations exceeding cutoffs were reviewed individually to ensure robustness.

For model validation, our dataset was evenly divided (50:50) divided into train data and test data. Models were built using train data, and final models were selected based on adjusted R-squared, AIC, and BIC. The linear regression assumptions for the final model was assessed, then tested, with coefficients and problematic observations from train and test data expected to be similar and statistically significant.

3. Results (265 words)

3.1 Description of Data

After cleaning the dataset by removing NA values, we ended up with 8760 observations, which we then split into 2190 training and 2190 testing observations (50:50) after a 50% random selection of 4380 observations. This is a reasonable balance between computational efficiency and representativeness, preserving key statistical properties.

Variable	Description
Rented_Bike_Count	Number of bikes rented in a given hour

Hour	Time of the day categorized into Late Night (00:00-06:00), Morning (06:00-12:00), Afternoon (12:00-18:00), and Evening (18:00-24:00)
Temperature	Average temperature during the hour (°C)
Humidity	Average relative humidity during the hour (%)
Wind_speed	Average wind speed during the hour (m/s)
Visibility	Average visibility during the hour (meters)
Dew_point_temperature	Temperature at which air becomes saturated with moisture during the hour(°C)
Solar_Radiation	Amount of solar radiation during the hour (MJ/m^2)
Rainfall	Amount of rainfall during the hour (mm)
Snowfall	Amount of snowfall during the hour (cm)
Seasons	Season during the hour: Spring, Summer, Fall, or Winter
Holiday	Whether the day is a holiday or not (Yes/No)
Functioning_Day	Whether the bike rental system was operational on that day (Yes/No)

Table 1: Description of Variables

Table 1 provides an overview of our variables. Table 2 summarizes the numeric variables of our interest showing the distribution of training data (upper) and testing data (bottom). Most of the variables in both distributions are roughly similar, therefore random split keeps the full data well-represented, minimizes risks of overfitting and provides better model predictions. Figure 1 provides a better visualization of our statistics.

	Rented_Bike_Count	Hour	Temperature	Humidity	Wind_speed	Dew_point_temperatur	Solar_Radiation	Rainfall	Snowfall
Min	0.000	0.000	-17.500	0.000	0.000	-30.500	0.000	0.000	0.000
1st Qu.	187.000	5.000	3.700	42.000	1.000	-4.200	0.000	0.000	0.000
Median	482.000	11.000	13.900	57.500	1.500	5.400	0.020	0.000	0.000
Mean	699.101	11.427	12.873	58.211	1.733	4.095	0.570	0.147	0.092
3rd Qu.	1067.500	17.000	22.500	74.000	2.300	14.500	0.958	0.000	0.000
Max	3556.000	23.000	39.400	98.000	7.400	26.800	3.520	18.500	8.800

	Rented_Bike_Count	Hour	Temperature	Humidity	Wind_speed	Dew_point_temperatur	Solar_Radiation	Rainfall	Snowfall
Min	0.000	0.000	-17.400	0.00	0.000	-30.60	0.000	0.000	0.000
1st Qu.	182.000	5.000	3.500	43.00	0.900	-4.70	0.000	0.000	0.000
Median	484.500	11.000	13.900	58.00	1.500	5.30	0.010	0.000	0.000
Mean	693.466	11.398	13.052	58.48	1.713	4.25	0.593	0.137	0.057
3rd Qu.	1052.000	17.000	23.000	75.00	2.400	15.10	0.990	0.000	0.000
Max	3245.000	23.000	38.000	98.00	6.000	26.80	3.520	29.500	5.100

Table 2: Numerical Summary of Training Data Variables and Testing Data Variables

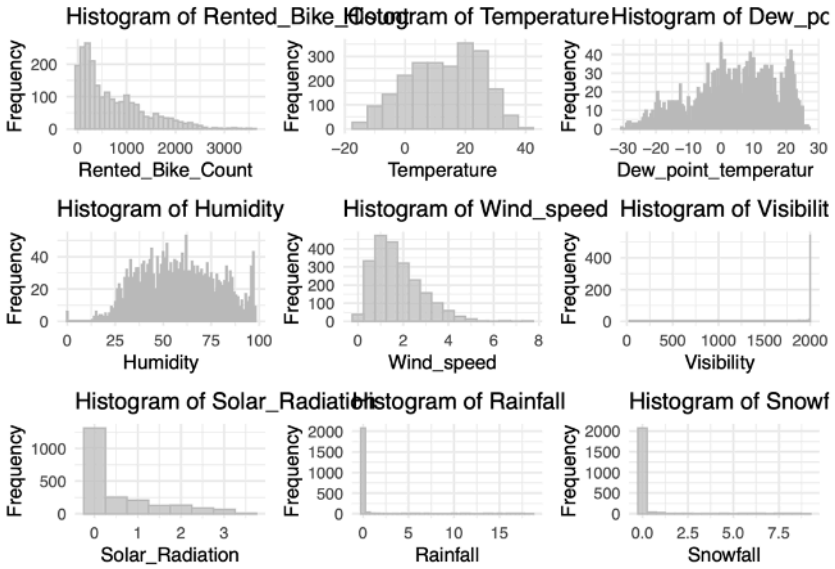


Figure 1: Histograms of Numerical Variable of Training Data

3.2 Model Construction

Initially, **Model 1** is developed by fitting a multiple linear regression using all predictors of interest on number of rented bikes to address our research question:

$$\begin{aligned}\hat{Y} = {} & 872.06718 + 35.49594 \cdot \text{Temperature} - 4.62688 \cdot \text{Dew\_point\_temperatur} \\ & - 5.48636 \cdot \text{Humidity} + 20.30833 \cdot \text{Wind\_speed} + 0.02639 \cdot \text{Visibility} \\ & - 36.27540 \cdot \text{Solar\_Radiation} - 73.53272 \cdot \text{Rainfall} - 621.57474 \\ & \cdot \text{LateNight} - 311.26899 \cdot \text{Morning} - 349.59978 \cdot \text{Afternoon} \\ & - 161.14843 \cdot \text{Holiday}\end{aligned}$$

Referring to Appendix Figure 2, the pairwise scatterplot showed a very strong linear relationship between predictors: Temperature and Dew_point_temperatur. Also, by computing the VIF (Table 5), we found that This suggests that including both predictors in the model will lead to redundancy and bias in the regression coefficients. To address this issue, it is reasonable to remove the predictor Dew_point_temperatur, so that the model will reduce multicollinearity.

After removing the predictor Dew_point_temperatur, we reach our **Model 2**:

$$\begin{aligned}\hat{Y} = 973.60479 &+ 31.10513 \cdot \text{Temperature} - 6.59769 \cdot \text{Humidity} + 20.45815 \\ &\cdot \text{Wind\_speed} + 0.02593 \cdot \text{Visibility} - 33.80450 \cdot \text{Solar\_Radiation} \\ &- 73.07421 \cdot \text{Rainfall} - 622.30754 \cdot \text{LateNight} - 312.91803 \cdot \text{Morning} \\ &- 348.49487 \cdot \text{Afternoon} - 161.89391 \cdot \text{Holiday}\end{aligned}$$

Referring to Figure 2, Model 1 was evaluated using the plots of Residual vs. Fitted to detect deviation from constant variance assumptions. As evidenced by residuals not remaining constant across fitted values, homoscedasticity is violated. Additionally, the Normal Q-Q Plot reveals that the residuals deviate significantly from the diagonal line, indicating a lack of normality in their distribution. These findings suggest that transformations are needed to address these issues.

Box-Cox transformations are applied to address the lack of normality and constant variance. Based on maximum likelihood estimates, an optimal $\lambda = 0$ is chosen for Rented Bike Count, and $\lambda = -2$ for Rainfall (Figure 3). A constant of positive 1 is added to ensure all values were positive. Post-transformation, **Model 3** is developed:

$$\log(\hat{Y}) = 2.386 \cdot 10^{1} + 4.316 \cdot 10^{-1} \cdot \text{Temperature} - 6.484 \cdot 10^{-2} \cdot \text{Humidity} + 2.827 \cdot 10^{-2} \cdot \text{Wind\_speed} + 3.721 \cdot 10^{-4} \cdot \text{Visibility} - 1.516 \cdot 10^{-1} \cdot \text{Solar\_Radiation} - 2.889 \cdot \text{Rainfall}^{-2} - 8.546 \cdot 10^{0} \cdot \text{LateNight} - 3.230 \cdot 10^{0} \cdot \text{Morning} - 3.392 \cdot 10^{0} \cdot \text{Afternoon} - 3.601 \cdot 10^{0} \cdot \text{Holiday}$$

After implementing transformation, there is an improvement in the violation of the constant variance assumption. However, it is not significant enough to meet homoscedasticity. Residual plots suggested the lower bound of the Rented Bike Count still restrict variability at low demand levels, which cause the decreasing linear pattern in the plot of Residual vs. Fitted ranging from 10 to 30 (Figure 2). This dataset limitation is realized and discussed further later.

To check if we can simplify our final model based on Model 3, we removed all the insignificant predictors (p-value > 0.05) from Model 3 to make Model 4:

$$\log(\hat{Y}) = 24.76509 + 0.43070 \cdot \text{Temperature} - 0.06976 \cdot \text{Humidity} - 28.94982 \cdot \text{Rainfall}^{-2} - 8.56234 \cdot \text{LateNight} - 3.37717 \cdot \text{Morning} - 3.65439 \cdot \text{Afternoon} - 3.58649 \cdot \text{Holiday}$$

A partial-F test is conducted for Model 3 and **Model 4**. We set the null hypothesis(H0): All the coefficients of Wind_speed, Visibility, Solar_Radiation are zero. Alternative hypothesis(Ha): At least one of the coefficients of Wind_speed, Visibility, Solar_Radiation does not equal zero. The result of our partial-F test yields p-value of 0.5535 > 0.05, we fail to reject H0 and conclude that additional variables in the full model do not significantly enhance the predictability of our model compared to the reduced model, so they are removed.

Building on our Model 4, stepwise Automated Selection Methods are used here to obtain our **Model 5**:

$$\log(\hat{Y}) = 24.76509 + 0.43070 \cdot \text{Temperature} - 0.06976 \cdot \text{Humidity} - 28.94982 \cdot \text{Rainfall}^{-2} - 8.56234 \cdot \text{LateNight} - 3.37717 \cdot \text{Morning} - 3.65439 \cdot \text{Afternoon} - 3.58649 \cdot \text{Holiday}$$

Table 3 compares Adjusted R-Square, AIC, and BIC for the five models. Model 4 and 5 are identical, and they have the fewest predictors, with the lowest AIC and BIC. Although Models 1 and 2 have a slightly bigger adjusted R-Square compared to Models 4/5, considering meeting linear regression assumption and statistical significance, they are not chosen. The slight reduction in the Adjusted R-Square is outweighed by significant improvements in AIC and BIC, this leads to our conclusion of the final model: Model 5.

Model	N.of.Predictors	Adjust.R.Square	AIC	BIC
Model_1	11	0.5121	32992.67	33066.66
Model_2	10	0.5122	32991.02	33059.32
Model_3	10	0.5076	14626.29	14694.59
Model_4	7	0.5078	14622.40	14673.62
Model_5	7	0.5078	14622.40	14673.62

Table 3: Model Comparison

3.3 Model Diagnostics and Validation

After conducting diagnostics of Model 5 on both the training and testing datasets, the following observations were noted (Table 4 and Table 5 in the Appendix):

1. All assumptions of linear regression are checked again for our final model.

2. The coefficient estimates for Model 5 are consistent across both the training and testing datasets (indicating stability in model performance).

3. The VIF for each predictor is below 5 in both datasets (there is no substantial multicollinearity concern).

4. The number of statistically significant predictors is the same in both the training & testing models (indicates consistency in the variables that impact the model outcomes).

However, In the training dataset, 52% of observations were identified as problematic, compared to 54.7% in the testing dataset (testing dataset might have slightly more outliers).

Metric	Model5_train	Model5_test	Model5_train_ratio	Model5_test_ratio
Leverage	181	207	0.0826484	0.0945205
Small_Outliers	101	104	0.0461187	0.0474886
Large_Outliers	24	17	0.0109589	0.0077626
Cooks	0	0	0.0000000	0.0000000
DFFITS	123	122	0.0561644	0.0557078
Total_DFBETAS	711	749	0.3246575	0.3420091
Total_Problematic_Observations	1140	1199	0.5205479	0.5474886

Term	Estimate_Train	Std_Error_Train	Pr_Train	Estimate_Test	Std_Error_Test	Pr_Test
Intercept	24.76509	0.57993	< 2e-16	26.392344	0.564275	< 2e-16
Temperature	0.43070	0.01291	< 2e-16	0.440388	0.012534	< 2e-16
Humidity	-0.06976	0.00896	1.06e-14	-0.095323	0.008653	< 2e-16
Rainfall_transformed	-28.94982	1.84066	< 2e-16	-25.901090	1.708935	< 2e-16
LateNight	-8.56234	0.42725	< 2e-16	-8.396633	0.413824	< 2e-16
Morning	-3.37717	0.42101	1.69e-15	-4.152670	0.408383	< 2e-16
Afternoon	-3.65439	0.43192	< 2e-16	-4.412398	0.425073	< 2e-16
Holiday_Holiday	-3.58649	0.69805	3.03e-07	-3.964966	0.622652	2.33e-10

Table 4: Summary of Problematic Observations & Coefficient Comparison of Model 5

4. Conclusion and Limitation (355 words)

Final Model (Model5):

$$\log(\hat{Y}) = 24.76509 + 0.43070 \cdot \text{Temperature} - 0.06976 \cdot \text{Humidity} - 28.94982 \cdot \text{Rainfall}^{-2} - 8.56234 \cdot \text{LateNight} - 3.37717 \cdot \text{Morning} - 3.65439 \cdot \text{Afternoon} - 3.58649 \cdot \text{Holiday}$$

Conclusion: Our research identifies key factors affecting bike-sharing demand, providing valuable insights into usage patterns. The model shows that, holding all other factors constant, an increase in temperature by one unit increases bike rentals by 0.43% , reflecting higher user activity in moderate weather. Conversely, for a one unit increase in Rainfall^-2, bike demand will decrease by 28.94%, capturing the significant deterrent effect of

precipitation. These findings align with Gao and Chen (2022), who also identified temperature and rainfall as critical factors.

Time of day also plays a significant role. Compared to the reference period, demand decreases by 8.56% during LateNight, 3.37% in the Morning, and 3.65% in the Afternoon, highlighting evening and peak times as periods of higher activity. Similarly, holidays reduce demand by 3.59%, indicating fewer trips on non-working days. Humidity further decreases demand by 0.07% for each unit increase, reflecting its adverse effect on riding comfort.

These results answer the research question by clearly quantifying the effects of weather and temporal factors on bike demand. They provide a foundation for policymakers to optimize bike distribution during favorable weather and peak times, ensuring efficient resource allocation and promoting sustainable urban mobility.

Limitation: One notable limitation in our research is that we do not meet all linear assumptions. The residual vs. fitted plot shows heteroscedasticity, the Q-Q plot reveals deviations from normality. These can lead to unreliable significance tests, incorrect standard errors, and potentially misleading conclusions. These patterns are realized but cannot be addressed after we implement all possible techniques. This is because while linear regression provides good interpretability, it assumes a linear relationship between predictors and the response variable, however, in later analysis, demand of rented bikes yields diminishing returns with increasing temperature, the number of bikes rented will rise first and drop afterwards. Also, the fact that the response variable cannot be negative creates a hard boundary for residuals when the fitted values are small, this could cause lower value prediction to be imprecise. Lastly, influential points and outliers were not removed, though retaining them may introduce biases, this also enhances external validity and the applicability of the findings to real-world situations.

Appendix

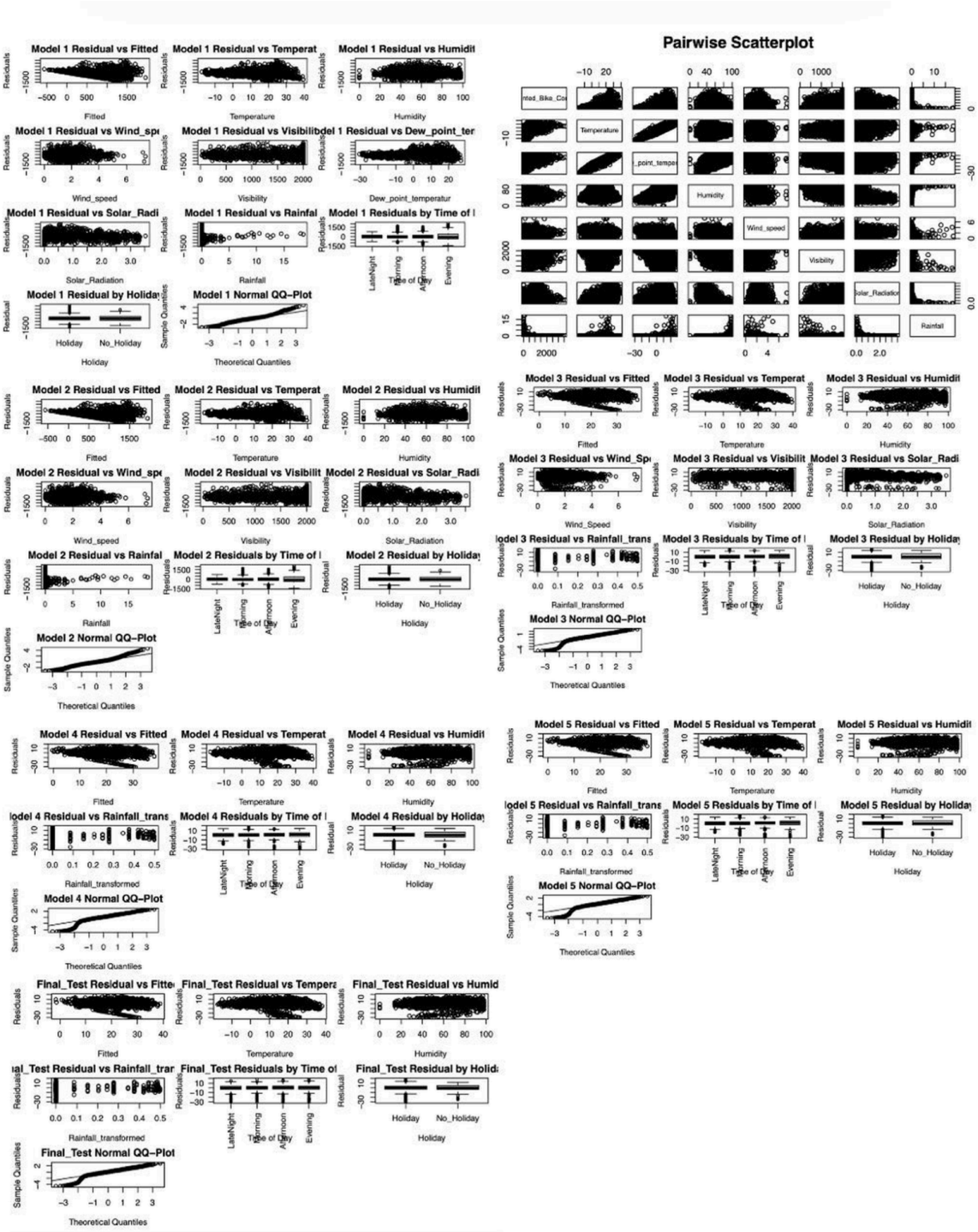


Figure 2: Residual Plos & Normal QQ Plots for Each Model & Pairwise Scatterplot

Variable	VIF
Temperature	86.096
Dew_point_temperatur	111.107
Humidity	18.228
Wind_speed	1.319
Visibility	1.569
Solar_Radiation	3.222
Rainfall	1.099
LateNight	1.748
Morning	2.051
Afternoon	2.695
Holiday_Holiday	1.007

Variable	VIF
Temperature	1.609
Humidity	2.524
Wind_speed	1.319
Visibility	1.566
Solar_Radiation	3.084
Rainfall	1.092
LateNight	1.745
Morning	2.036
Afternoon	2.688
Holiday_Holiday	1.006

Variable	Model5_Training	Model5_Testing
Afternoon	1.7001	1.6812
Holiday_Holiday	1.0051	1.0097
Humidity	1.5663	1.3587
LateNight	1.6328	1.6202
Morning	1.5702	1.5483
Temperature	1.1200	1.1122

Table 5: VIF summary (Model1, Model2, Model5_Training & Model5_Testing)

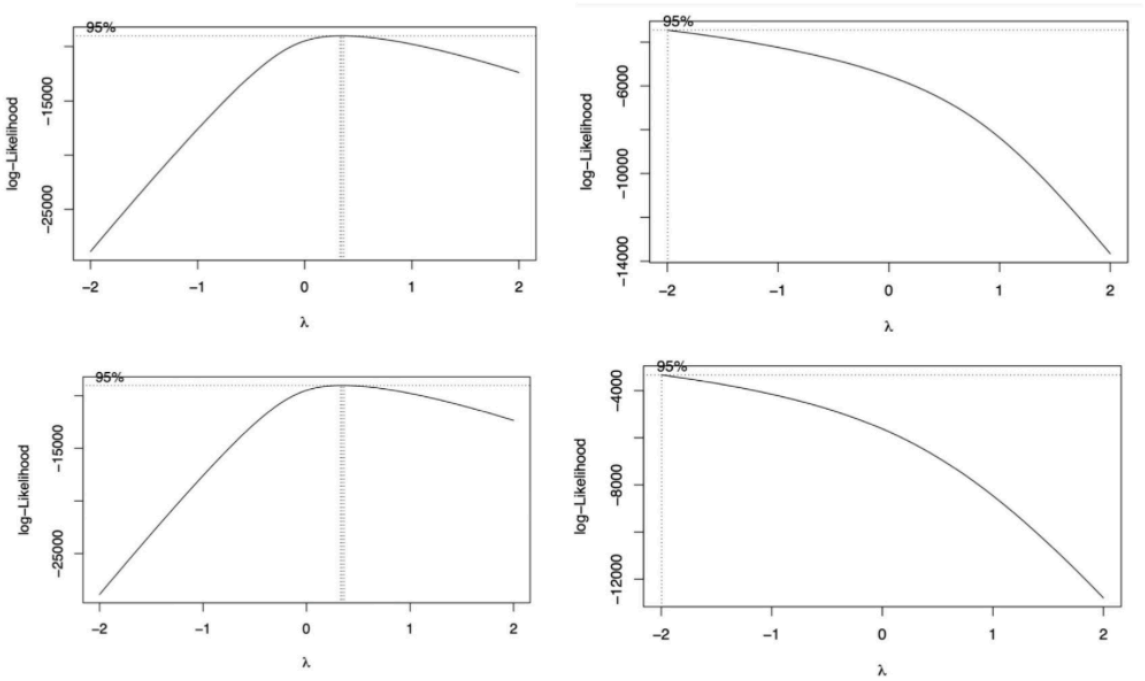


Figure 3: Graph for Box-Cox Transformation for Rented_Bike_Count (left) and Rainfall (right) both in training data and testing data

Reference

El-Assi, W., Salah Mahmoud, M., & Nurul Habib, K. (2017). Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in Toronto. *Transportation (Dordrecht)*, *44*(3), 589–613. https://doi.org/10.1007/s11116-015-9669-z

Shahane, S. (n.d.). Seoul Bike Sharing Demand Prediction [Data set]. Kaggle. Retrieved November 28, 2024, from https://www.kaggle.com/datasets/saurabhshahane/seoul-bike-sharing-demand-prediction

Stienmetz, J. L., Massimo, D., & Ferrer-Rosell, B. (2022). Using Machine Learning Methods to Predict Demand for Bike Sharing. In *Information and Communication Technologies in Tourism 2022*. Springer International Publishing AG.

V E, S., & Cho, Y. (2020). A rule-based model for Seoul Bike sharing demand prediction using weather data. *European Journal of Remote Sensing*, *53*(1), 166–183. https://doi.org/10.1080/22797254.2020.1725789