Final project for DS510

Gwangwoo Kim, 20235472

June 2024

1 Introduction

Currently, rental bikes are being introduced in many urban cities to enhance mobility and comfort. Ensuring that rental bikes are available and accessible to the public at the right time is crucial, as it reduces waiting times. Consequently, maintaining a stable supply of rental bikes throughout the city has become a major concern. The key challenge is accurately predicting the number of bikes required each hour to ensure a stable supply.

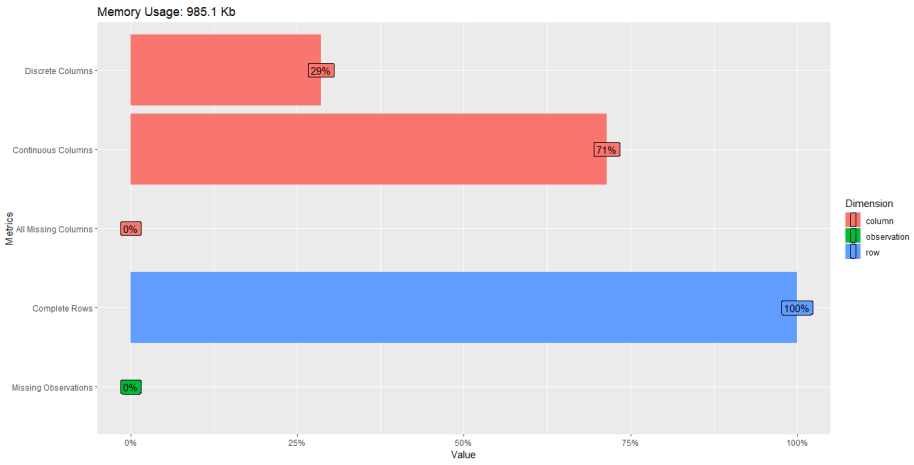
2 Regression analysis

2.1 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a common methodology for visually analyzing data using various parameters, allowing for effective data summarization. EDA was performed using the DataExplorer package in R.

```
# A tibble: 1 × 9
   rows columns discrete_columns continuous_columns all_missing_columns total_missing_values complete_rows total_observations memory_usage
  <int>   <int>            <int>              <int>               <int>                <int>         <int>              <int>        <dbl>
1  8760      14                4                 10                   0                    0          8760             122640      1008752
```

(a) A summary of data



(b) Visualization

Figure 1: These are basic data summaries. The number of instances is 8760, there are 14 features, four of them are discrete variables, and the rest are continuous variables. Fortunately, there are no missing values in all features!

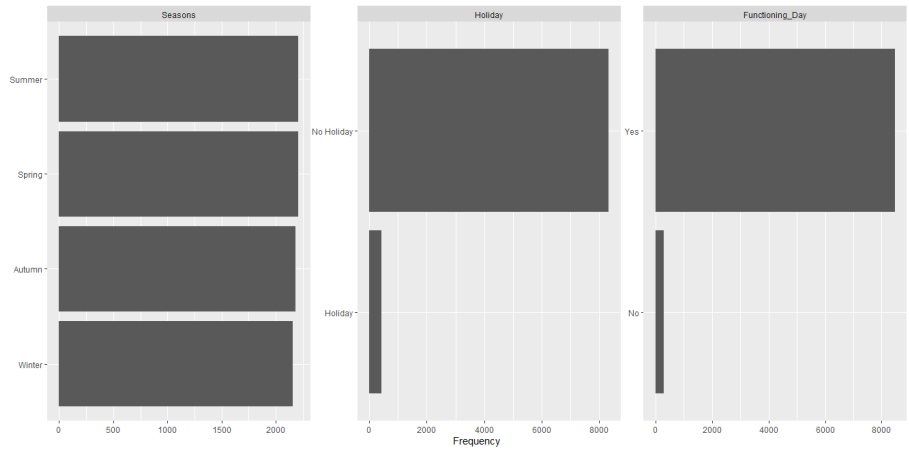
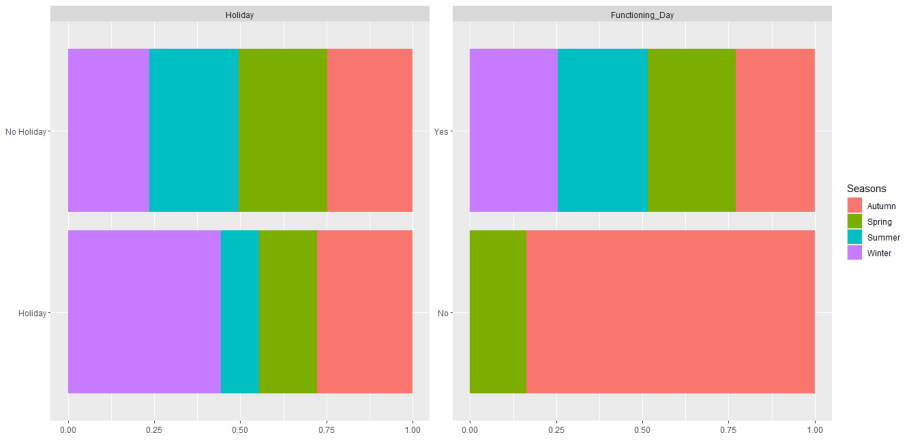
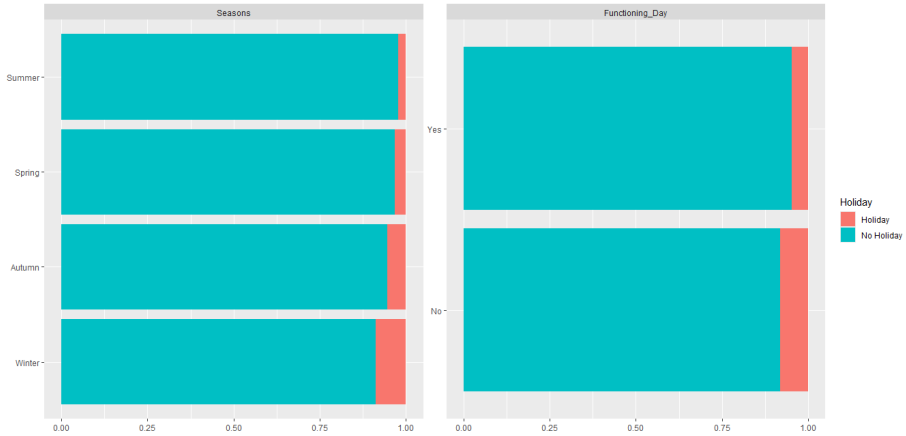


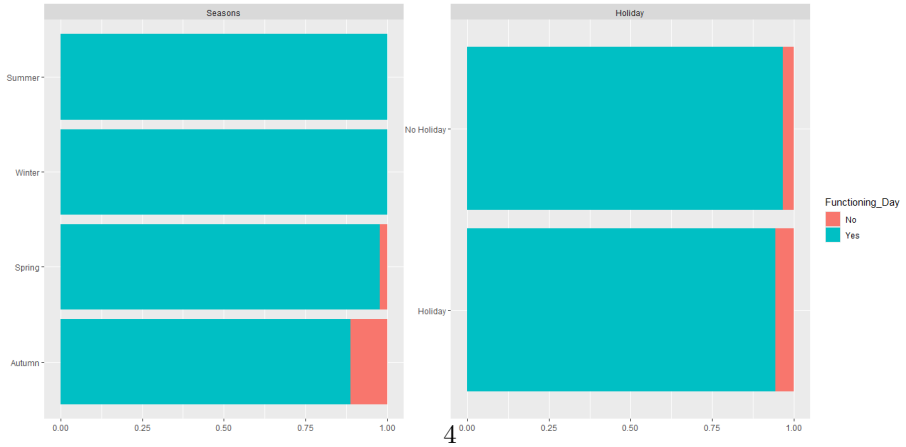
Figure 2: Histograms for discrete variables. The Seasons variable is balanced, while the other variables (Holiday and Functioning_Day) are not. Such imbalanced classes might not be reflected in (simple) models and could be considered outliers.



(a) Conditioned on Seasons



(b) Conditioned on Holiday



(c) Conditioned on Functioning_Day

Figure 3: Histograms conditioned on each discrete variable. Note that the Summer and Winter classes of the Seasons variable do not have instances labeled by Functioning_Day.

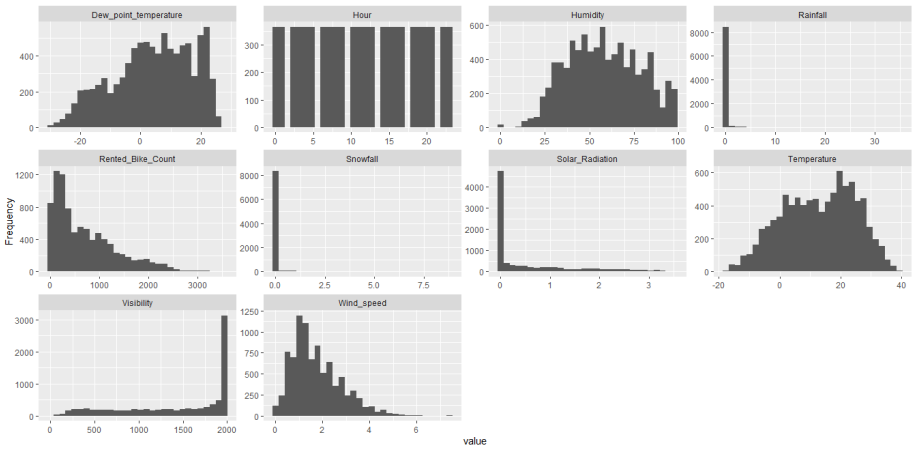


Figure 4: Histograms for continuous variables. In particular, the Rented_Bike_Count variable is our target variable. If we use a model assuming the response variable's normality, the target variable's normality should be tested.

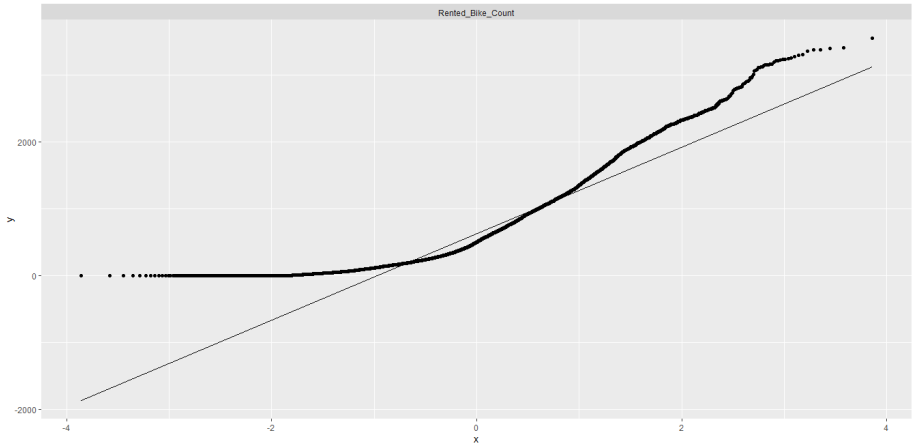
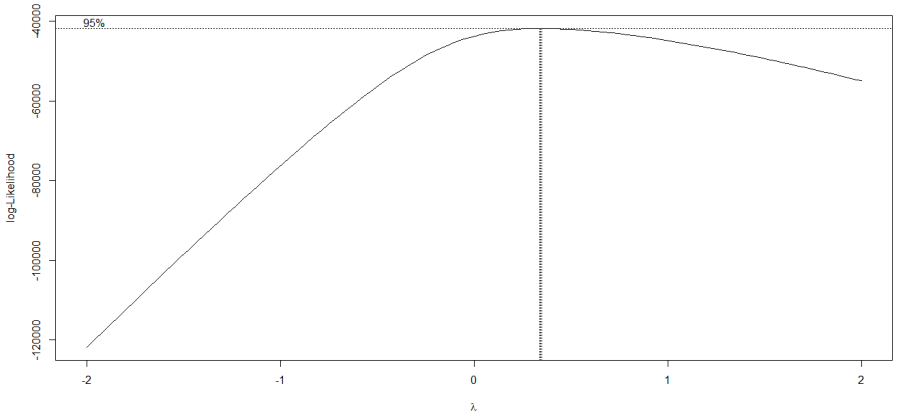
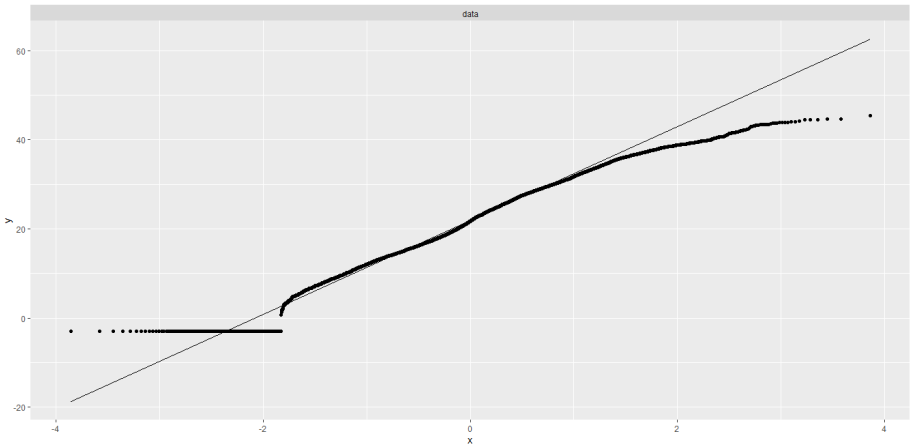


Figure 5: The QQ-plot of the target variable. The plot shows the need for a variable transformation of the target variable.



(a) The box-cox transformation



(b) The QQ-plot of the transformed data

Figure 6: We applied the box-cox transformation to make the target variable closer to a normal distribution. However, since the given data includes zeros, we added one. (a) The λ required for the box-cox transformation was estimated to be approximately 0.34 as the optimal value. (b) The result after applying the transformation using the optimal λ.

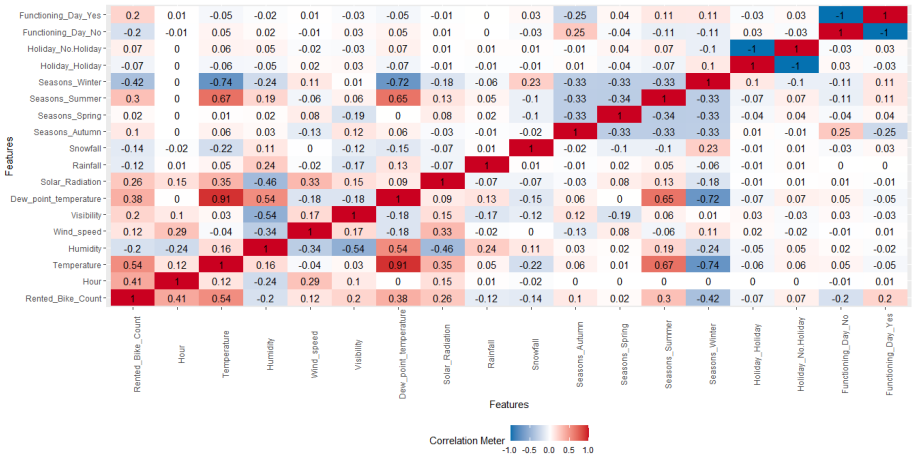


Figure 7: Correlation heatmap for all features. Note that the Dew_point_temperature is highly correlated to the Temperature variable (0.92). To reduce multicollinearity, it seems advisable not to use the Temperature (or Dew_point_temperature) variable.

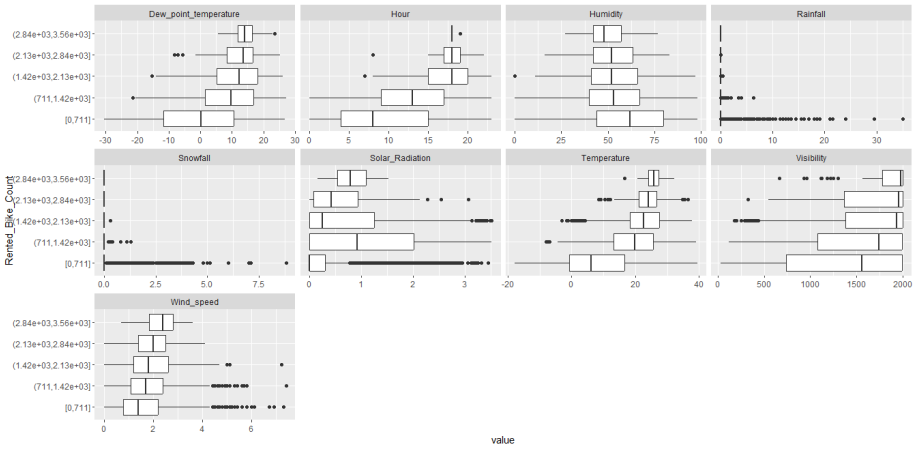


Figure 8: The y-axis in each boxplot represents the target variable divided into five equal parts. These figures allow us to visually inspect the results of Figure 7. For instance, an increase in the Temperature is observed with an increase in the target variable.

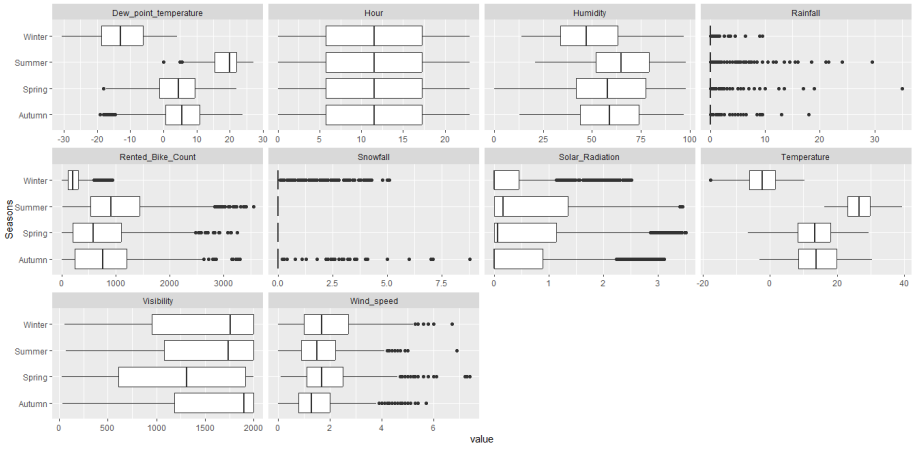


Figure 9: To examine the season effect, boxplots whose y-axis is the Seasons variable are shown, but it is hard to check visually.

2.2 Model specification and assumptions

We begin with a simple (multiple) linear regression. Given a data set $\{y_i, x_{i1}, ..., x_{ip}\}_{i=1}^n$, the model takes the form

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i \quad \forall i. \tag{1}$$

Herein, the ϵ_i is the error term. This variable captures all other factors that influence the dependent variable y other than the regressors x.

Standard linear regression models with standard estimation techniques make many assumptions about the predictor variables, the response variables, and their relationship.

- **Linearity**: The relationship between the dependent variable and the independent variables should be linear.
- **Independence**: The observations should be independent. This implies that the residuals (errors) are not correlated.
- **Homoscedasticity**: The errors of the response variables are uncorrelated.
- **Normality**: The residuals should be normally distributed (more specifically, $e_i \sim N(0, \sigma^2)$). This assumption is important for conducting hypothesis tests and constructing confidence intervals.
- **No Multicollinearity**: There should be no strong correlations among the independent variables.

2.3 Estimation of a linear regression model

We have insights from the results of the EDA, but let's proceed with modeling without using them. Assuming the conditions presented in Section 2.2, we obtain the following results.

```
Call:
lm(formula = Rented_Bike_Count ~ Hour + Temperature + Humidity +
    Wind_speed + Visibility + Dew_point_temperature + Solar_Radiation +
    Rainfall + Snowfall + Seasons + Holiday + Functioning_Day,
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1215.94  -274.37   -57.39   211.09  2282.82

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)           -8.205e+01  9.861e+01  -0.832 0.405404
Hour                   2.753e+01  7.347e-01  37.470  < 2e-16 ***
Temperature            1.607e+01  3.662e+00   4.388 1.16e-05 ***
Humidity              -1.081e+01  1.030e+00 -10.498  < 2e-16 ***
Wind_speed             1.920e+01  5.095e+00   3.769 0.000165 ***
Visibility             1.029e-02  9.881e-03   1.042 0.297635
Dew_point_temperature  1.116e+01  3.835e+00   2.911 0.003608 **
Solar_Radiation       -7.731e+01  7.593e+00 -10.182  < 2e-16 ***
Rainfall              -5.848e+01  4.270e+00 -13.694  < 2e-16 ***
Snowfall               3.269e+01  1.121e+01   2.918 0.003534 **
SeasonsSpring         -1.353e+02  1.388e+01  -9.749  < 2e-16 ***
SeasonsSummer         -1.545e+02  1.721e+01  -8.976  < 2e-16 ***
SeasonsWinter         -3.661e+02  1.971e+01 -18.574  < 2e-16 ***
HolidayNo Holiday      1.176e+02  2.160e+01   5.442 5.40e-08 ***
Functioning_DayYes     9.321e+02  2.665e+01  34.974  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 432.8 on 8745 degrees of freedom
Multiple R-squared:  0.5504,    Adjusted R-squared:  0.5497
F-statistic: 764.8 on 14 and 8745 DF,  p-value: < 2.2e-16
```

Figure 10: The summary of the resulting model. The coefficient of Visibility feature is very small and the associated statistical test tells us that we can reject the coefficient could be zero.

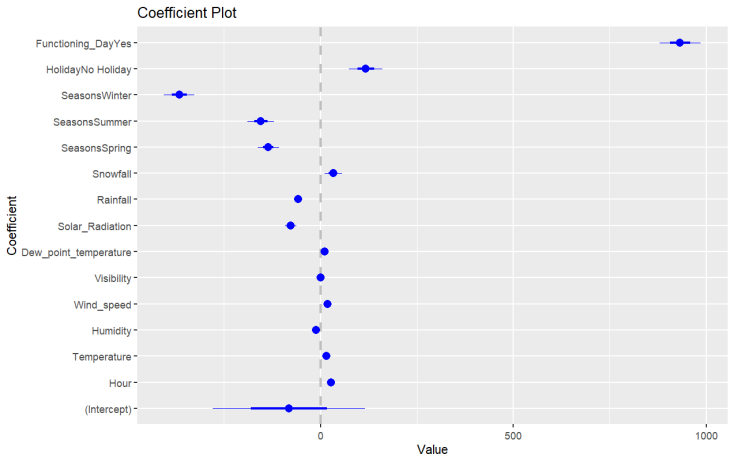


Figure 11: The estimated coefficients and their confidence intervals.

2.4 Diagnostics

We have fitted a linear regression model, but we need to check if this model satisfies the assumptions presented in Section 2.2. Above all, the R^2 or R^2_{adj} values presented in Figure 10 is lower than expected.

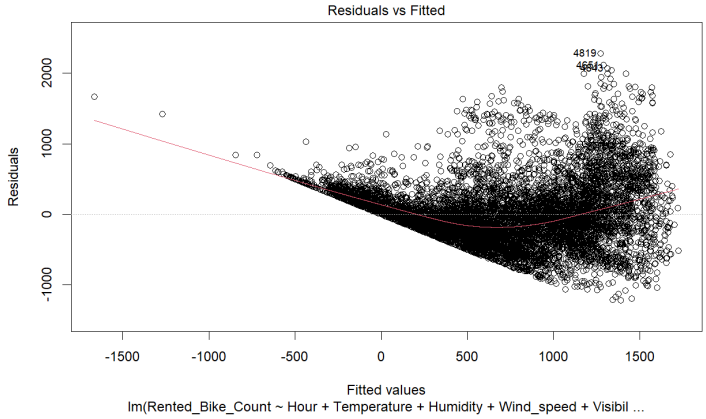


Figure 12: This is a scatterplot of the residuals against the predicted values. We can see that there is a tendency (red curve) between the fitted values and the residuals, which may violate the Linearity or/and Normality assumptions.

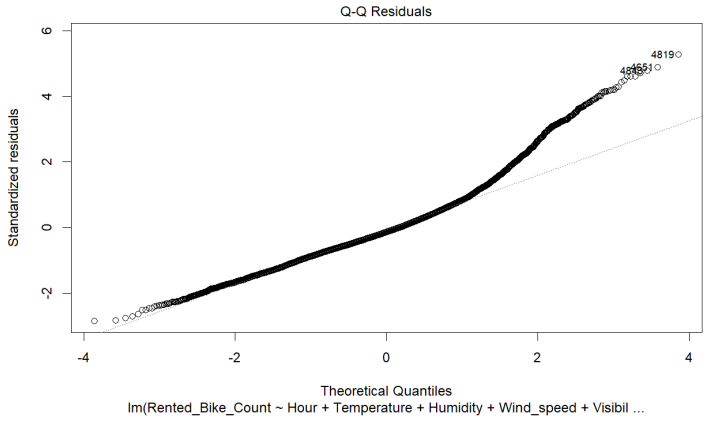


Figure 13: The QQ-plot of the residuals clearly shows that the Normality assumption does not hold on the current setting.

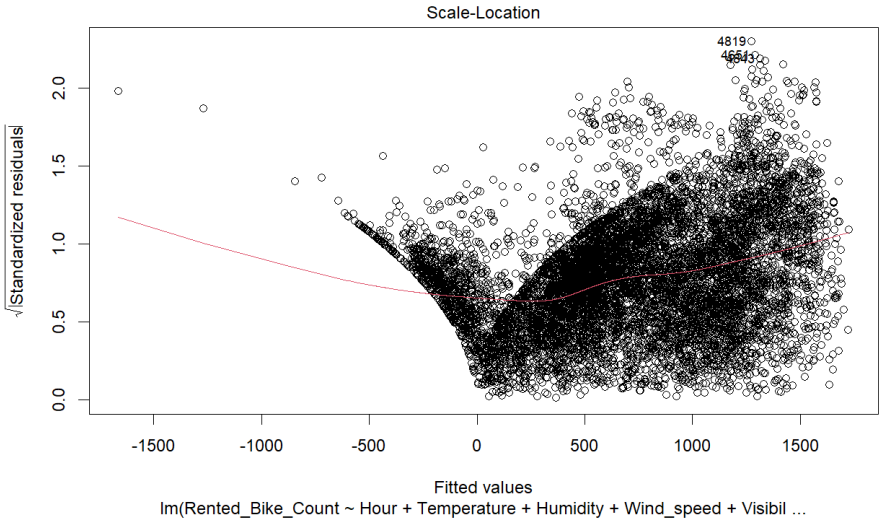


Figure 14: To check homoscedasticity, a scatterplot of the (standardized) residuals against the predicted values. The model does not satisfy homoscedasticity.

	GVIF	Df	GVIF^(1/(2*Df))
Hour	1.209577	1	1.099808
Temperature	89.477069	1	9.459232
Humidity	20.553911	1	4.533642
Wind_speed	1.303644	1	1.141772
Visibility	1.689144	1	1.299671
Dew_point_temperature	117.298694	1	10.830452
Solar_Radiation	2.034617	1	1.426400
Rainfall	1.085306	1	1.041780
Snowfall	1.119845	1	1.058227
Seasons	5.526992	3	1.329683
Holiday	1.023340	1	1.011603
Functioning_Day	1.080974	1	1.039699

Figure 15: Finally, the multicollinearity can be checked by calculating Variance Inflation Factor (VIF). Usually, a VIF greater than 10 indicates a high correlation. In our case, Dew_point_temperature, Humidity, Temperature are greater than 10. Note that this result is consistent with Figure 7.

To remedy these problems, the target data y is transformed to y_{bc} (See Figure 6) and the Dew_point_temperature feature is removed because of Figure

7 and Figure 15. After that, we re-fit the same model on the remedied dataset, and the results are as follows.

```
Call:
lm(formula = Rented_Bike_Count_bc ~ Hour + Temperature + Humidity +
    Wind_speed + Visibility + Solar_Radiation + Rainfall + Snowfall +
    Seasons + Holiday + Functioning_Day, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-24.655  -3.289  -0.053   3.287  37.261

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -6.380e+00  6.361e-01 -10.030  < 2e-16 ***
Hour                3.590e-01  9.223e-03  38.927  < 2e-16 ***
Temperature         3.463e-01  1.092e-02  31.722  < 2e-16 ***
Humidity           -1.232e-01  4.625e-03 -26.645  < 2e-16 ***
Wind_speed          3.503e-02  6.395e-02   0.548    0.584
Visibility          2.338e-05  1.239e-04   0.189    0.850
Solar_Radiation    -4.866e-01  9.328e-02  -5.217 1.86e-07 ***
Rainfall           -1.316e+00  5.328e-02 -24.708  < 2e-16 ***
Snowfall           -5.778e-03  1.403e-01  -0.041    0.967
SeasonsSpring      -2.420e+00  1.742e-01 -13.887  < 2e-16 ***
SeasonsSummer      -2.190e+00  2.153e-01 -10.170  < 2e-16 ***
SeasonsWinter      -6.314e+00  2.475e-01 -25.512  < 2e-16 ***
HolidayNo Holiday   2.248e+00  2.713e-01   8.285  < 2e-16 ***
Functioning_DayYes  2.843e+01  3.346e-01  84.954  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.436 on 8746 degrees of freedom
Multiple R-squared:  0.6959,    Adjusted R-squared:  0.6955
F-statistic:  1540 on 13 and 8746 DF,  p-value: < 2.2e-16
```

Figure 16: The summary of the re-fitted model. It has improved in almost every aspect; standard error of the coefficients and residuals, R^2, and R^2_{adj} values. Interestingly, the R^2 value improved despite the reduction in the number of features.

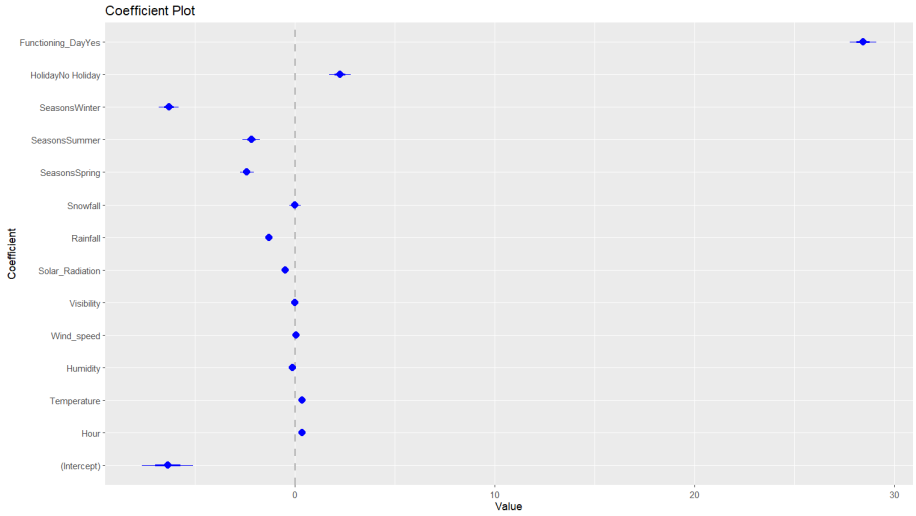
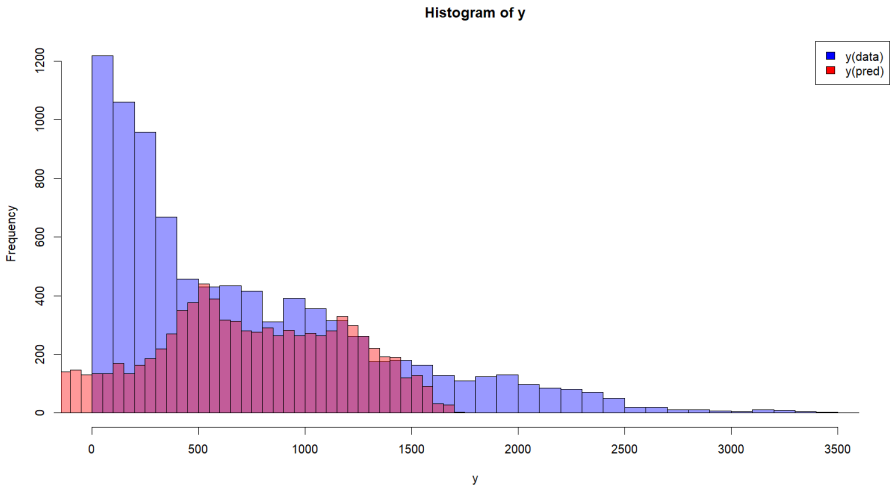
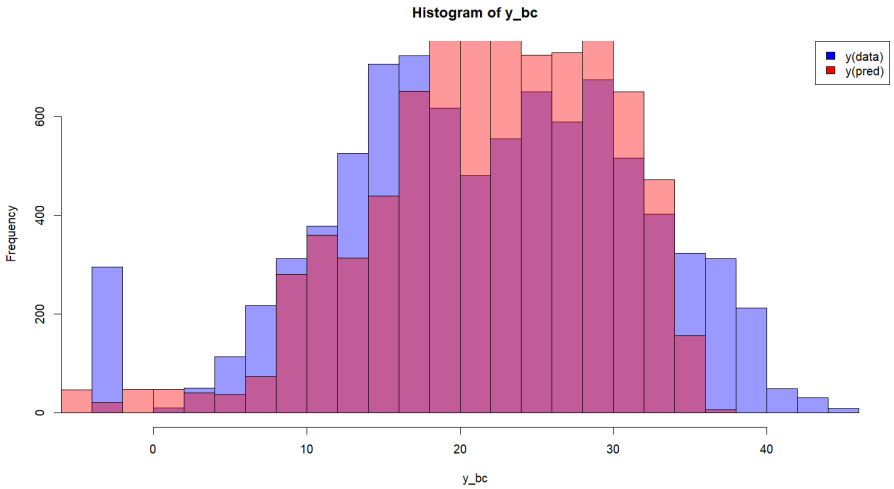


Figure 17: A visualization of the estimated coefficients. As said in Figure 16, the width of their confidence intervals decreased compared to Figure 11.



(a) Histograms of y and $\hat{y}$



(b) Histograms of y_{bc} and $\hat{y_{bc}}$

Figure 18: To emphasize the results in Figure 16, we compared the estimated values of the previous model and the current model using the training data. The training data of the previous model is non-negative, whereas the current model's training data may include negative values (because of the box-cox transformation).

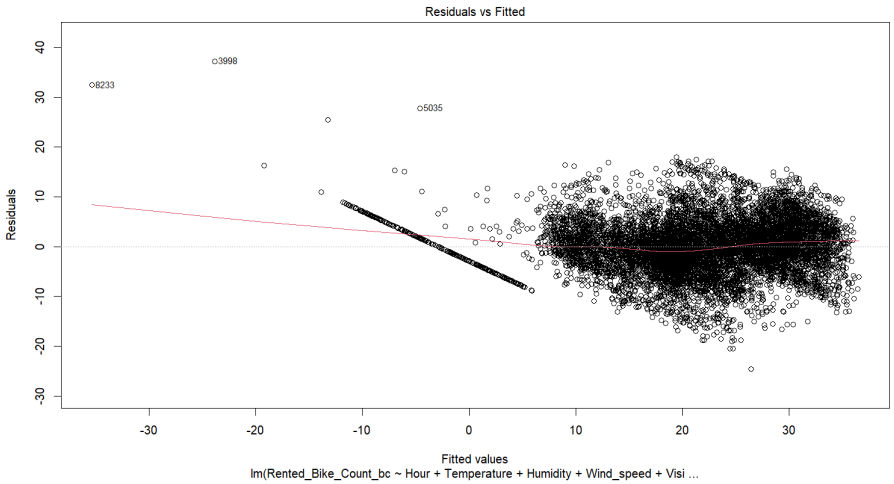


Figure 19: A scatterplot of the residuals against the predicted values. The closer the red curve is to horizontal and near zero, the better. It seems to have improved compared to Figure 12. The estimated values around zero show some strange behavior in the residuals. This will be addressed later (Section 2.7).

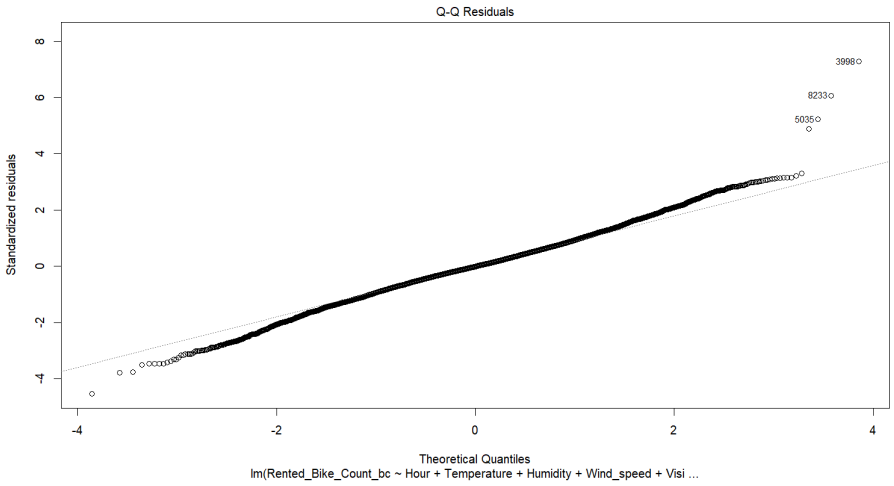


Figure 20: The QQ-plot of the residuals. It has improved compared to Figure 13. There would be some outliers or influential points.

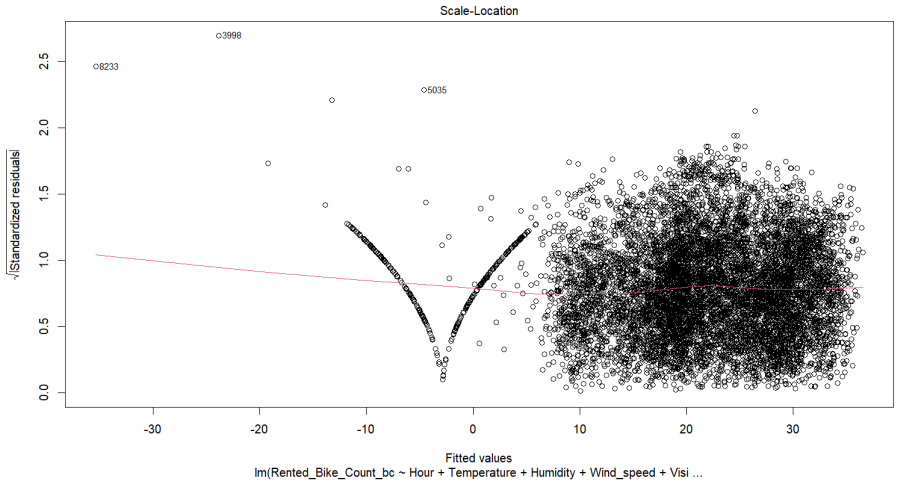


Figure 21: Compared to Figure 14, the points in the right area seem to satisfy homoscedasticity. On the other hand, those with predicted values around 0 exhibit strange behavior similar to Figure 19.

2.5 Discussion on the season effect

In this section, we will test whether we can omit the Seasons variable or not. Let $\beta_{Seasons1}, \beta_{Seasons2}, \beta_{Seasons3}$ denote the coefficients in the linear model (1) (Remark that the Seasons variable has four categories, resulting in three coefficients). The hypothesis test is

$$H_0 : \beta_{Seasons1} = \beta_{Seasons2} = \beta_{Seasons3} = 0 \quad \text{versus} \quad H_1 : \neg H_0,$$

and its associated test statistic is

$$F^* = \frac{SSR(X_{Seasons}|X_{others})/p - q}{SSE(all)/n - p} \sim F(p - q, n - p),$$

where p is the number of features in the full model, q is the number of features in the reduced model, and n is the number of the data. By the R implementation, we obtain $F^* = 436.2541$, and since $F(2, 8746, 0.95) = 2.99$, we can reject H_0, that is, we can not omit the Seasons variable.

2.6 Variable selection

Variable selection was performed using the stepwise forward selection method. Variables were selected in the direction of decreasing AIC.

```
Step:  AIC=29670.48
Rented_Bike_Count_bc ~ Hour + Temperature + Humidity + Solar_Radiation +
    Rainfall + Seasons + Holiday + Functioning_Day

                  Df Sum of Sq    RSS   AIC
<none>                         258456 29670
+ Wind_speed       1        10 258446 29672
+ Visibility       1         2 258454 29672
+ Snowfall         1         0 258456 29672
- Solar_Radiation  1       875 259331 29698
- Holiday          1      2025 260481 29737
- Rainfall         1     18084 276540 30261
- Seasons          3     26494 284949 30519
- Temperature      1     30100 288556 30633
- Humidity         1     32667 291123 30711
- Hour             1     48017 306473 31161
- Functioning_Day  1    213523 471978 34944

Call:
lm(formula = Rented_Bike_Count_bc ~ Hour + Temperature + Humidity +
    Solar_Radiation + Rainfall + Seasons + Holiday + Functioning_Day,
    data = df)

Coefficients:
       (Intercept)                Hour         Temperature            Humidity
           -6.2481              0.3601              0.3460             -0.1240
   Solar_Radiation            Rainfall       SeasonsSpring       SeasonsSummer
           -0.4788             -1.3154             -2.4148             -2.1796
     SeasonsWinter   HolidayNo Holiday  Functioning_DayYes
           -6.3131              2.2448             28.4181
```

Figure 22: The final result from variable selection. An interesting point is that it aligns with coefficients tested to be zero in Figure 16. Based on these results, we removed variables Wind_speed, Visibility, and Snowfall. We also re-fitted the model. Because their coefficients were very small, we won't list the results of the refitted model.

2.7 Ouliers and influential points

To identify outlying y observations, we calculated Studentized Deleted Residuals (or Externally Studentized Residuals) and tested whether their values are greater than $t(1 - \frac{0.95}{2n}, n - p - 1) = 3.873$ (Bonferroni critical value). Recall that by variable selection, $p = 11$.

The outlier indices according to the test results are 3998, 5035, 6502, and 8233. These indices are represented in Figures 19, 20, and 21. Thus, calculating the SDR effectively detected the outliers.

Similarly, outliers for x also need to be detected. This was done using the hat matrix (or leverages) and a critical value of $\frac{2p}{n}$. As a result, 785 data points were detected (too many to list). Here are some interesting findings about these outliers.

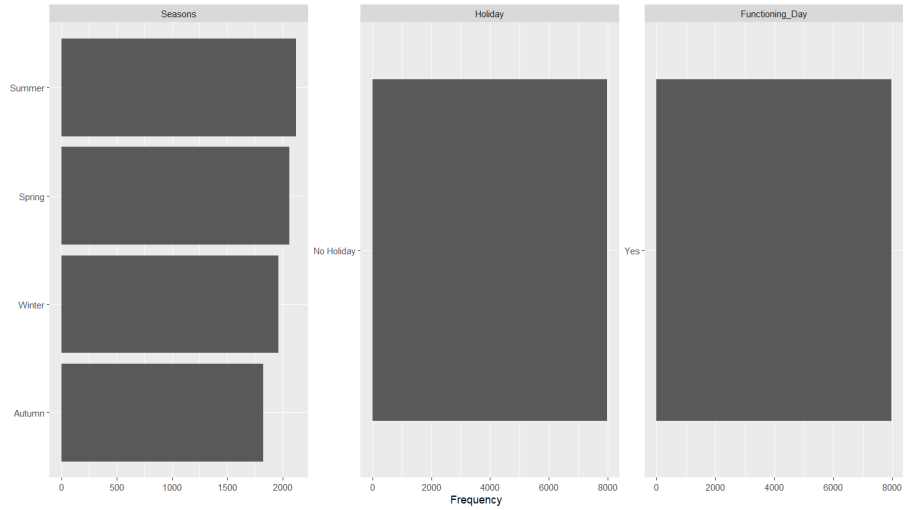


Figure 23: The data after removing outliers. Upon redoing the EDA on this data, we find that one class of the Holiday and Functioning_day variables has disappeared. This indicates that the data for a specific class was distributed significantly differently from the rest of the data. To address this, we could consider refitting the model with an interaction term. However, since the data for this specific class was sparse (10%), we simply removed the outliers and refitted the model.

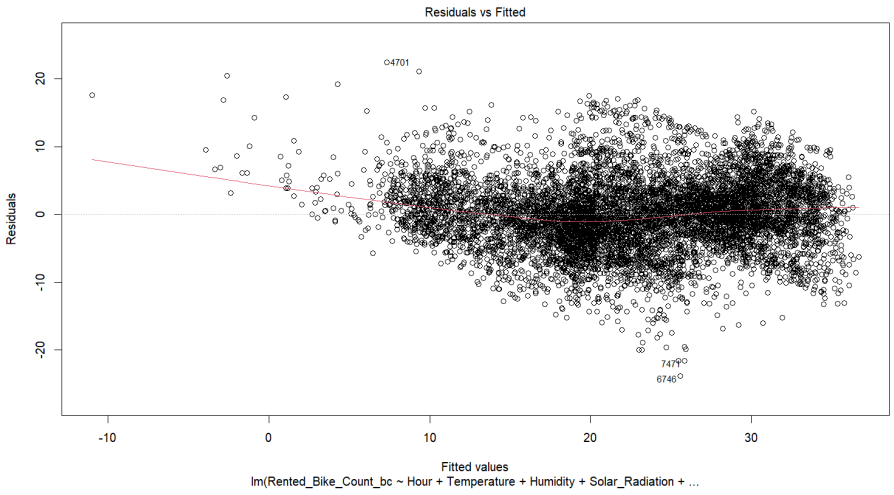


Figure 24: For the re-fitted model, we draw a scatterplot of the residuals against the predicted values. Compared to Figure 19, the strange points have finally been removed.

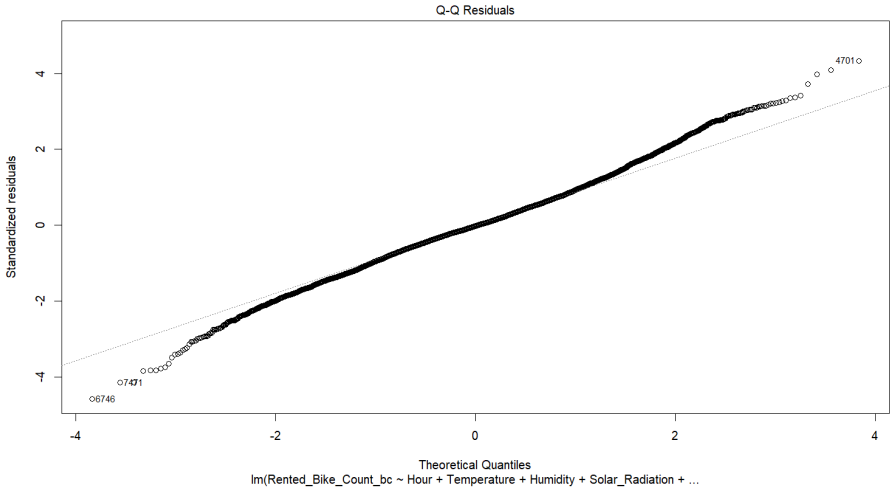


Figure 25: The QQ-plot of the residuals of the re-fitted model. It has improved compared to Figure 20. The 4701, 7471, and 6746 observations may be expected to be influential points.

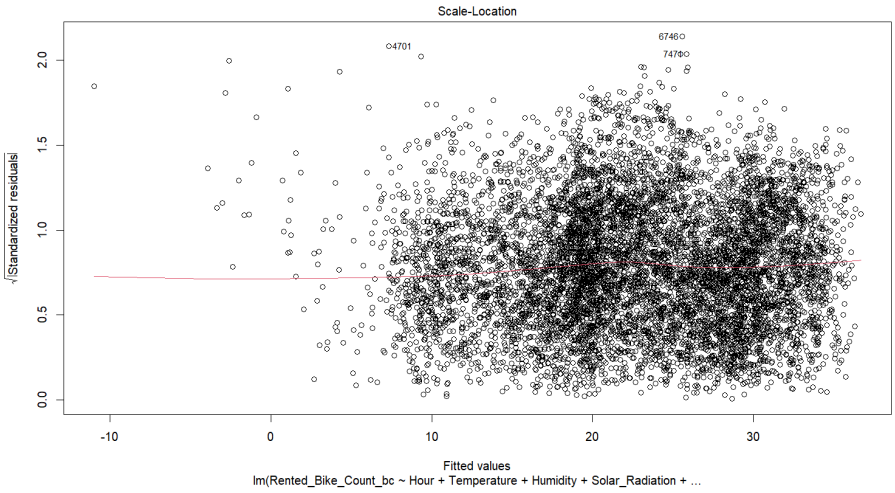
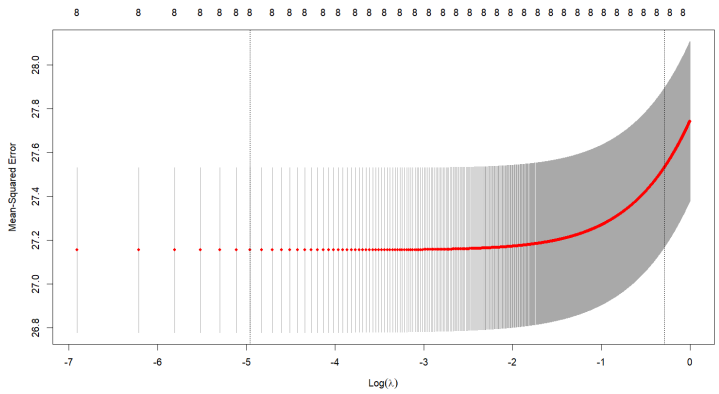


Figure 26: Compared to Figure 21, all points in the figure seem to satisfy homoscedasticity!

```
> # influential
> inf = as.data.frame(cbind(
+   "DFFITS" =dffits(fit_4),
+   "D" =cooks.distance(fit_4)
+ ))
>
> n = 7975
> th = 2*sqrt(p/n)
> influential_indices_DFFFIT = which(inf$DFFITS >= th)
> th = 4/n
> influential_indices_Cooks = which(inf$D >= th)
> influential_indices = union(influential_indices_DFFFIT, influential_indices_Cooks)
>
> 4701 %in% influential_indices
[1] TRUE
> 6746 %in% influential_indices
[1] TRUE
> 7471 %in% influential_indices
[1] TRUE
```

Figure 27: To detect influential points, we calculated DFFITS and Cook's distance, setting the respective thresholds to $2\sqrt{\frac{p}{n}}$ and $\frac{4}{n}$ (commonly used values for large samples). The points identified in Figure 25 were indeed influential.

2.8 Estimation of ridge regression



(a) Optimization of λ

```
> bestlam
[1] 0.007007007
> out=glmnet(X,Y,alpha=0,lambda=bestlam)
> predict(out,type="coefficients",s=bestlam)
9 x 1 sparse Matrix of class "dgCMatrix"
                      s1
(Intercept)     23.4187409
Hour             0.3939611
Temperature      0.3492252
Humidity        -0.1101362
Solar_Radiation -0.4934182
Rainfall        -5.1297297
SeasonsSpring   -2.2827838
SeasonsSummer   -2.2425669
SeasonsWinter   -6.2248842
> coefficients(fit_4)
    (Intercept)            Hour     Temperature        Humidity Solar_Radiation
     23.4199673       0.3939203       0.3506966      -0.1103213      -0.5001060
       Rainfall   SeasonsSpring   SeasonsSummer   SeasonsWinter
      -5.1320228      -2.2881611      -2.2659783      -6.2140952
```

(b) The result of the ridge regression.

Figure 28: (a) The process of optimizing the λ to be used in ridge regression, resulting in approximately 0.007. (b) After fitting the ridge regression using the optimal lambda, we compared the estimated coefficients with the original ones. Since the λ was very small, there seems to be little difference from the original values.

2.9 K-fold cross-validation

We split the data into five sub-data, using one part as the test dataset and the rest as the training dataset. After training, we can make predictions using the test dataset and calculate the MSE between y_{test} and $y_{pred}(x_{test})$. This process can be repeated five times. Computing the average of these MSE values obtained through repetition constitutes k-fold Cross-Validation method. For linear regression, we obtained a result of 27.15593, and for ridge regression, we obtained 27.15575. As noted in Section 2.8, the λ values are low in this experiment (0.01 0.02). Therefore, there may not be a significant difference in MSE.

3 Conclusion

In this study, we conducted EDA to visualize the data and gain some insights. Through linear regression, we were able to enhance the model by diagnostics, variable selection, and outlier detection. Additionally, comparing with ridge regression suggests that our model is not ill-conditioned and the estimation is stable.