Stats 101A, Project

Shabib Alam – 106038360

04/24/24

Contents

Problem Description

Currently Rental bikes are introduced in many urban cities for the enhancement of mobility comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

Data Description

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour, Seasons, Holiday and Functioning.Day information.**Loading the data from SeouBikeData.csv dataset, and get the details**

```
##  Rented.Bike.Count      Hour       Temperature..C.   Humidity...
##  Min.   :   0.0    Min.   : 0.00   Min.   :-17.80   Min.   : 0.00
##  1st Qu.: 191.0    1st Qu.: 5.75   1st Qu.:  3.50   1st Qu.:42.00
##  Median : 504.5    Median :11.50   Median : 13.70   Median :57.00
##  Mean   : 704.6    Mean   :11.50   Mean   : 12.88   Mean   :58.23
##  3rd Qu.:1065.2    3rd Qu.:17.25   3rd Qu.: 22.50   3rd Qu.:74.00
##  Max.   :3556.0    Max.   :23.00   Max.   : 39.40   Max.   :98.00
##  Wind.speed..m.s. Visibility..10m. Dew.point.temperature..C.
##  Min.   :0.000    Min.   :  27     Min.   :-30.600
##  1st Qu.:0.900    1st Qu.: 940     1st Qu.: -4.700
##  Median :1.500    Median :1698     Median :  5.100
##  Mean   :1.725    Mean   :1437     Mean   :  4.074
##  3rd Qu.:2.300    3rd Qu.:2000     3rd Qu.: 14.800
##  Max.   :7.400    Max.   :2000     Max.   : 27.200
##  Solar.Radiation..MJ.m2.  Rainfall.mm.     Snowfall..cm.       Seasons
##  Min.   :0.0000           Min.   : 0.0000   Min.   :0.00000   Length:8760
##  1st Qu.:0.0000           1st Qu.: 0.0000   1st Qu.:0.00000   Class :character
##  Median :0.0100           Median : 0.0000   Median :0.00000   Mode  :character
##  Mean   :0.5691           Mean   : 0.1487   Mean   :0.07507
##  3rd Qu.:0.9300           3rd Qu.: 0.0000   3rd Qu.:0.00000
##  Max.   :3.5200           Max.   :35.0000   Max.   :8.80000
##    Holiday          Functioning.Day
##  Length:8760        Length:8760
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

Data Cleaning

Removing Date Variable from dataset

In my analysis, **does not require information about the specific dates** but rather focuses on other variables such as “Hour”, “Temperature_C”, “Humidity”, “Wind_speed_m_s”, “Visibility_10m”, “Dew_point_temperature_C”, “Solar_Radiation_MJ_m2”, “Rainfall_mm”, “Snowfall_cm”, “Seasons”, “Holiday”, and “Functioning_Day”. So, I have data without data.After that I need to check others stuff to see for data cleaning.

Finding NA, infinite, duplicate value

There are **no missing values** in any of the columns of the seoul_bike_clean dataset.There are **no duplicate values** in the seoul_bike_clean dataset. Most of numeric columns (Rented.Bike.Count, Hour, Temperature..C., Humidity…, Wind.speed..m.s., Visibility..10m., Dew.point.temperature..C., and Solar.Radiation..MJ.m2.,Rainfall.mm.) have finite values, indicating but there are some **infinite values** present in these columns such as , **Snowfall..cm., Seasons, Holiday, Functioning.Day**.

Convert infiinite value into finite value

By using as.factor function, I converted nonfinite value into finite value from **Snowfall..cm., Seasons, Holiday, Functioning.Day**. Now overall, it seems that the seoul_bike_clean dataset is free from missing values, duplicates, and infinite values in the numeric columns, which is good for further analysis or modeling.

Multicolinearity Test

```
##                                GVIF Df GVIF^(1/(2*Df))
## Hour                       1.209577  1        1.099808
## Temperature..C.           89.477069  1        9.459232
## Humidity...               20.553911  1        4.533642
## Wind.speed..m.s.           1.303644  1        1.141772
## Visibility..10m.           1.689144  1        1.299671
## Dew.point.temperature..C. 117.298694  1       10.830452
## Solar.Radiation..MJ.m2.    2.034617  1        1.426400
## Rainfall.mm.               1.085306  1        1.041780
## Snowfall..cm.              1.119845  1        1.058227
## Holiday                    1.023340  1        1.011603
## Functioning.Day            1.080974  1        1.039699
## Seasons                    5.526992  3        1.329683
```

High GVIF values, such as those observed for Temperature..C. and Dew.point.temperature..C., suggest potential multicollinearity issues that may require further investigation or remediation techniques, such as variable transformation or stepwise function . Conversely, variables with low GVIF values are less affected by multicollinearity concerns and can be considered reliable predictors in the regression model.

Stepwise Regression

Before Stepwise

I conducted stepwise regression to streamline the model by eliminating non-essential variables. The goal was to enhance model efficiency and interpretability by focusing on the most influential predictors.

```
backward_model <- step(lm_model, direction = "backward")
```

```
## Start:  AIC=106366.7
## Rented.Bike.Count ~ Hour + Temperature..C. + Humidity... + Wind.speed..m.s. +
##     Visibility..10m. + Dew.point.temperature..C. + Solar.Radiation..MJ.m2. +
##     Rainfall.mm. + Snowfall..cm. + Holiday + Functioning.Day +
##     Seasons
##
##                             Df Sum of Sq        RSS    AIC
## - Visibility..10m.           1    203228 1638368966 106366
## <none>                                   1638165738 106367
## - Dew.point.temperature..C.  1   1587705 1639753443 106373
## - Snowfall..cm.              1   1594789 1639760527 106373
## - Wind.speed..m.s.           1   2661081 1640826819 106379
## - Temperature..C.            1   3607143 1641772881 106384
## - Holiday                    1   5548264 1643714003 106394
```

```
## - Solar.Radiation..MJ.m2.    1  19419052 1657584790 106468
## - Humidity...                1  20643815 1658809554 106474
## - Rainfall.mm.               1  35130814 1673296552 106551
## - Seasons                    3  95027170 1733192908 106855
## - Functioning.Day            1 229135968 1867301706 107512
## - Hour                       1 263012687 1901178425 107669
##
## Step:  AIC=106365.8
## Rented.Bike.Count ~ Hour + Temperature..C. + Humidity... + Wind.speed..m.s. +
##     Dew.point.temperature..C. + Solar.Radiation..MJ.m2. + Rainfall.mm. +
##     Snowfall..cm. + Holiday + Functioning.Day + Seasons
##
##                             Df Sum of Sq        RSS    AIC
## <none>                                   1638368966 106366
## - Snowfall..cm.              1   1572079 1639941045 106372
## - Dew.point.temperature..C.  1   1659150 1640028116 106373
## - Wind.speed..m.s.           1   2826411 1641195377 106379
## - Temperature..C.            1   3523101 1641892067 106383
## - Holiday                    1   5511535 1643880502 106393
## - Solar.Radiation..MJ.m2.    1  21049622 1659418588 106476
## - Humidity...                1  23182043 1661551009 106487
## - Rainfall.mm.               1  35348843 1673717809 106551
## - Seasons                    3  97001442 1735370408 106864
## - Functioning.Day            1 228985543 1867354509 107510
## - Hour                       1 263241463 1901610429 107669
```

```
summary(backward_model)
```

```
##
## Call:
## lm(formula = Rented.Bike.Count ~ Hour + Temperature..C. + Humidity... +
##     Wind.speed..m.s. + Dew.point.temperature..C. + Solar.Radiation..MJ.m2. +
##     Rainfall.mm. + Snowfall..cm. + Holiday + Functioning.Day +
##     Seasons, data = seoul_bike)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1232.45  -274.71   -59.16   211.47  2278.27
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -46.6952    92.5841  -0.504 0.614025
## Hour                        27.4780     0.7330  37.487  < 2e-16 ***
## Temperature..C.             15.8574     3.6565   4.337 1.46e-05 ***
## Humidity...                -11.0807     0.9961 -11.124  < 2e-16 ***
## Wind.speed..m.s.            19.7038     5.0726   3.884 0.000103 ***
## Dew.point.temperature..C.   11.3943     3.8287   2.976 0.002928 **
## Solar.Radiation..MJ.m2.    -78.8827     7.4415 -10.600  < 2e-16 ***
## Rainfall.mm.               -58.6288     4.2680 -13.737  < 2e-16 ***
## Snowfall..cm.               32.4539    11.2029   2.897 0.003778 **
## HolidayNo Holiday          117.1694    21.6013   5.424 5.98e-08 ***
## Functioning.DayYes         931.7363    26.6496  34.963  < 2e-16 ***
## SeasonsSpring             -138.5285    13.5316 -10.237  < 2e-16 ***
## SeasonsSummer             -153.5290    17.1863  -8.933  < 2e-16 ***
```

```
## SeasonsWinter             -369.7508    19.3863 -19.073  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 432.8 on 8746 degrees of freedom
## Multiple R-squared:  0.5504, Adjusted R-squared:  0.5497
## F-statistic: 823.6 on 13 and 8746 DF,  p-value: < 2.2e-16
```

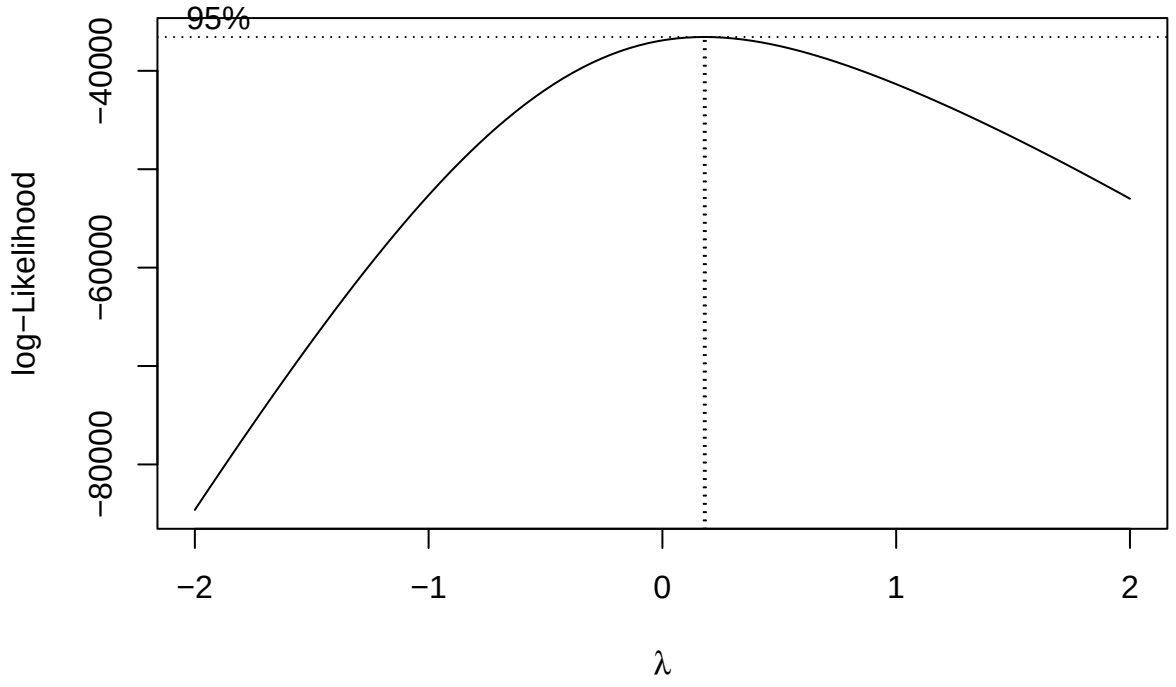
After Stepwise

The stepwise regression procedure helped refine the model by eliminating the least significant predictor, Visibility..10m., while retaining other relevant variables. This streamlined model can provide more accurate predictions of rented bike counts based on the remaining predictors.

Transformation

Box Cox Transformation

The Box-Cox transformation is a technique used to stabilize the variance and improve the normality of the residuals in linear regression models. By applying this transformation to the response variable, we aim to address issues such as non-normality in the model's residuals, and make model better.




```
##
```

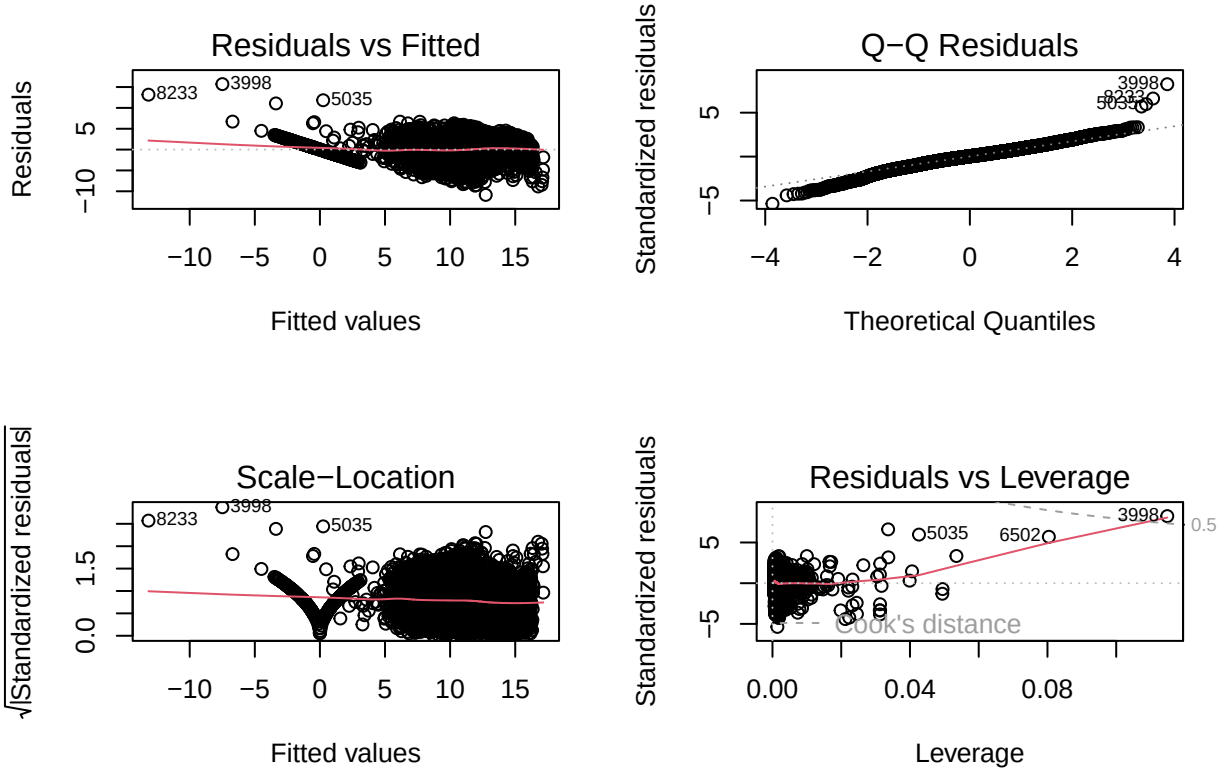
```
## Call:
## lm(formula = Rented.Bike.Count_transformed ~ Hour + Temperature..C. +
##     Humidity... + Wind.speed..m.s. + Dew.point.temperature..C. +
##     Solar.Radiation..MJ.m2. + Rainfall.mm. + Snowfall..cm. +
##     Holiday + Functioning.Day + Seasons, data = seoul_bike)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8652  -1.1305   0.0726   1.2166  15.7000
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2.623919   0.433360   6.055 1.46e-09 ***
## Hour                        0.129359   0.003431  37.703  < 2e-16 ***
## Temperature..C.            -0.035145   0.017115  -2.053   0.0401 *
## Humidity...                -0.089955   0.004662 -19.294  < 2e-16 ***
## Wind.speed..m.s.           -0.002559   0.023744  -0.108   0.9142
## Dew.point.temperature..C.   0.172438   0.017921   9.622  < 2e-16 ***
## Solar.Radiation..MJ.m2.    -0.061306   0.034832  -1.760   0.0784 .
## Rainfall.mm.               -0.532974   0.019977 -26.679  < 2e-16 ***
## Snowfall..cm.              -0.004969   0.052438  -0.095   0.9245
## HolidayNo Holiday           0.918683   0.101109   9.086  < 2e-16 ***
## Functioning.DayYes         12.694877   0.124739 101.772  < 2e-16 ***
## SeasonsSpring              -0.908297   0.063338 -14.341  < 2e-16 ***
## SeasonsSummer              -0.863077   0.080444 -10.729  < 2e-16 ***
## SeasonsWinter              -2.367323   0.090742 -26.089  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.026 on 8746 degrees of freedom
## Multiple R-squared:  0.7257, Adjusted R-squared:  0.7253
## F-statistic:  1780 on 13 and 8746 DF,  p-value: < 2.2e-16


##
## Call:
## lm(formula = Rented.Bike.Count ~ Hour + Humidity... + Visibility..10m. +
##     Dew.point.temperature..C. + Rainfall.mm. + Holiday + Functioning.Day +
##     Seasons, data = seoul_bike)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1268.08  -271.24   -70.95   208.94  2315.61
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.113e+02  4.570e+01   2.435   0.0149 *
## Hour                        2.900e+01  7.122e-01  40.717  < 2e-16 ***
## Humidity...                -1.237e+01  3.596e-01 -34.386  < 2e-16 ***
## Visibility..10m.            2.812e-02  9.654e-03   2.912   0.0036 **
## Dew.point.temperature..C.   2.462e+01  8.632e-01  28.525  < 2e-16 ***
## Rainfall.mm.               -5.758e+01  4.262e+00 -13.508  < 2e-16 ***
## HolidayNo Holiday           1.155e+02  2.173e+01   5.315 1.09e-07 ***
## Functioning.DayYes          9.323e+02  2.680e+01  34.790  < 2e-16 ***
## SeasonsSpring              -1.365e+02  1.377e+01  -9.909  < 2e-16 ***
```

```
## SeasonsSummer              -1.407e+02  1.724e+01  -8.162 3.76e-16 ***
## SeasonsWinter              -3.737e+02  1.952e+01 -19.148  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 435.5 on 8749 degrees of freedom
## Multiple R-squared:  0.5446, Adjusted R-squared:  0.5441
## F-statistic:  1046 on 10 and 8749 DF,  p-value: < 2.2e-16
```

Since p value is lower of Temperature, windspeed, and snowfall, I remove those variable to make my model more effective.

Diagnostic Plot

The diagnostic plot displays four diagnostic plots: **Residuals vs Fitted, Normal Q-Q plot, Scale-Location plot, and Residuals vs Leverage**. These plots are used to assess the assumptions of linear regression.This is also better than before that tell us that model became better after doing transformation.



Conclusion

In this analysis, a linear regression model was initially built using the variables Hour, Temperature..C., Humidity..., Wind.speed..m.s., Visibility..10m., Dew.point.temperature..C., Solar.Radiation..MJ.m2., Rainfall.mm., Snowfall..cm., Holiday, Functioning.Day, and Seasons to predict the Rented.Bike.Count in

Seoul. The model's diagnostics revealed issues of **multicollinearity**, especially with Temperature..C., Dew.point.temperature..C., and Seasons showing high Variance Inflation Factors (VIFs). Subsequently, a **backward stepwise** regression was performed to address multicollinearity, resulting in the **removal of the Visibility..10m.** variable. Then, a **Box-Cox transformation** was applied to the response variable Rented.Bike.Count to address heteroscedasticity and non-normality in the residuals. The transformed model demonstrated an**improvement in the adjusted R-squared value from 0.5497 to 0.7253, indicating a better fit** to the data. Since p value is lower of Temperature, windspeed, and snowfall, I remove those variable to make my model more effective. Additionally, diagnostic plots showed improved linearity, homoscedasticity, and normality assumptions in the transformed model compared to the original model. Overall, these steps helped in **refining the model's performance and addressing issues** related to multicollinearity and non-normality in the data.