# DC Bike Sharing Analysis

**Chia Yen Chen, Ibrahim Mohamed, Jiyoung Lee, Shalini Sah, Yu Jiyun Chang**

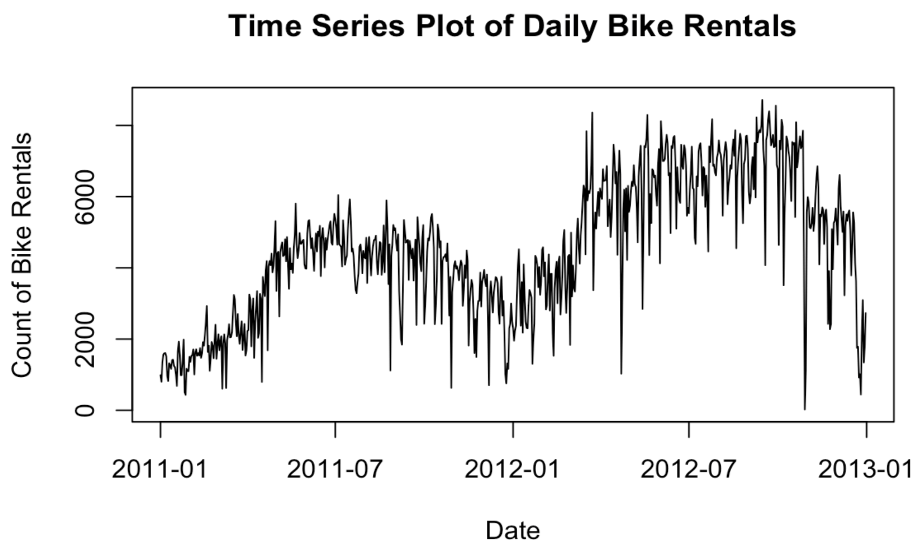
**(Group 1)**

## 1. Introduction and Overview

The DC bike-sharing dataset spans from January 1st, 2011 to December 31st, 2012, and involves 9 variables. These include counts of bike share rentals, date day, month, holiday, weekday, and weather features. Our project focuses on the counts of bike share rentals as the dependent variable. In Section 2 (univariate time series model), we will consider the month variable and sine and cosine pairs as the independent variables, while in Section 3 (time series regression model), we will use weather features as the independent variables. Additionally, our holdout sample will consist of 106 observations, which is 15% of the total dataset.

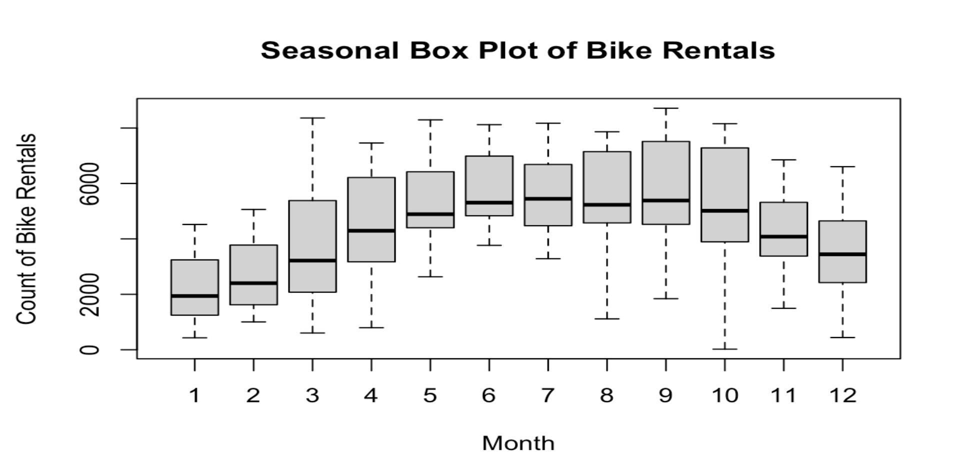
We started examining **Exhibit 1**, a Time Series Plot of Daily Bike Rentals, we observed some fluctuations in bike rental counts over time. These fluctuations suggest the presence of both trend and seasonality in the data. The plot does not follow a flat line, which implies the series is non-stationary. Furthermore, during certain periods, particularly around the middle of the year, there's a visible spike in rentals, likely an indication of seasonal trends such as better weather and increased outdoor activities.

**Exhibit 1:** *Time Series Plot of Daily Bike Rentals*



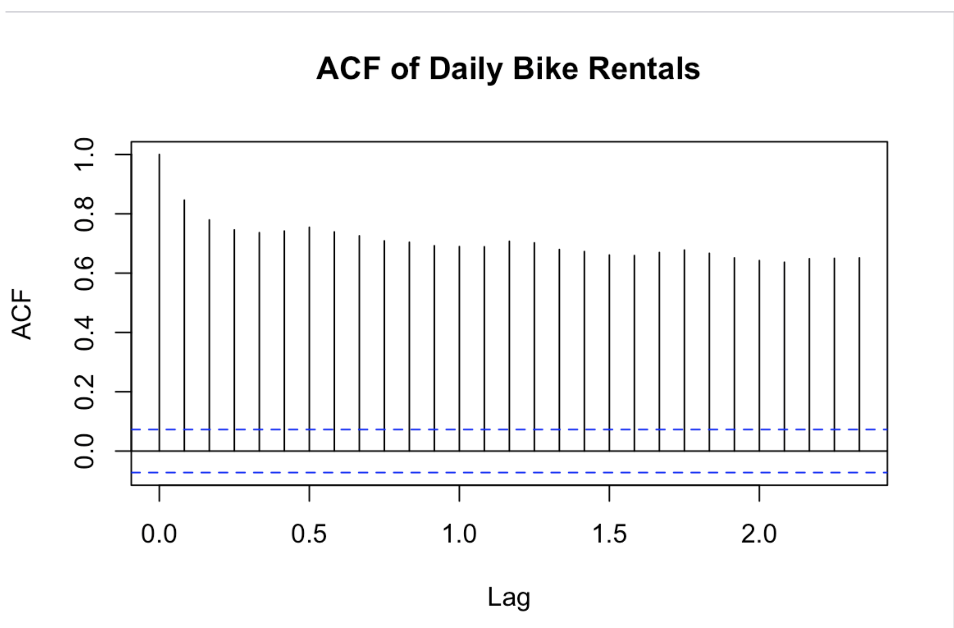

Moving to **Exhibit 2**, the Seasonal Box Plot of Bike Rentals, we can see the changeability of bike rentals throughout different months. The box plots show the typical values and range of rentals, with each box showing the middle 50% of the data and the line inside marking the middle value. From this visual, it's clear that rental patterns are not constant across the year. The higher median and larger range during warmer months suggest a surge in rental activity, possibly due to favorable weather conditions that encourage biking. Conversely, the lower median during the colder months indicates a decrease, likely due to the less conducive weather for outdoor activities.

**Exhibit 2:** *Seasonal Box Plot of Bike Rentals*




Lastly **Exhibit 3**, the Autocorrelation Function (ACF) of Daily Bike Rentals, reveals the correlation between the counts of bike rentals over a series of time lags. The plot shows a gradually declining pattern, indicating a strong positive correlation that decreases as the lags increase. This suggests that past values have an influence on future values in the series. The slow decay of the autocorrelation reconfirms that the series is nonstationary.

**Exhibit 3:** *Autocorrelation Function (ACF) of Daily Bike Rentals*




Now that we've looked at these patterns, we're ready to dive into Section 2 and 3 for more in depth analysis of the Daily Bike Rentals.

**2. Univariate Time-series models.**

## 2.1 Deterministic Time Series Models (Seasonal Dummies and Trend, Cyclical Trend)

### Seasonal Dummies with Trends

```
Call:
lm(formula = n_CNT ~ time + as.factor(n_MONTH))

Residuals:
    Min      1Q  Median      3Q     Max 
-6433.9  -441.9   135.6   609.2  3451.6 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           903.6547   130.7060   6.914 1.11e-11 ***
time                    6.4115     0.2127  30.141  < 2e-16 ***
as.factor(n_MONTH)2   267.6610   178.8674   1.496   0.1350    
as.factor(n_MONTH)3  1134.4347   175.3922   6.468 1.93e-10 ***
as.factor(n_MONTH)4  1731.5257   177.4226   9.759  < 2e-16 ***
as.factor(n_MONTH)5  2400.8490   176.8029  13.579  < 2e-16 ***
as.factor(n_MONTH)6  2627.8905   179.2875  14.657  < 2e-16 ***
as.factor(n_MONTH)7  2223.6503   179.1448  12.413  < 2e-16 ***
as.factor(n_MONTH)8  2125.6356   180.6807  11.765  < 2e-16 ***
as.factor(n_MONTH)9  2032.1820   183.8048  11.056  < 2e-16 ***
as.factor(n_MONTH)10 1269.3402   184.3557   6.885 1.34e-11 ***
as.factor(n_MONTH)11  424.2023   209.1335   2.028   0.0429 *  
as.factor(n_MONTH)12 -330.8109   216.6608  -1.527   0.1273    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 974 on 662 degrees of freedom
Multiple R-squared:  0.7577,	Adjusted R-squared:  0.7533 
F-statistic: 172.5 on 12 and 662 DF,  p-value: < 2.2e-16
```

**Exhibit 4:** *Seasonal Dummies Model*

As seen in **Exhibit 4**, the p-value associated with time is less than 0.05, meaning that there exists a trend term in this model, and the p-values associated with each month are also less than 0.05, indicating that there is a seasonality in the model. Also, the multiple R-squared is 0.7577, which is satisfied. We will further investigate this in the next section by calculating its MAPE and correlatio

*Cyclical Trend:*

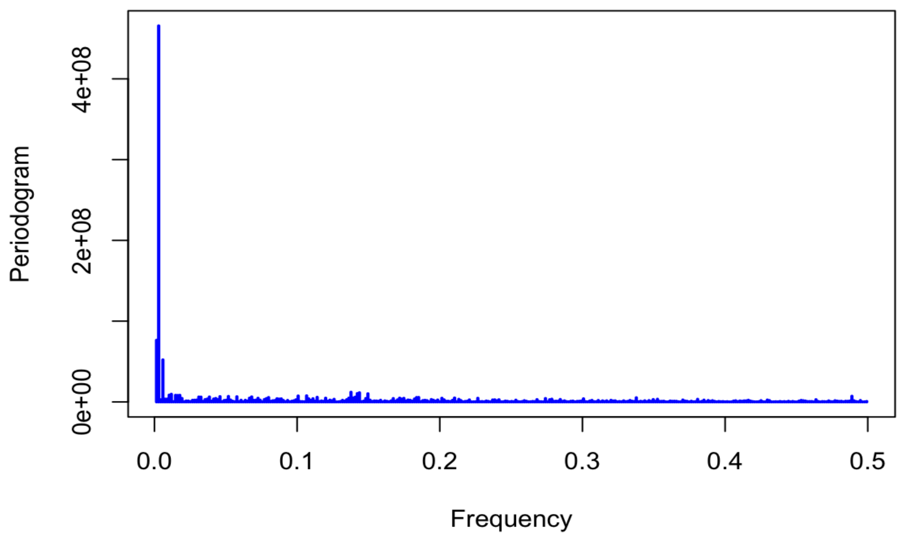



**Exhibit 5:** *Frequency of Cyclical Model*

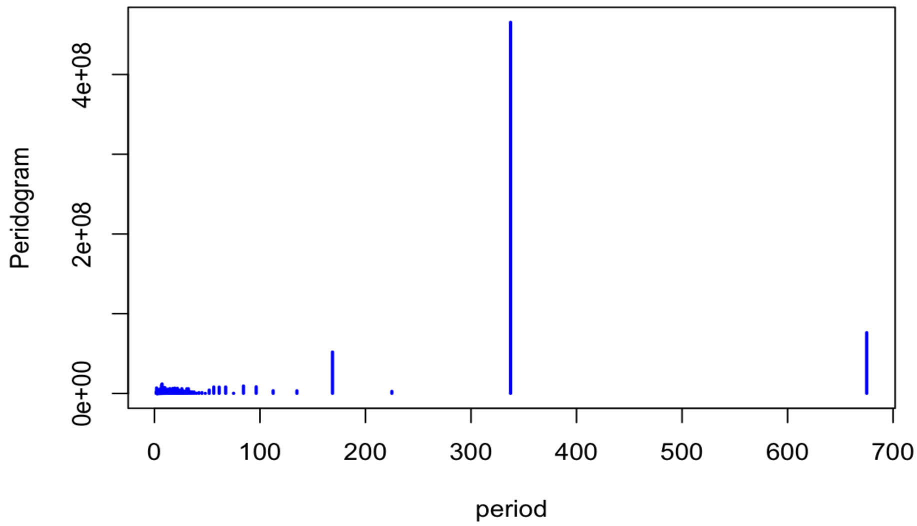



**Exhibit 6:** *Period of Cyclical Model*

| | period | frequency | amplitude |
|---|---|---|---|
| [1,] | 675.000000 | 0.001481481 | 7.634832e+07 |
| [2,] | 337.500000 | 0.002962963 | 4.656379e+08 |
| [3,] | 225.000000 | 0.004444444 | 2.751548e+06 |
| [4,] | 168.750000 | 0.005925926 | 5.207984e+07 |
| [5,] | 135.000000 | 0.007407407 | 3.502209e+06 |
| [6,] | 112.500000 | 0.008888889 | 3.565971e+06 |
| [7,] | 96.428571 | 0.010370370 | 8.474886e+06 |
| [8,] | 84.375000 | 0.011851852 | 9.505479e+06 |
| [9,] | 75.000000 | 0.013333333 | 3.525000e+05 |
| [10,] | 67.500000 | 0.014814815 | 8.228072e+06 |
| [11,] | 61.363636 | 0.016296296 | 8.073377e+06 |
| [12,] | 56.250000 | 0.017777778 | 8.226449e+06 |
| [13,] | 51.923077 | 0.019259259 | 4.368144e+06 |
| [14,] | 48.214286 | 0.020740741 | 1.922530e+05 |
| [15,] | 45.000000 | 0.022222222 | 1.242386e+06 |
| [16,] | 42.187500 | 0.023703704 | 1.190091e+06 |
| [17,] | 39.705882 | 0.025185185 | 4.141979e+05 |
| [18,] | 37.500000 | 0.026666667 | 2.011438e+06 |
| [19,] | 35.526316 | 0.028148148 | 2.166398e+06 |
| [20,] | 33.750000 | 0.029629630 | 2.328648e+06 |

**Exhibit 7:** *Periodogram of Cyclical Model*

As seen in **Exhibit 7**, we observe the highest amplitude, which is 465637855 at Period 337.5. To create sine and cosine pairs, we consider periods associated with the top 6 highest amplitude by sorting the amplitude in a descending order.

```
        period   frequency amplitude
[1,] 337.500000 0.002962963 465637855
[2,] 675.000000 0.001481481  76348321
[3,] 168.750000 0.005925926  52079837
[4,]   7.258065 0.137777778  12021327
[5,]   6.958763 0.143703704  11349454
[6,]   7.031250 0.142222222  10442903
```

**Exhibit 8:** *Top 6 Amplitudes*

```
Call:
lm(formula = n_CNT ~ time + cos1 + sin1 + cos2 + sin2 + cos3 +
    sin3 + cos4 + sin4 + cos5 + sin5 + cos6 + sin6)

Residuals:
    Min      1Q  Median      3Q     Max
-5799.3  -395.9   118.8   565.9  2799.4

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2197.1743   140.8044  15.604  < 2e-16 ***
time            6.8400     0.4033  16.961  < 2e-16 ***
cos1        -1059.0836    49.9313 -21.211  < 2e-16 ***
sin1         -567.6951    66.1043  -8.588  < 2e-16 ***
cos2          364.2059    49.9313   7.294 8.63e-13 ***
sin2          184.6996   100.0027   1.847 0.065202 .
cos3            6.1943    49.9313   0.124 0.901309
sin3         -423.2475    54.4251  -7.777 2.87e-14 ***
cos4          -99.9742    49.9313  -2.002 0.045667 *
sin4         -160.9466    49.9373  -3.223 0.001331 **
cos5         -122.7573    49.9313  -2.459 0.014206 *
sin5          134.5470    49.9366   2.694 0.007232 **
cos6         -173.6605    49.9313  -3.478 0.000538 ***
sin6          -25.4048    49.9367  -0.509 0.611104
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 917.3 on 661 degrees of freedom
Multiple R-squared:  0.7854,    Adjusted R-squared:  0.7812
F-statistic: 186.1 on 13 and 661 DF,  p-value: < 2.2e-16
```

**Exhibit 9:** *Cyclical Model with Sine and Cosine Pairs*

According to the top 6 amplitudes, the periods are provided in **Exhibit 8**, associating with the Harmonic 2, 1, 4, 93, 97, 96 in the overall periodogram. We then create the sine and cosine pairs based on these values to run a cyclical model. As seen in **Exhibit 9**, most p-values associated with each cosine and sine pair are less than 0.05, meaning that the coefficients are different than 0, and that the periods associated with these pairs contribute to the model.

## 2.2 Comparison of "candidate" models in terms of fit and hold-out sample.

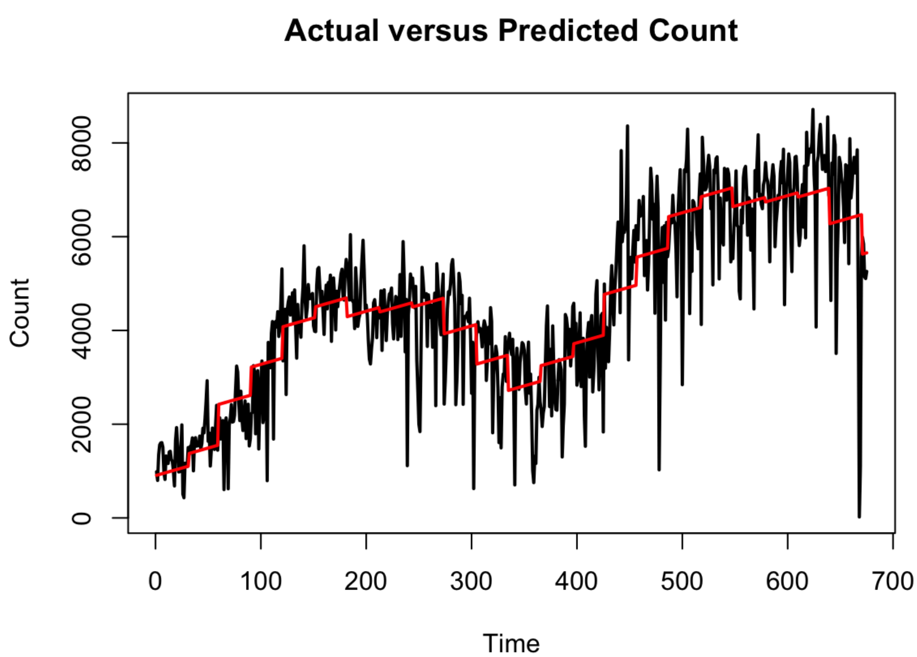



**Exhibit 10:** *Plot of Actual Versus Predicted of Seasonal Dummies Model*

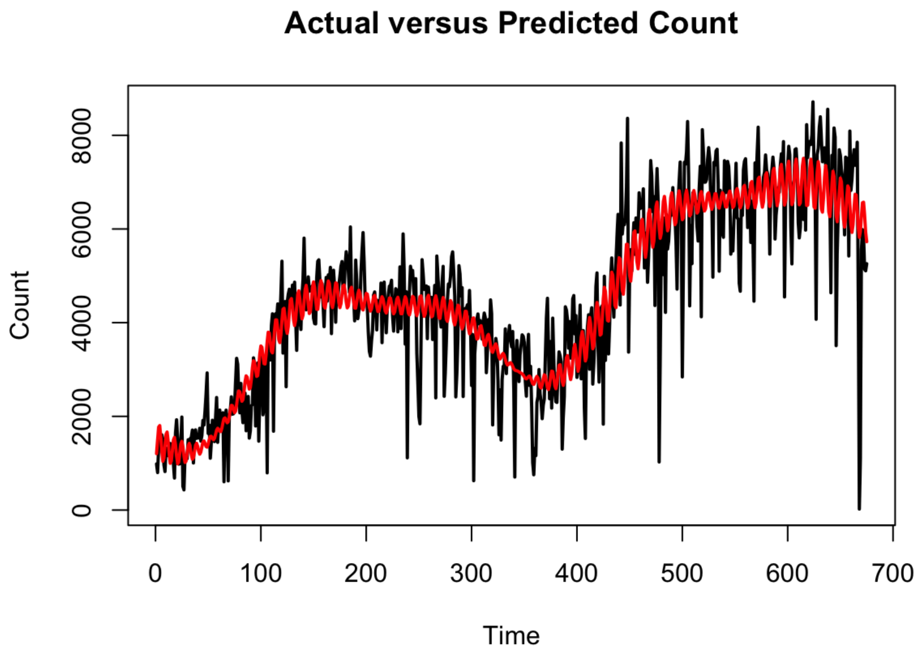



**Exhibit 11:** *Plot of Actual Versus Predicted of Cyclical Model*

Based on the plots, we can see that the plot of the actual and predicted counts of rental for the cyclical model fits relatively better than the seasonal dummies model. To consider statistically, you can refer to the following.

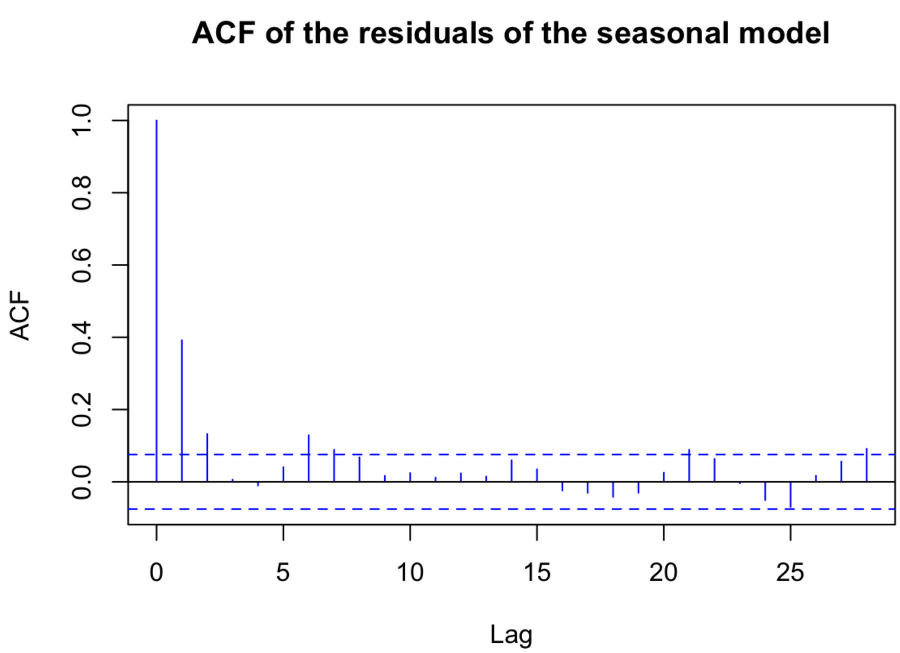
| | MAPE (Training) | MAPE (Out-of-sample) | Correlation |
|---|---|---|---|
| **Seasonal Dummies with Trends** | 66.37% | 74.27% | 0.8705 |
| **Cyclical Model** | 60.73% | 79.28% | 0.8862 |

The lower MAPE for the training dataset of the cyclical model, which is 60.73%, suggests that this model fits better compared to the seasonal dummies model, whose MAPE is 66.37%.

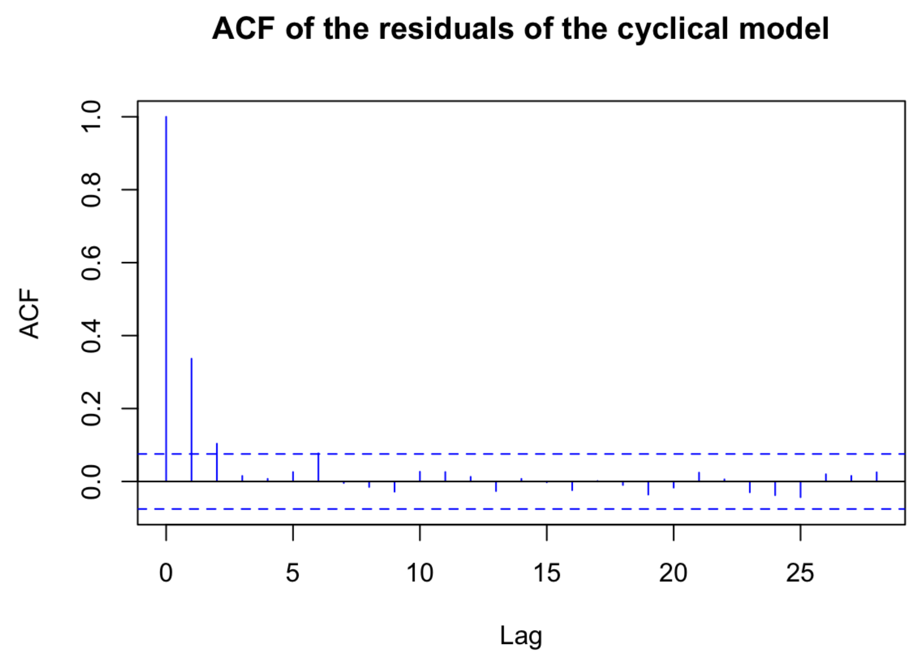
However, the MAPE for the out-of-sample dataset of the cyclical model, which is 79.28%, is higher than the MAPE of the seasonal dummies model, meaning that the cyclical model performs worse when applied to the out-of-sample dataset. This suggests that while the model captures certain cyclic patterns well in the training data, it might be overfitting to the training data or may not be capturing all relevant factors influencing the data. However, the correlation of the cyclical model is slightly higher than that of the seasonal dummies model, indicating that the cyclical model demonstrates a stronger relationship between predictions and observations. Therefore, we then look at the residuals of the model for further analysis.

## 2.3 Looking at residuals of the model(s).

**Seasonal Dummies with Trends**



**Cyclical Model**




The residuals of both models are not white noise, as evidenced by the Box-Pierce test p-value being lower than 0.05, but stationary after Lag 1.

## 3. Time Series Regression Models

### 3.1 Discussion of independent variables. Correlation analysis and scatter plots

**Correlation matrix:**

| | temp | hum | windspeed | cnt |
|---|---|---|---|---|
| temp | 1.0000000 | 0.1269629 | -0.1579441 | 0.6274940 |
| hum | 0.1269629 | 1.0000000 | -0.2484891 | -0.1006586 |
| windspeed | -0.1579441 | -0.2484891 | 1.0000000 | -0.2345450 |
| cnt | 0.6274940 | -0.1006586 | -0.2345450 | 1.0000000 |

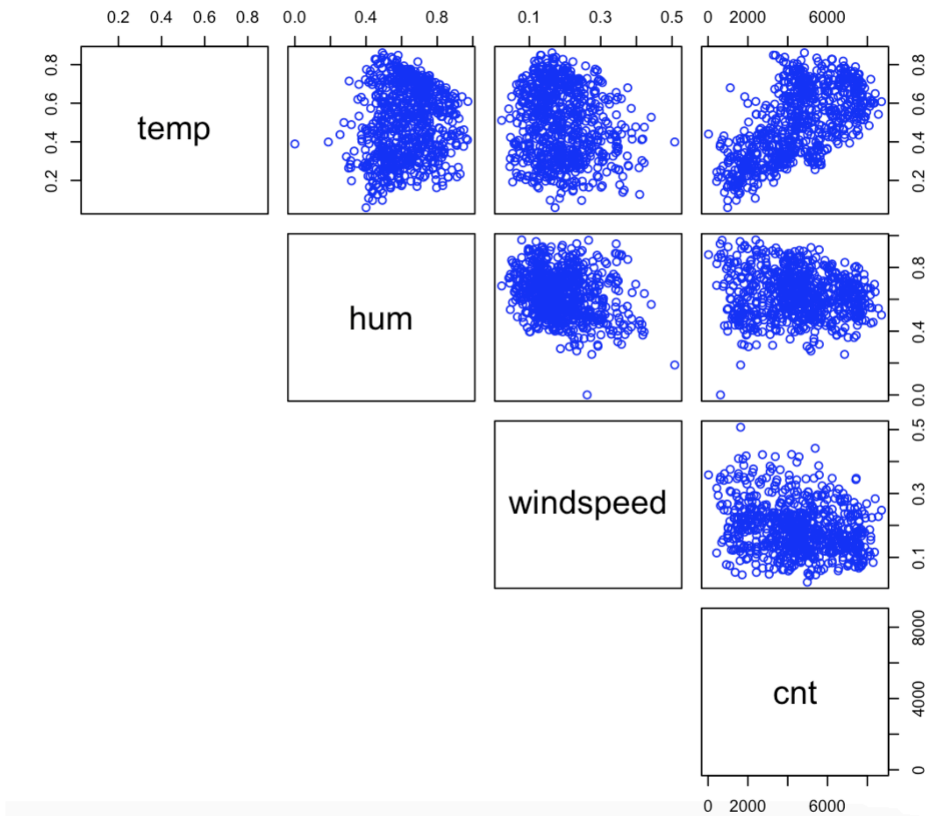
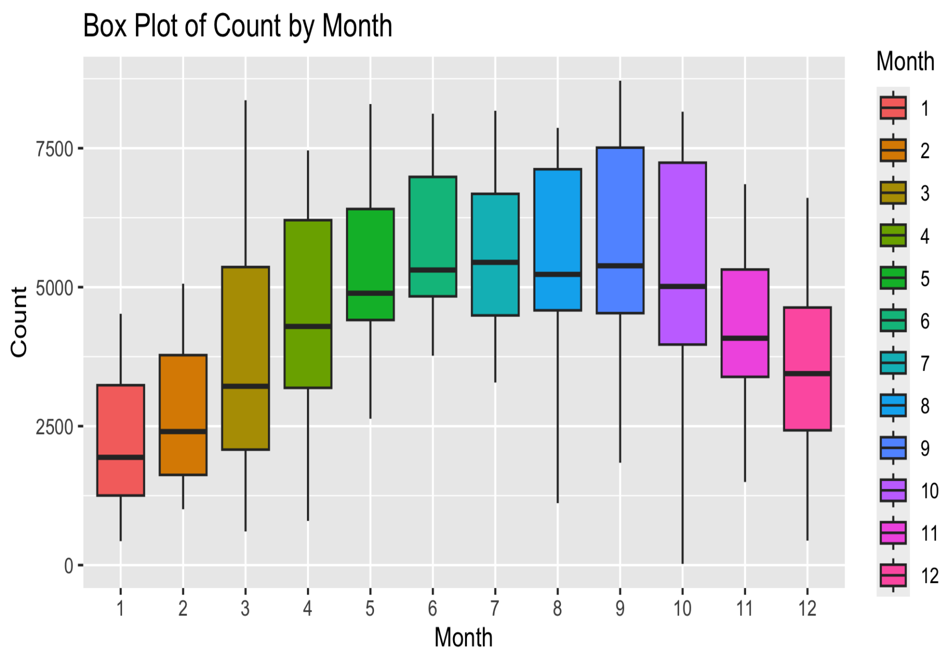
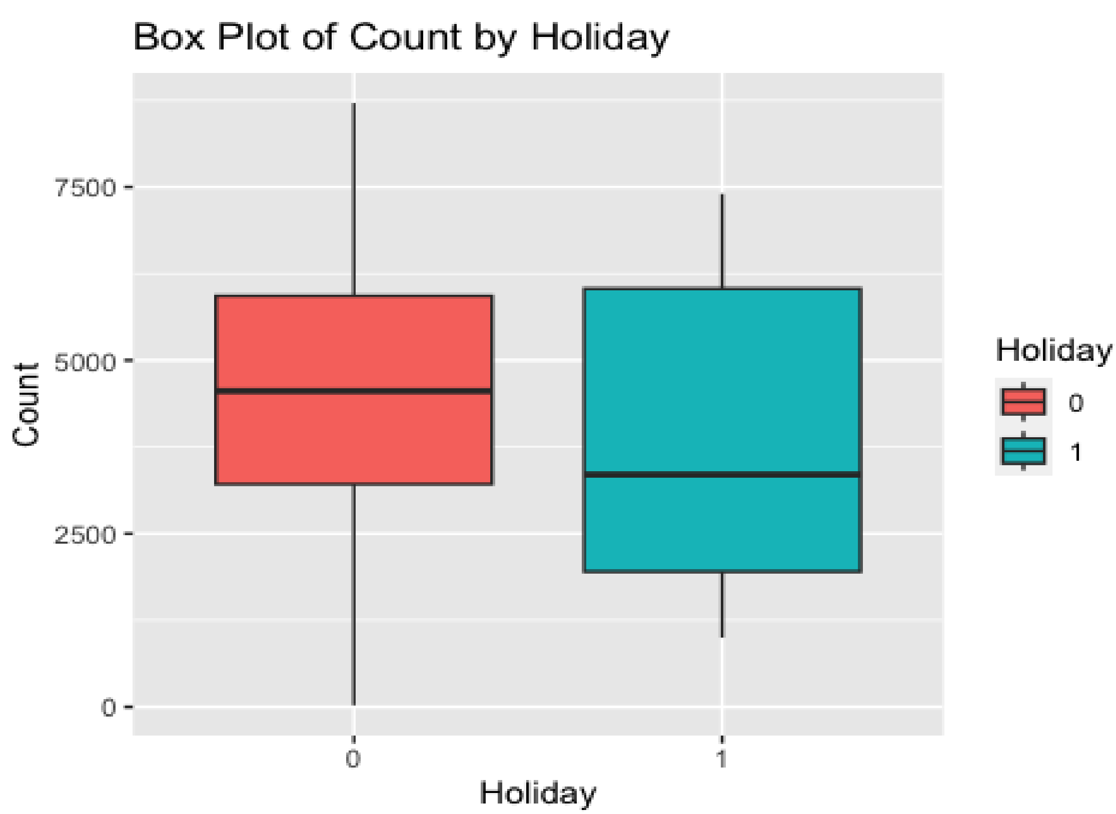



**Exhibit 12:** *Correlation Analysis and Scatter Plots*

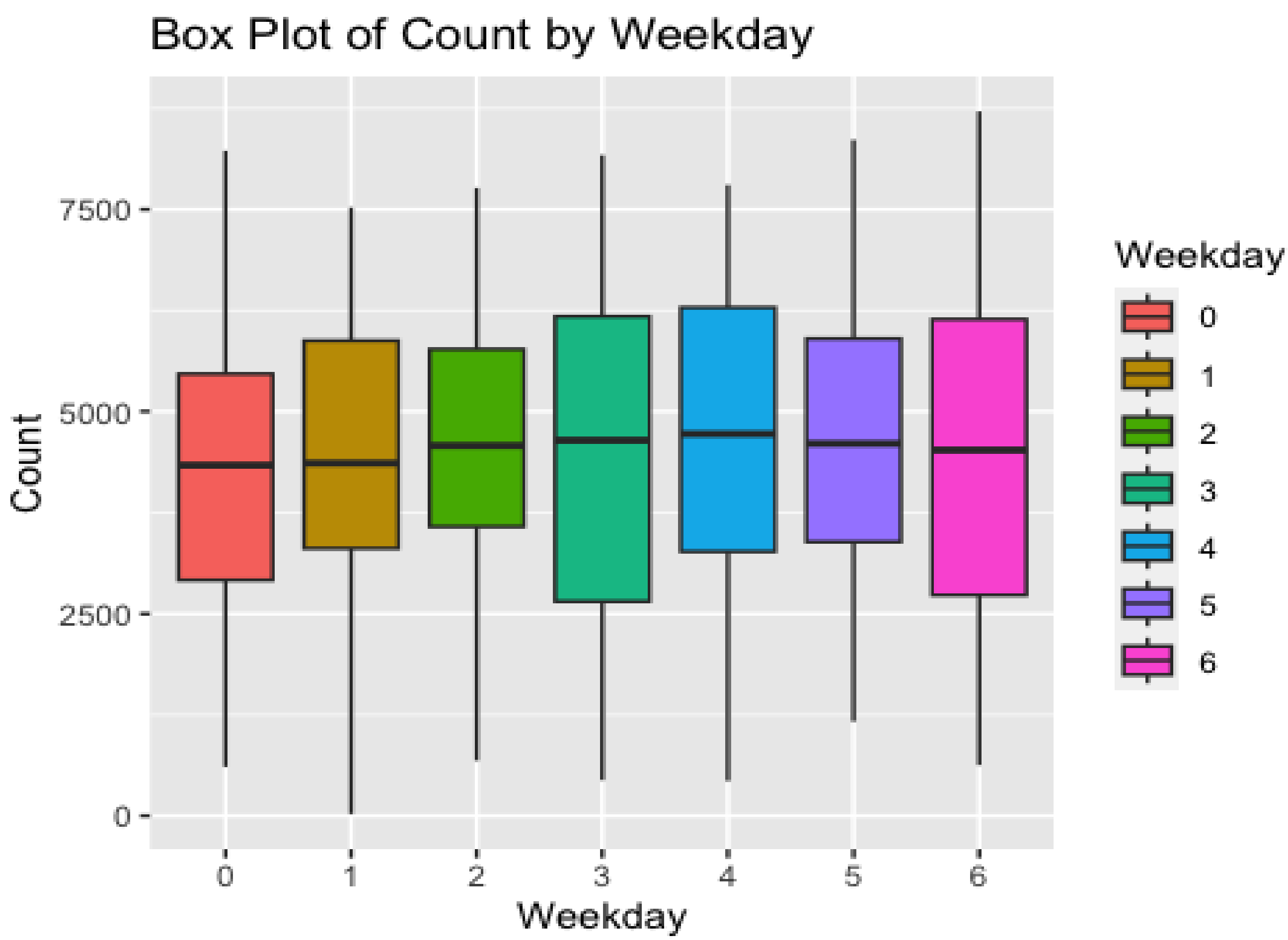
In the correlation matrix, we checked how each variable is positively and negatively correlated with each other, as seen in **Exhibit 12**. We have focused primarily on the relationship between count and other variables to determine how their unique trends contribute. Comparing month, weekday, and temperature showed a positive correlation. Since the month, holiday, and weather situation are based on the categorical data set, the following regression models will be implemented with the following box plots.



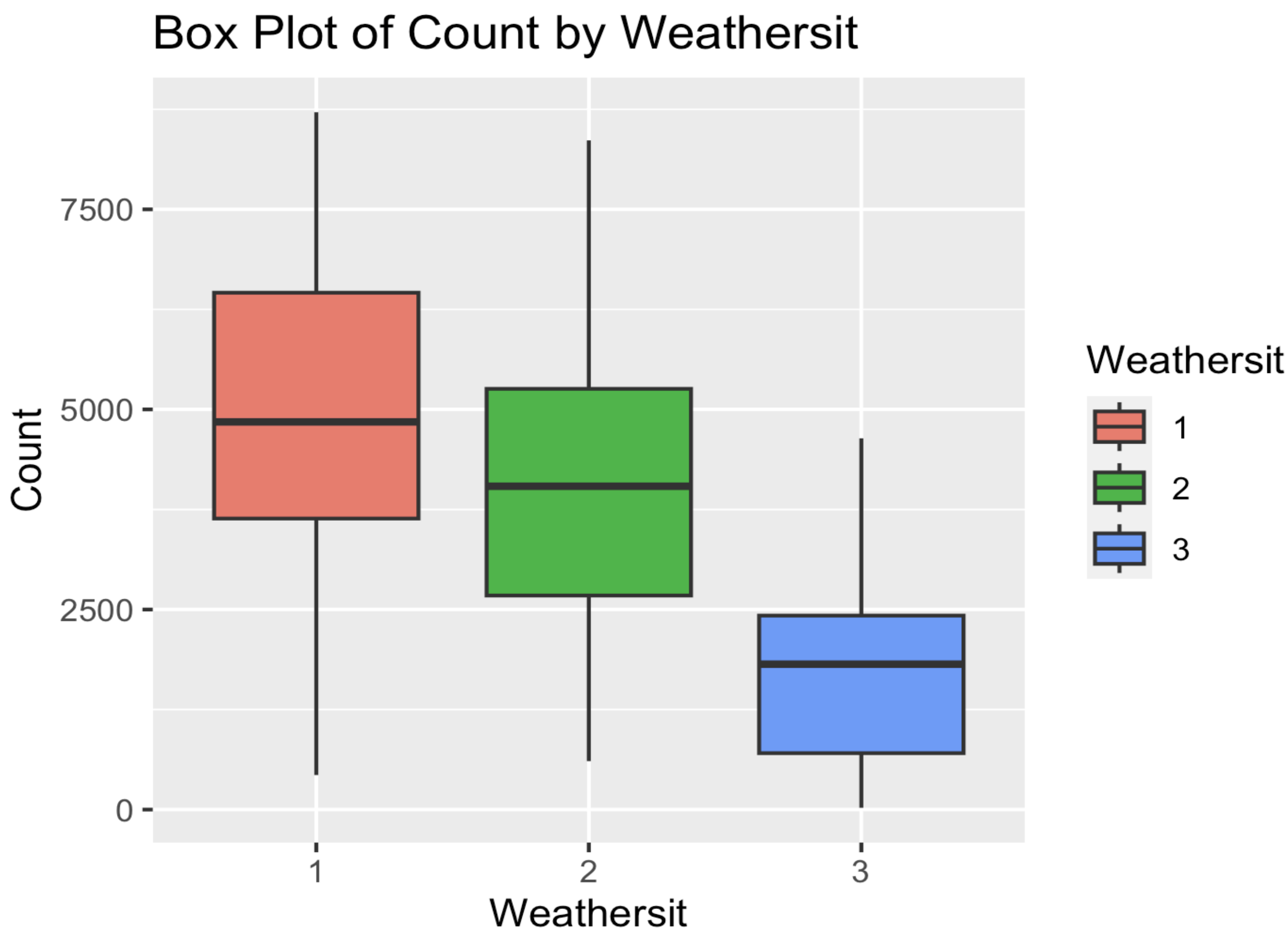

Our data analysis has revealed a distinct relationship between Bike rentals and month. Specifically, we've found that March, April, September, and October exhibit a significant degree of variability, indicating potential seasonal patterns in bike rentals.



The Plot shows the relationship between variable bike counts and Holidays. By checking the Box plot results we could check that customers are more willing to pick up and drop them not during the holiday seasons.




The plot shows the relationship between the weekdays and bike rentals. From the box plots and data set from 1/1/2011, we see that there is not much difference between the weekdays for the bike rental by checking their counts which all show a range between 2,500 to 5,000.




The box plot shows the relationship between bike rentals and weather situations. We could check from the graph that when the weather improves (1: excellent, 2: worse, 3: worst), customers are willing to rent more bicycles; however, when the weather gets worse, the situation goes vice versa.

## 3.2 Comparison of "candidate" models in terms of fit and hold-out sample

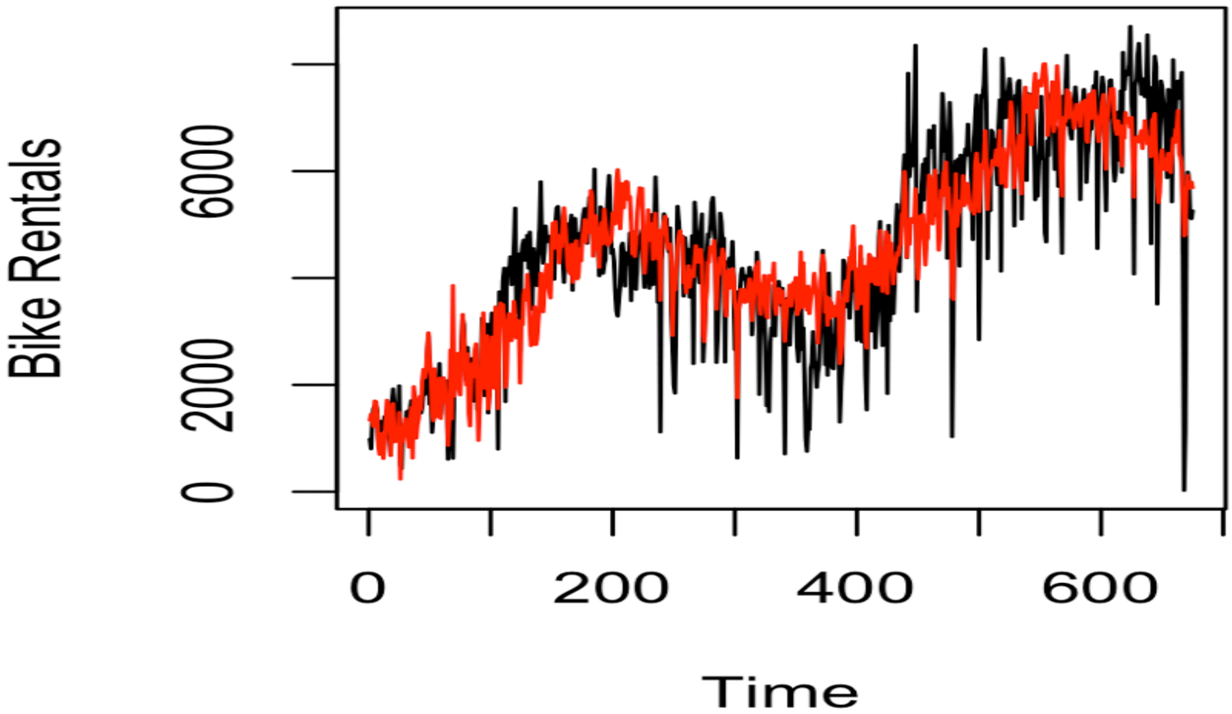



**Exhibit 13:** *Plot of Actual Versus Predicted of Regression Model*

**Exhibit 13** depicts the actual and the predicted value of the regression model.

```
lm(formula = n_cnt ~ time + n_temp + n_hum + n_wind, data = pro)

Residuals:
    Min      1Q  Median      3Q     Max 
-4764.8  -505.6    69.3   563.4  2796.8 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2349.8833   231.6468  10.144  < 2e-16 ***
time            5.7306     0.1916  29.916  < 2e-16 ***
n_temp       5109.9788   207.0025  24.686  < 2e-16 ***
n_hum       -2782.1197   255.0215 -10.909  < 2e-16 ***
n_wind      -3325.7920   484.1633  -6.869 1.48e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 915.3 on 670 degrees of freedom
Multiple R-squared:  0.7834,	Adjusted R-squared:  0.7821 
F-statistic: 605.9 on 4 and 670 DF,  p-value: < 2.2e-16
```

**Exhibit 14:** *Regression Model*

As seen in **Exhibit 14**, all the coefficients of the regression model are significant, since p value for temperature, humidity and wind speed is less than .05.

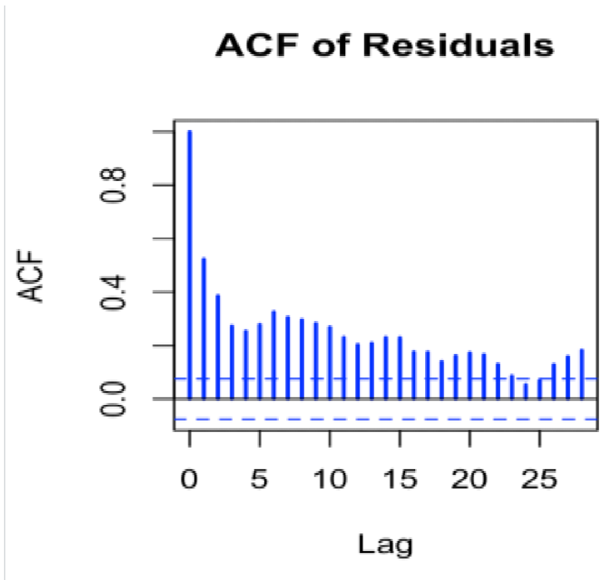
| | MAPE (Train) | MAPE (Out-of-sample) | Correlation |
|---|---|---|---|
| **Time Series Regression model** | **52.83569%** | **72.48018%** | **0.8843642** |

From the regression model we found that the MAPE for in sample data is 53% which is less than the holdout sample of 72%.The correlation coefficient in Actual and Predicted value is 88% which suggests a good correlation.

RMSE of the Time Series Regression Model :911.9317

## 3.3 Looking at residuals of the model(s).

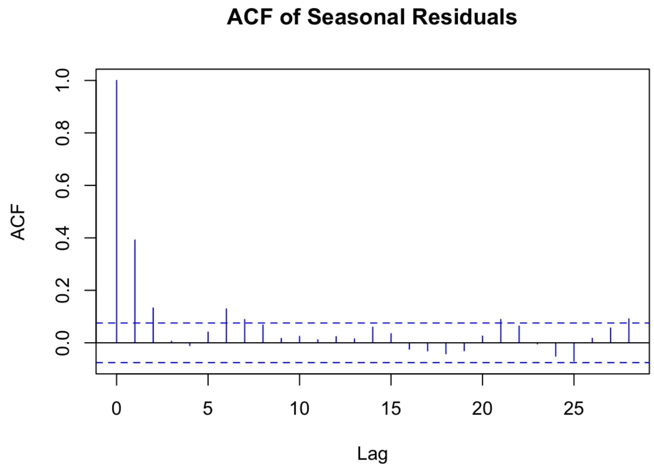
| Residuals of the Regression model | Box-Pierce p-value | White Noise or not |
|---|---|---|

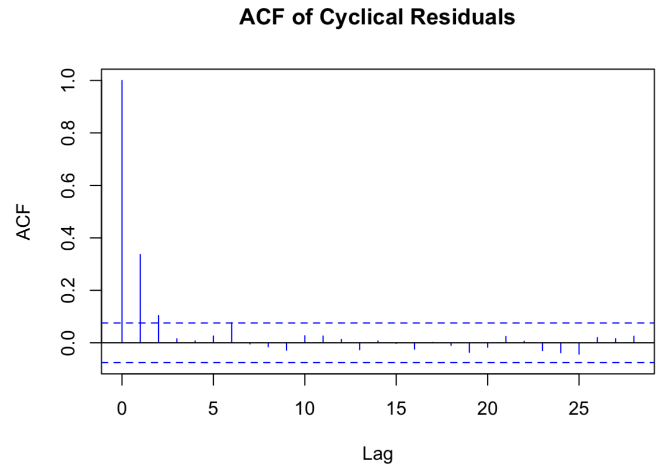
| | | |
|---|---|---|
|  | <2.2e-16 | **Reject the null hypothesis → Not WN** |

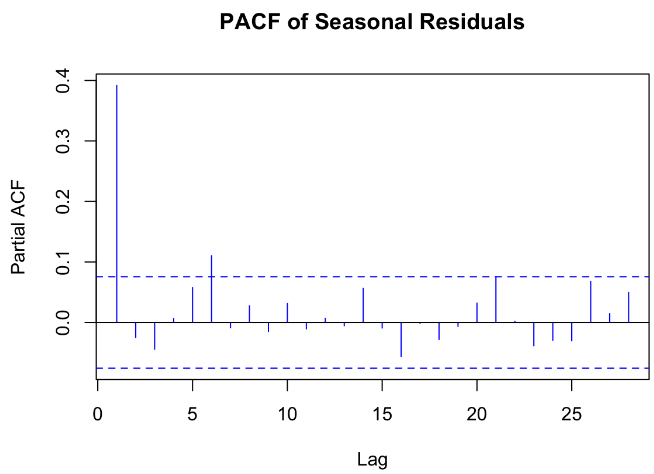
**4. Stochastic Time Series Models**

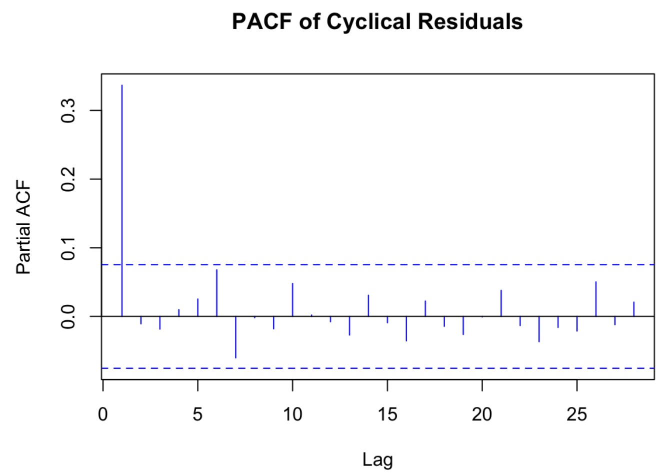
**4.1 Analysis and modeling of deterministic time series model residuals**

| Seasonal Dummies Model | Cyclical Model |
|---|---|




As seen in the above graphs, the ACF decays quickly, and PACF chops off to 0 after lag 1. Therefore, we will consider the **AR(1)**

**AR(1)**

```
Series: cr
ARIMA(1,0,0) with non-zero mean

Coefficients:
         ar1     mean
      0.3363  -0.0753
s.e.  0.0362  49.5270

sigma^2 = 732653:  log likelihood = -5514.59
AIC=11035.17   AICc=11035.21   BIC=11048.72
```

process.

```
Series: resforfinal1
ARIMA(1,0,0) with non-zero mean

Coefficients:
         ar1     mean
      0.3913  -0.0958
s.e.  0.0354  56.0629

sigma^2 = 789872:  log likelihood = -5539.99
AIC=11085.98   AICc=11086.01   BIC=11099.52
```

```
        Box-Pierce test

data:  fit_arar$resid
X-squared = 19.809, df = 20, p-value = 0.47
```

Since the ratio of phi1 and s.e., which is 0.3913/0.0354, is greater than 2, we can conclude that the coefficient at lag 1 is statistically different from 0.

Moreover, according to the Box-Pierce Test, the p-value, 0.47, is greater than 0.05, meaning that the residuals from this AR model are white noise after we model the series as AR(1).

```
        Box-Pierce test

data:  cc$resid
X-squared = 11.568, df = 20, p-value = 0.9301
```

Since the ratio of phi1 and s.e., which is 0.3363/0.0362, is greater than 2, we can conclude that the coefficient at lag 1 is statistically different from 0.

Moreover, according to the Box-Pierce Test, the p-value, 0.9301, is greater than 0.05, meaning that the residuals from this AR model are white noise after we model the series as AR(1).

**MA(2)**

```
Series: res3
ARIMA(0,0,2) with non-zero mean

Coefficients:
         ma1     ma2     mean
      0.3376  0.1040  -0.0973
s.e.  0.0380  0.0376  47.3936

sigma^2 = 733652:  log likelihood = -5514.54
AIC=11037.09   AICc=11037.15   BIC=11055.15
```

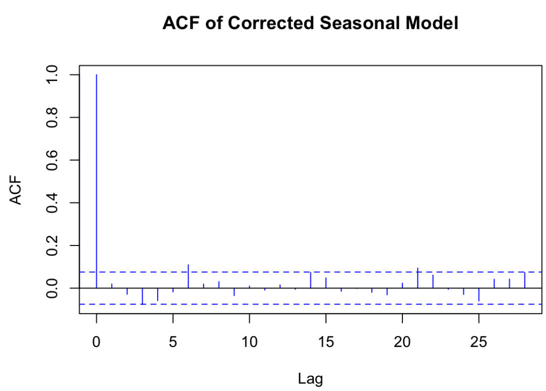
```
        Box-Pierce test

data:  res_armi
X-squared = 0.0046962, df = 1, p-value = 0.9454
```

Also, we can use the MA(2) model, since the t-stats are 0.3376/0.0380 and 0.1040/0.0376, which are greater than 2, we can conclude that the t-stat is significant.

Additionally, according to the Box-Pierce Test, the p-value, 0.9454, is greater than 0.05, meaning that the residuals from this MA model are white noise after we model the series as MA(2).

**<u>Corrected Model:</u>**

**<u>ARIMA(2,0,1)</u>**



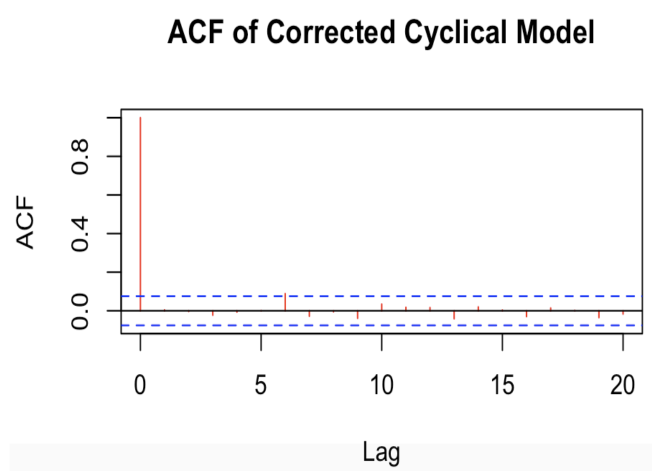

The ACF falls within 2 standard error bounds, meaning that the series is white noise.

```
Series: n_CNT
Regression with ARIMA(2,0,1) errors

Coefficients:
         ar1      ar2      ma1  intercept    time
      1.3176  -0.3238  -0.9000  2026.4730  7.0990  -18.0626
s.e.  0.0430   0.0421   0.0188   833.9941  2.0245   45.2946

sigma^2 = 826387:  log likelihood = -5553.85
AIC=11121.7   AICc=11121.87   BIC=11153.3
```

After trying the ARIMA process, we find that ARIMA(2,0,1) is an appropriate model. The t-stats are all greater than 2, indicating the

**<u>Corrected Model:</u>**

**<u>ARIMA(1,0,0)</u>**




The ACF falls within 2 standard error bounds, meaning that the series is white noise.

```
Series: n_usage
Regression with ARIMA(1,0,0) errors

Coefficients:
         ar1  intercept    time      cos1      sin1        cos2       sin2     cos3
      0.3364  2200.0345  6.8301  363.2670  182.5726  -1060.0224  -568.7695   5.2559
s.e.  0.0362   195.6294  0.5599   70.0012  139.2343     69.9943    92.3593  69.9667
           sin3       cos4       sin4       cos5      sin5       cos6      sin6
      -423.8064  -100.5792  -161.5714  -174.2441  -26.0417  -123.3338  133.9063
s.e.    76.2405    56.5054    56.5532    55.9129   55.9578    55.7166   55.7606

sigma^2 = 747083:  log likelihood = -5514.59
AIC=11061.17   AICc=11062   BIC=11133.41
```

After trying the ARIMA process, we find that ARIMA(1,0,0) is an appropriate model. The t-stats are all greater than 2, indicating the significance, thus validating the use of this model.

```
        Box-Pierce test

data:  sfit_corrected$resid
X-squared = 23.899, df = 20, p-value = 0.2468
```

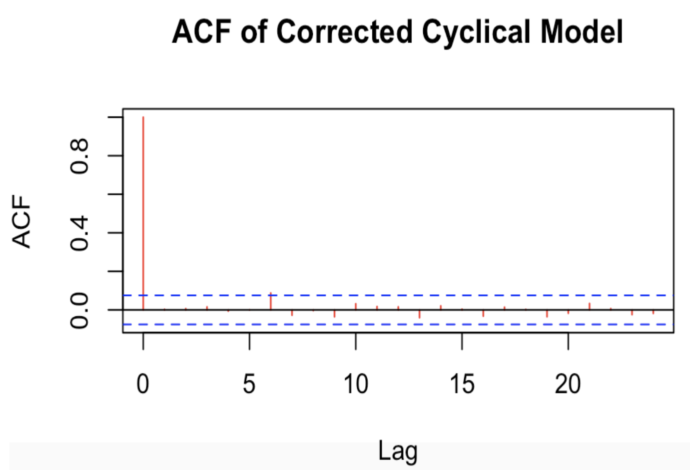
Additionally, we assess the residuals using the Box-Pierce test. The p-value is greater than 0.05, leading to the conclusion that they present white noise.

significance, thus validating the use of this model.

```
        Box-Pierce test

data:  cfit_corrected$resid
X-squared = 11.56, df = 20, p-value = 0.9304
```

Additionally, we assess the residuals using the Box-Pierce test. The p-value is greater than 0.05, leading to the conclusion that they present white noise.

**ARIMA(0,0,2)**




The ACF falls within 2 standard error bounds, meaning that the series is white noise.

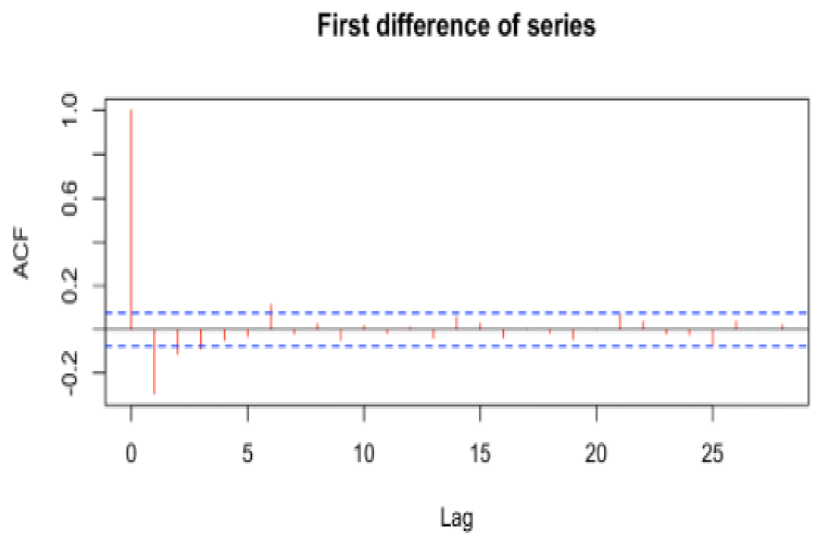
```
Series: n_usage
Regression with ARIMA(0,0,2) errors

Coefficients:
         ma1     ma2  intercept    time      cos1      sin1        cos2       sin2
      0.3376  0.1040  2198.1207  6.8363  363.5864  183.8950  -1059.7032  -568.1047
s.e.  0.0380  0.0376   187.6344  0.5371   66.9885  133.4738     66.9852    88.4625
        cos3       sin3       cos4       sin4       cos5      sin5       cos6      sin6
      5.5750  -423.4670  -100.4378  -161.3747  -174.1135  -25.8431  -123.2067  134.1055
s.e.  66.9721    72.9832    58.3358    58.3823    57.8268   57.8706    57.6551   57.6982

sigma^2 = 748124:  log likelihood = -5514.54
AIC=11063.09   AICc=11064.02   BIC=11139.84
```

| | |
|---|---|
| | After trying the ARIMA process, we find that ARIMA(0,0,2) is an appropriate model. The t-stats are all greater than 2, indicating the significance, thus validating the use of this model.<br><br>`Box-Pierce test`<br><br>`data:  cfit_corrected$resid`<br>`X-squared = 11.14, df = 20, p-value = 0.9425`<br><br>Additionally, we assess the residuals using the Box-Pierce test. The p-value is greater than 0.05, leading to the conclusion that they present white noise. |

**4.2 Analysis and modeling of regression model residuals**

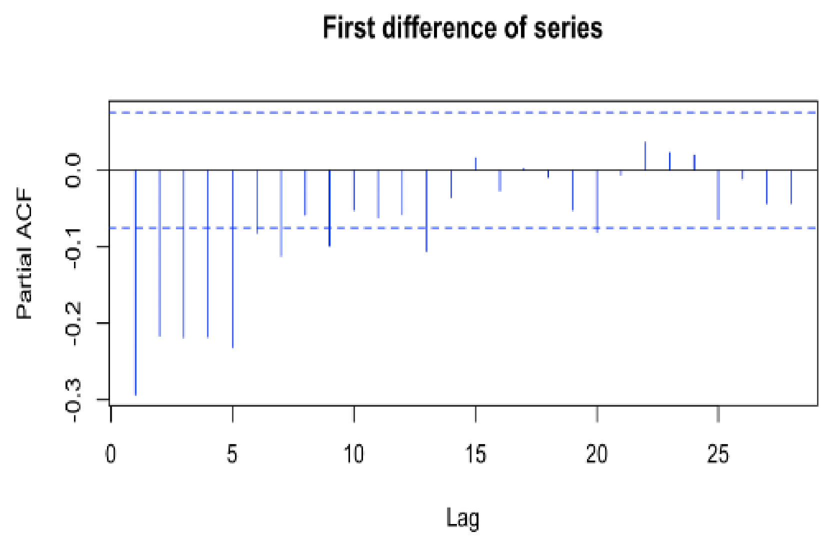
| | |
|---|---|
| Model of Regression Residual | Since the series obtained from the Regression model is nonstationary, we differentiate the series. |

The difference series ACF is chopped after lag 2.




PACF is decaying quickly.

We fit an MA(2) with corrected residuals.




Regression coefficients of the corrected Model

Fitting the model with ARIMA(0,1,2) we found that the absolute value of coefficients of the model are greater than 1.96.

**Regression Coefficients with Residuals**

```
Regression with ARIMA(0,1,2) errors

Coefficients:
          ma1      ma2              n_temp       n_hum      n_wind
      -0.6154  -0.2044  5.2233  4850.1841  -3306.1891  -3181.0677
s.e.   0.0358   0.0350  5.3138   480.7439    239.9195    413.9787

sigma^2 = 580735:  log likelihood = -5426.53
AIC=10867.06   AICc=10867.23   BIC=10898.65
```

Hence, all the coefficients in the model are significant.

We further found the accuracy of the model for In sample and out of Sample.

**For holdout  Sample**

Correlation Coefficient = 0.78642

RMSE: 1002.372

MAPE :

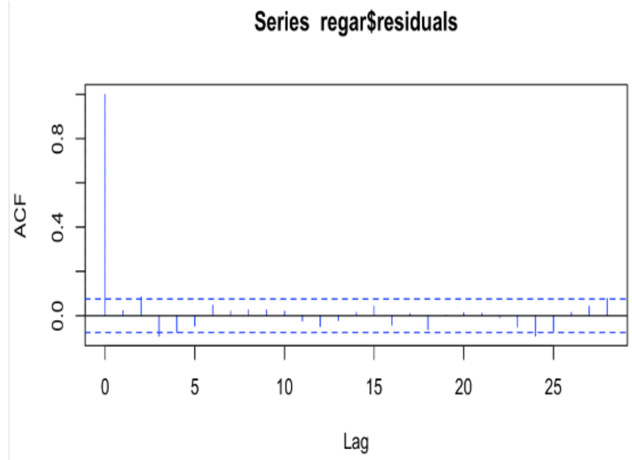
29.55824%

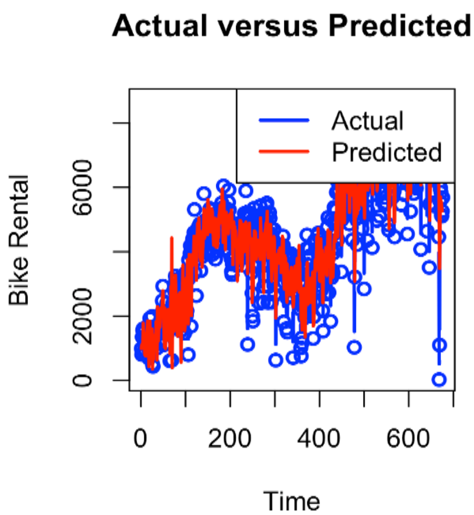
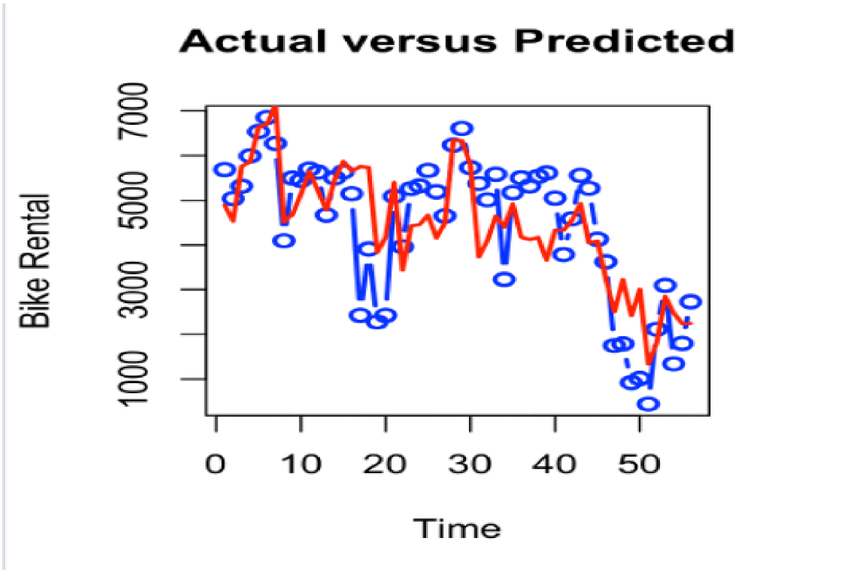
**For In Sample**

Correlation Coefficient =0.7864226

RMSE: 758.0977

MAPE : 49.49%



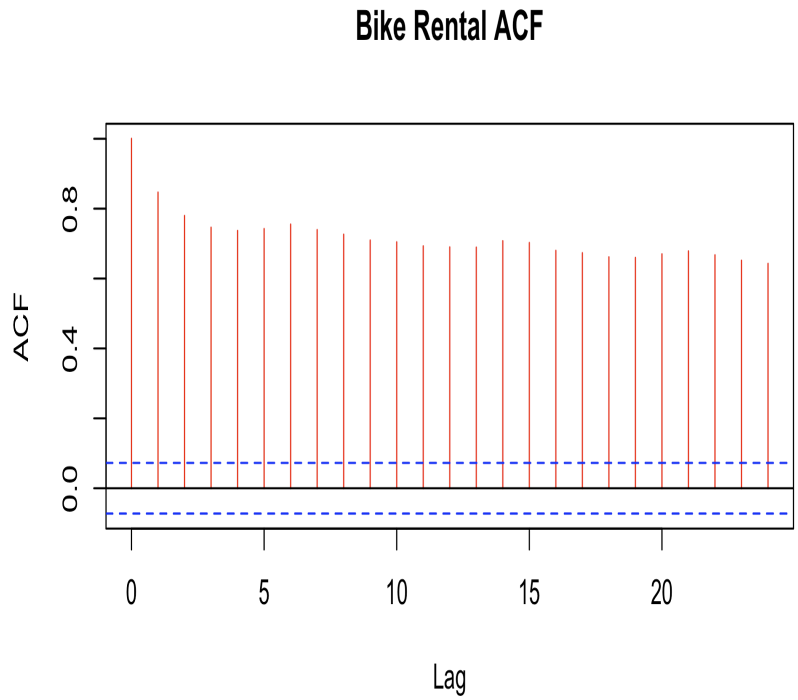
<table>
<tr>
<td>From Box Pierce test,<br>p=.08408 and ACF is also<br>between 2se bounds hence it’s<br>a White Noise Series</td>
<td>Box-Pierce test<br><br>data: regar$residuals<br>X-squared = 33.857, df = 24, p-value = 0.08724</td>
</tr>
<tr>
<td>Actual Vs predicted for In sample<br></td>
<td>Actual Vs Predicted for Holdout Sample<br></td>
</tr>
</table>

## 4.3 ARIMA models (for the variable of interest)

Model ARIMA(1,1,1)

Bike Rental residuals as shown in the plot are decaying slowly so we differenced the series and further modeled it with ARIMA(1,1,1).

Since in the section 4.2 value for Box plot p value is near to .05 so we tried fitting ARIMA(1,1,1) to the Bike rental series.



The Model Box plot is now .3557 and ACF is between 2 standard error bounds.

**In sample:**

As the p value is greater than the .05 so we can say that we cannot reject the null hypothesis and the series is a white Noise series.

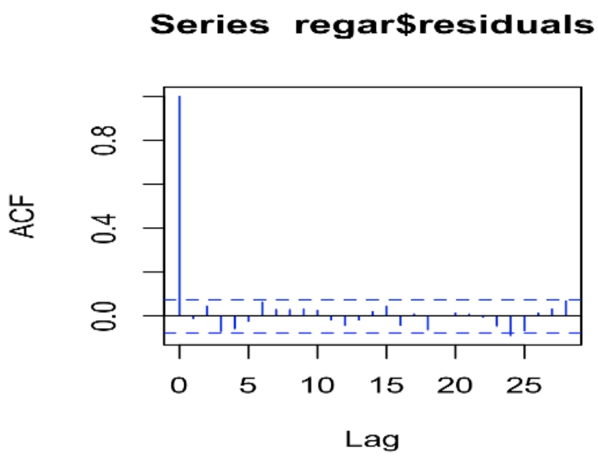
MAPE: 48.91067%

RMSE : 753.9765

Correlation Coefficient:

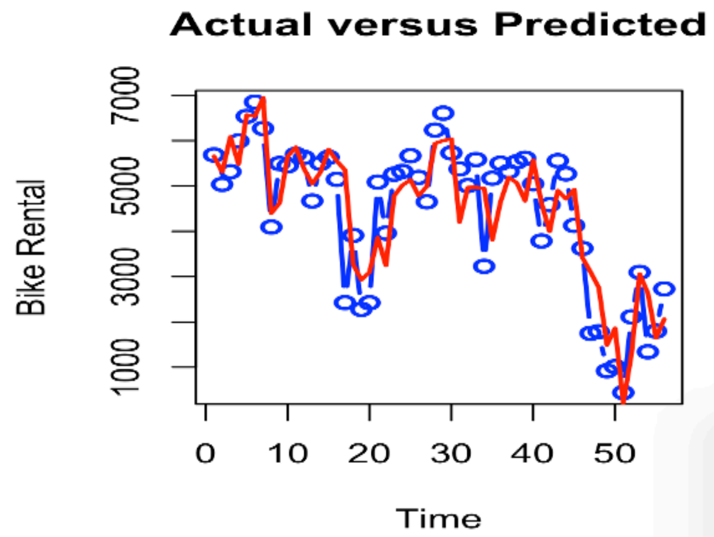
0.9233345

```
Regression with ARIMA(1,1,1) errors

Coefficients:
         ar1      ma1              n_temp       n_hum      n_wind
      0.3116  -0.8912  5.3554  4961.0642  -3330.6928  -3124.6883
s.e.  0.0477   0.0235  4.6501   485.6595    241.1232    410.4867

sigma^2 = 574438:  log likelihood = -5422.87
AIC=10859.75   AICc=10859.92   BIC=10891.34
```




Box-Pierce test

data: regar$residuals
X-squared = 25.911, df = 24, p-value = 0.3577

| **Hold Out SAMPLE:**<br><br>MAPE:20.436%<br>RMSE: 778.9141<br>Correlation Coefficient:<br>0.8774671 |  |
| --- | --- |



**5. Conclusion summary of your findings and comparison of deterministic and stochastic model performance based on the hold-out sample.**

In conclusion, our study provides insights into the deterministic and regression model that explain the forecast of rental counts.

In terms of the deterministic model, we conclude that the seasonal dummies model demonstrates a significant trend term and seasonality, as analyzed by low p-values associated with time and each month, and that the cyclical model captures cyclical patterns well, as evidenced by low p-values associated with time and sine and cosine pairs. However, the higher out-of-sample MAPE for the cyclical model suggests potential overfitting or inadequacy in capturing all relevant factors. Therefore, analyzing the residuals, we model the seasonal dummies model using ARIMA(2,0,1) process, and the cyclical model using AR(1) or MA(2) process, as the t-stats imply the significance of the coefficient associated with lags, and both models' residuals are considered as white noise, as analyzed by Box-Pierce test p-values greater than 0.05. This means that the models adequately capture the underlying structure of the data, providing a foundation for further refinement and application in rental count forecasting.

As for the regression model, we difference the series and conclude that the model with 3 independent variables, temperature, humidity, and wind speed, From the Regression models we found the MAPE to be 52% for in sample and 72.48018% for out sample but we fitted regression model with residuals with ARIMA(0,1,2) we found that the MAPE for Insample is 49.49% and for hold out sample is 29.5% .Thus we can say that by fitting the regression model with residuals the MAPE is reduced which means the model is improved by fitting the regression model with residuals. Further we see the RMSE of the regression model is 911.91 while the RMSE for corrected regression model is 753.97 for in sample and 1005 for hold out may be due to overfitting. Due to this overfitting we also have implemented ARIMA(1,1,1) which shows Insample RMSE is 753.9765 and MAPE of 48.91067% and also the holdout sample Accuracy is also improved, RMSE is 778.9141 and MAPE of 20.49%. This suggests that after taking the difference of the residuals and also fitting the model with residuals the performance of the model is increased. ARIMA(1,1,1) error is appropriate, as evidenced by the significant t-stats and residuals falling within the standard error bounds and MAPE and RMSE values.

Overall, our findings suggest that refining regression models through residual fitting and incorporating ARIMA processes can enhance forecasting accuracy, mitigating potential overfitting issues and improving model performance for rental count predictions.