# PROJECT REPORT FOR REGRESSION ANALYSIS OF BIKE RENTAL DATA

**SRINIDHI ALWALA**

# <u>Case Study: Bike Rental Data set</u>

1. Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousand bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

**<u>Motivation:</u>**

Covid-19 pandemic has led to a dramatic loss of human life, transport, working and living conditions. Bike rental demand has been reduced leading to decrease in business drastically. A bike rental Company wanted to restart its business once pandemic gets settled. They wanted to research using previous dataset and identify factors effecting Bike Rental Business.

**<u>Objective:</u>**

Identifying variables which are significant in predicting demand of Bike Rentals using Regression Analysis.

Regression Analysis is a very efficient method in identifying factors having impact on particular topic. This process allows you to confidently determine which factors matter most, which factors can be ignored, and how these factors influence each other which helps in taking better business decisions in predicting the future demand.

In this case study we are going to find the relationship between variables in data set and identify important factors which impact Bike Rentals increasing their demand

2. This data set has been taken from the following URL:

*https://www.kaggle.com/c/bike-sharing-demand/overview/description*

**<u>3. Data Description:</u>**

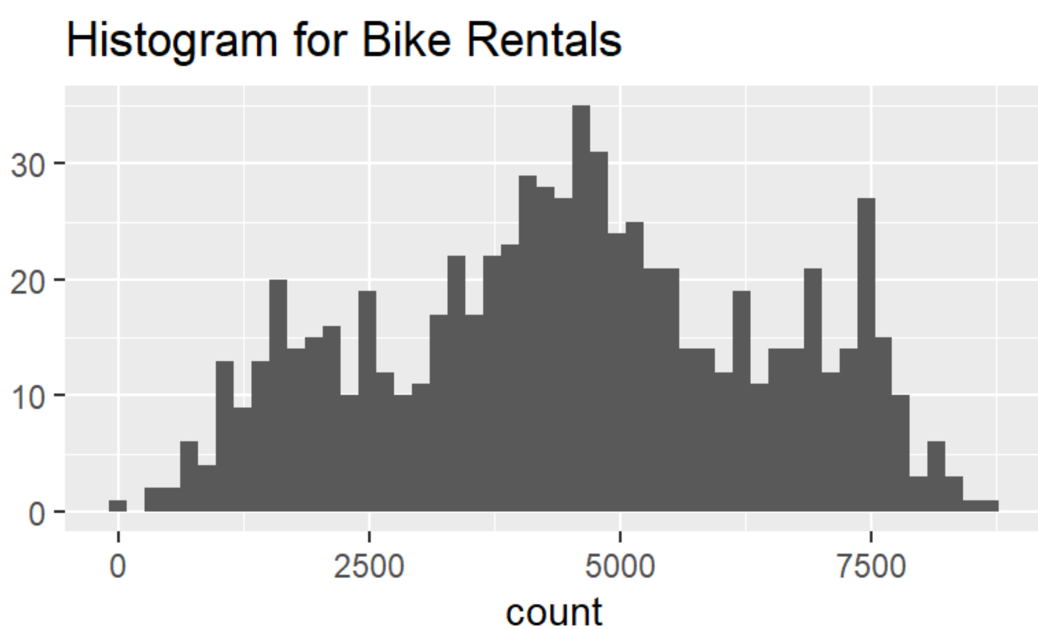
**<u>Table1:</u>**

| **Name** | **Type** | **Details** |
|---|---|---|
| **Instant** | **Categorial** | **Record index** |
| **dteday** | **Categorial** | **Date** |
| **season** | **Categorial** | **Season (1: spring,2: summer,3: fall,4: winter)** |
| **yr** | **Categorial** | **Year (0:2018,1:2019)** |
| **mnth** | **Categorial** | **Month (1 to 12)** |
| **holiday** | **Categorial** | **Holiday or no holiday** |
| **weekday** | **Categorial** | **Day of the week** |

| | | |
|---|---|---|
| **workingday** | **Categorial** | **If day is neither weekend nor holiday is 1, else 0** |
| **weathersit** | **Categorial** | **1: clear, few clouds, partly cloudy**<br>**2: Mist+cloudy, mist+broken clouds, mist few clouds, mist**<br>**3: Light snow, light rain+thunderstorm+scattered clouds**<br>**4: heavy rain+ice pallets+thunderstorm+mist, snow+fog** |
| **temp** | **Continuous** | **Temperature in Celsius** |
| **atemp** | **Continuous** | **Feeling temperature in Celsius** |
| **hum** | **Continuous** | **Humidity** |
| **windspeed** | **Continuous** | **Wind speed** |
| **casual** | **Continuous** | **Count of casual users** |
| **registered** | **Continuous** | **Count of registered users** |
| **cnt** | **Continuous (Response Variable)** | **Count of total rental bikes including both casual and registered** |

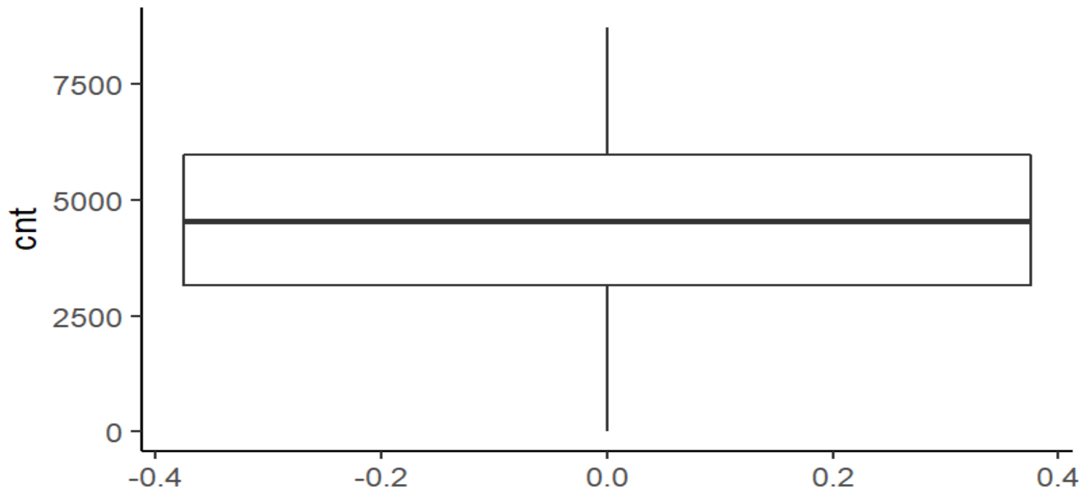
## 4. Descriptive Analysis on Response Variable

**Fig1: Histogram**




From the above histogram we can conclude that it is approximately following normal distribution pattern.

**Fig2: Box plot:**




We can conclude that there are no outliers.

**Fig 3: Mean, Median, Mode:**

```
     nbr.val     nbr.null       nbr.na          min
7.300000e+02 0.000000e+00 0.000000e+00 2.200000e+01
         max        range          sum       median
8.714000e+03 8.692000e+03 3.290845e+06 4.548500e+03
        mean      SE.mean CI.mean.0.95          var
4.508007e+03 7.165501e+01 1.406748e+02 3.748141e+06
     std.dev     coef.var
1.936012e+03 4.294607e-01
```

**Data Processing and Model fitting:**

There are no NA values or missing values in the data set. We have 730 observations and 16 variables in the data set.

**Step1:** We have divided the total into two parts. 90% of data to training set and 10% data to test set.

Training set has 657 observations and test set has 73 observations.

We fit the full MLR model on training set to train.

**R Code used:**

```
set.seed(123)
indx=sample(1:nrow(data1),0.9*nrow(data1))
traindata=data1[indx,]
testdata=data1[-indx,]
dim(traindata)
dim(testdata)
```

**Step2: Fitting MLR model**
**R Code:**

```
mod1=lm((cnt)~as.factor(season)+as.factor(weekday)+as.factor(workingday)+as.factor(yr)+as.factor(mnth)+as
.factor(holiday)+as.factor(weathersit)+temp+atemp+windspeed+hum,data=traindata)
```

**Summary result:**

```
Coefficients:
                    Estimate Std. Error t value
(Intercept)            1907.267    252.605   7.550
as.factor(season)2       829.915    186.776   4.443
as.factor(season)3       771.912    218.453   3.534
as.factor(season)4      1535.030    188.693   8.135
as.factor(weekday)1      442.422    505.011   0.876
as.factor(weekday)2      370.781    502.073   0.738
as.factor(weekday)3      583.853    496.239   1.177
as.factor(weekday)4      675.034    504.648   1.338
as.factor(weekday)5      722.529    506.166   1.427
as.factor(weekday)6      -78.508    110.449  -0.711
as.factor(workingday)1  -777.177    501.789  -1.549
as.factor(yr)1          2027.197     60.644  33.428
as.factor(mnth)2         108.894    150.855   0.722
as.factor(mnth)3         573.341    172.314   3.327
as.factor(mnth)4         523.242    258.774   2.022
as.factor(mnth)5         828.414    281.146   2.947
as.factor(mnth)6         752.308    295.994   2.542
as.factor(mnth)7         282.785    327.446   0.864
as.factor(mnth)8         706.359    314.628   2.245
as.factor(mnth)9        1186.974    275.826   4.303
as.factor(mnth)10        658.977    253.519   2.599
as.factor(mnth)11        -31.650    241.506  -0.131
as.factor(mnth)12       -160.889    189.284  -0.850
as.factor(holiday)1     -969.810    449.195  -2.159
as.factor(weathersit)2  -456.222     80.781  -5.648
as.factor(weathersit)3 -2047.232    201.824 -10.144
temp                     -25.063     57.122  -0.439
atemp                    111.330     50.423   2.208
windspeed                -38.085      6.632  -5.743
```

```
hum                    -14.176      3.035  -4.671

                    Pr(>|t|)
(Intercept)           1.54e-13 ***
as.factor(season)2     1.05e-05 ***
as.factor(season)3     0.000440 ***
as.factor(season)4     2.21e-15 ***
as.factor(weekday)1    0.381330
as.factor(weekday)2    0.460487
as.factor(weekday)3    0.239819
as.factor(weekday)4    0.181501
as.factor(weekday)5    0.153947
as.factor(weekday)6    0.477466
as.factor(workingday)1 0.121931
as.factor(yr)1          < 2e-16 ***
as.factor(mnth)2       0.470658
as.factor(mnth)3       0.000928 ***
as.factor(mnth)4       0.043600 *
as.factor(mnth)5       0.003333 **
as.factor(mnth)6       0.011273 *
as.factor(mnth)7       0.388135
as.factor(mnth)8       0.025112 *
as.factor(mnth)9       1.95e-05 ***
as.factor(mnth)10      0.009561 **
as.factor(mnth)11      0.895775
as.factor(mnth)12      0.395658
as.factor(holiday)1    0.031229 *
as.factor(weathersit)2 2.47e-08 ***
as.factor(weathersit)3  < 2e-16 ***
temp                   0.660989
atemp                  0.027611 *
windspeed              1.45e-08 ***
hum                    3.67e-06 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 756.7 on 627 degrees of freedom
Multiple R-squared:  0.8522,     Adjusted R-squared:  0.8454
F-statistic: 124.7 on 29 and 627 DF, p-value: < 2.2e-16
```

So, from the above table we see that 85% of the total variability in the response variable is being explained by MLR.

Adjusted $R^2$ is 84% after being adjusted for redundant predictors.

We see that there is no much difference in $R^2$ and Adj $R^2$ that means not many redundant predictors present in model.

P value <0.05 so we reject $H_0$

A low p-value confirms that the overall model is highly significant in predicting the response variable.

Interpretation of parameter estimates:

The results obtained from above MLR model:

We can see 7 categorical variables and 4 continuous variables

**Categorical variables:**

1. Season = 4-1 = 3 predictors
2. Year = 2-1 = 1 predictor
3. Month = 12-1 = 11 predictors
4. Holiday = 2-1 = 1 predictor
5. Weekday = 7-1 = 6 predictors
6. Working day = 2-1 = 1 predictor
7. Weather conditions = 4-1 = 3 predictors

**Continuous variables:**

1. Temp= 1 predictor
2. Atemp = 1 predictor
3. Humidity = 1 predictor
4. Windspeed = 1 predictor

Total predictors = 29 + 1(intercept) = 30 predictors.

Season, weather situation, month, windspeed, humidity are the most significant predictors of response variable as they have p value <0.05 denoted by (***).

For any categorical variable, (CATEGORY) = 0 denotes the base level for the categorical variable. All the estimates for the remaining categories are compared using this base category.

For instance, if the year has been increased to 1 year, then average count of Bike rentals would be 2027.

**Correlation Coefficient:**

Performed correlation test using below R code:

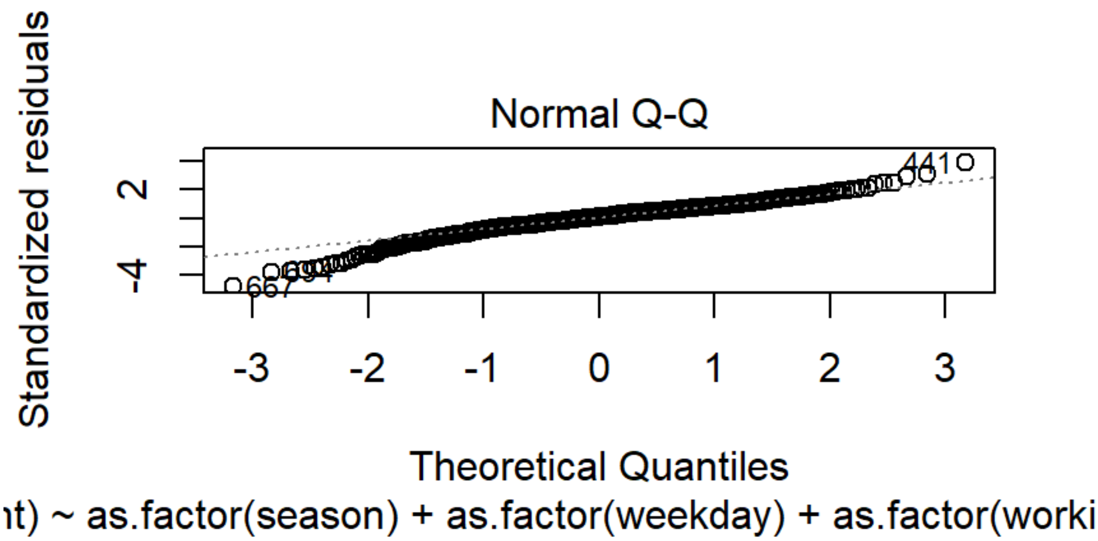
*cor(testdata$cnt,pred_val)*

we got correlation coefficient of 0.88 which indicates strong correlation between the predictive and actual response values.

**RMSE:** It is the deviation between actual and predicted response values. Lesser RMSE is good model. We got RMSE of 950.84 which way less than mean. This means model is working well.

**QQ Plot/ Shapiro Wilks test:**

**Fig4:**




The QQ plot of residuals can be used to visually check the normality assumption. The normal probability plot of residuals should approximately follow a straight line. In our model, most part of the points fall approximately along this reference line; so, we can assume normality.

**<u>Shapiro test result:</u>**

```
        Shapiro-Wilk normality test

data:  mod1$residuals
W = 0.96029, p-value = 2.469e-12
```

The Shapiro–Wilk test with a p-value < 0.05 indicates residuals' departure from normality. A Box-Cox transformation of the response variable may be used as a remedy.

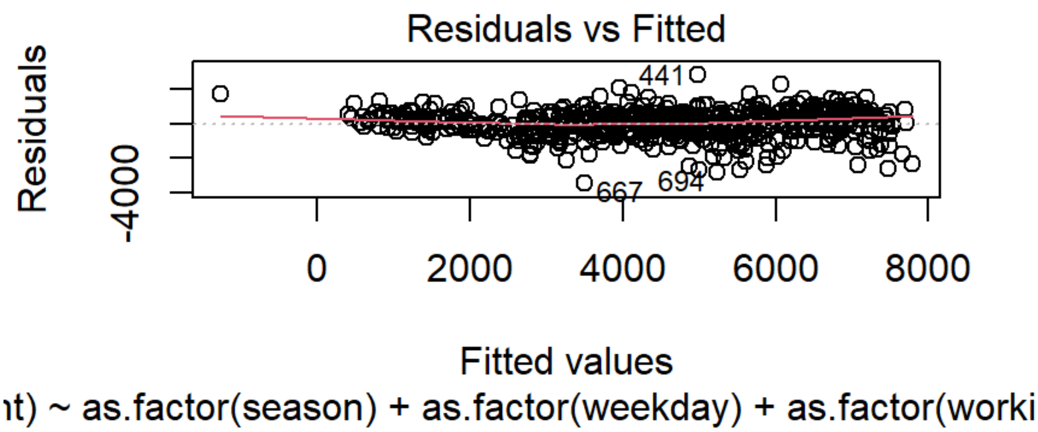
**<u>Box Cox R Code:</u>**

```
install.packages("MASS")
library(MASS)
bc = boxcox(mod1,lamda=seq(-5,5))
best.lam = bc$x[which(bc$y==max(bc$y))]
best.lam

##Adjust model by taking the response variable to the power of lamda
adjusted_mod1=lm((cnt)^0.79~as.factor(season)+as.factor(weekday)+as.factor(workingday)+as.factor(yr)+as.factor(mnth)+as.factor(holiday)+as.factor(weathersit)+temp+atemp+windspeed+hum,data=data1)

###perform Shapiro test on adjusted model
shapiro.test(adjusted_mod1$residuals)
```

So even after performing Box Cox transformation we got pvalue <0.05 hence continue with the original model.

**Homoscedasticity Check:**

**Fig 5:**




As there is no much difference in width from the centre line of the plot we can clearly conclude that homoscedasticity assumption is not violated.

**Independence check:**

Independence check is done by performing **Durbin Watson test**

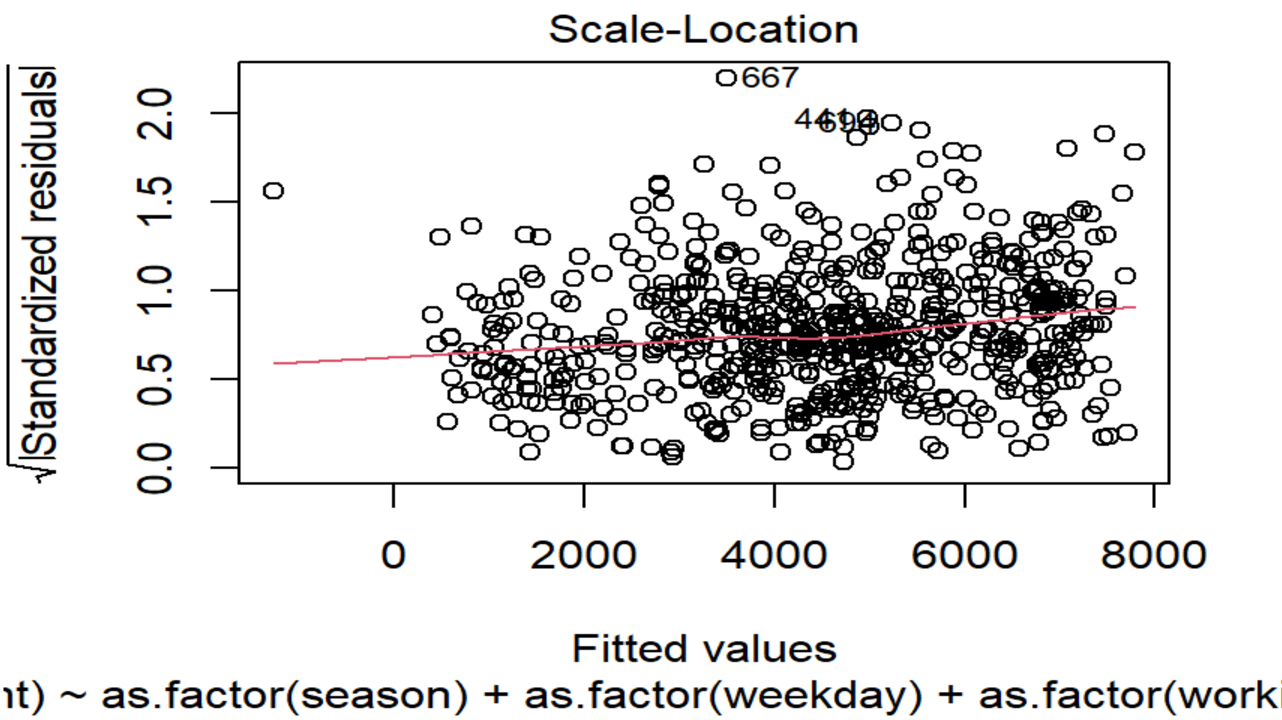
**Fig 6:**

```
        Durbin-Watson test

data:  mod1
DW = 1.9093, p-value = 0.1222
alternative hypothesis: true autocorrelation is greater tha
n 0
```

As p- value >0.05 we do not reject $H_0$ that means residuals are not autocorrelated. They are independent.

**Error Check:**

**Fig 7:**




From the plot we can see that none of the points crossed threshold 3. So, no significant outliers present is data set.

**Multicollinearity Check:**

**R code:**

*install.packages("car")*
*library(car)*
*vif(mod1)*

A VIF greater than 10 indicates multicollinearity.

The general process of handling multicollinearity is as follows:
Build the model with all the features.
Drop the feature with the highest VIF from the model.
Refit the model with the remaining predictors.
Repeat the process until no significant multicollinearity is left in the data.

**Results:**

**Fig 8:**

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| as.factor(season) | 3.457548 | 3 | 1.229687 |
| as.factor(workingday) | 1.081665 | 1 | 1.040031 |
| as.factor(yr) | 1.033135 | 1 | 1.016432 |
| as.factor(holiday) | 1.067091 | 1 | 1.033001 |
| as.factor(weathersit) | 1.850784 | 2 | 1.166377 |
| atemp | 3.302734 | 1 | 1.817343 |
| windspeed | 1.215348 | 1 | 1.102428 |
| hum | 1.938062 | 1 | 1.392143 |

**Variable Selection:**

When building a multiple linear regression model, you may have a few potential predictor variables; selecting the right ones is an extremely important exercise.
Using redundant variables may be expensive in terms of cost and time and would give very little yield. Including them may also lead to standard error inflation. We used Step wise method to identify important predictors impacting response variable by removing redundant variables.

**R code:**

```
library(MASS)
step.model=stepAIC(mod4,direction="both")
summary(step.model)
```

**Result:**

**Fig 9:**

```
                             Pr(>|t|)
(Intercept)                  3.48e-11 ***
as.factor(season)2            < 2e-16 ***
as.factor(season)3           8.96e-09 ***
as.factor(season)4            < 2e-16 ***
as.factor(workingday)1 0.007686 **
as.factor(yr)1                < 2e-16 ***
as.factor(holiday)1          0.004660 **
as.factor(weathersit)2 2.92e-07 ***
as.factor(weathersit)3  < 2e-16 ***
atemp                         < 2e-16 ***
windspeed                    9.53e-09 ***
hum                          0.000107 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 811.5 on 645 degrees of freedom
Multiple R-squared:  0.8252,    Adjusted R-squared:  0.8222
```

Applying Step-wise method, we obtain following set of predictors:
Season 2, Season 3, Season 4, workingday 1, yr 1, holiday 1, weathersit 2, weathersit 3, atemp, windspeed, humidity.

**Final MLR model:**

After checking for :

1. Model assumptions (normality, homoscedasticity and independence)
2. Outliers
3. Multicollinearity

Removing redundant predictors with stepwise regression

We obtained 11 significant predictors impacting response variable they are:
Season 2, Season 3, Season 4, workingday 1, yr 1, holiday 1, weathersit 2, weathersit 3, atemp, windspeed, humidity.

$R^2$ and Adj $R^2$ values are 0.825 and 0.822 respectively indicating strong adequacy.

As the p-value of F-test is less than 0.05 overall model is significant.

**Conclusion:**

After final MLR model we are left with Season 2, Season 3, Season 4, workingday 1, yr 1, holiday 1, weathersit 2, weathersit 3, atemp, windspeed, humidity predictors which are significant in predicting or impacting count of Rental Bikes.

In terms of season summer, fall and winter count of Rental bikes are high.
Working day is one of the significant predictors. People use Rental Bike more on working days for travelling to work, college and other. They will be resting at home when it is not a holiday spending time with family.

If we keep all the factors constant and increase humidity by 1g/kg then the count is decreased by 12units.
When coming to windspeed if we increase it by 1m/s then count is decreased by 38 units.

Season 2, season 3, season 4 and atemp has positive impact which increases count of Rental Bikes if we increase them by 1 unit.