BIKE RENTAL SYSTEM

BUSN 6324 : PREDICTIVE ANALYTICS FOR MANAGERS
Summer 1 - 2018

Submitted To - Prof. Nizar Zaarour
Submitted By – Arpit Rawat

Executive Summary / Introduction

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.
Apart from interesting real-world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into
a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important events in the city could be detected via monitoring these data.
The core data set is related to the two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA which is publicly available in http://capitalbikeshare.com/system-data. Further the data aggregated on daily basis and then extracted and added the corresponding weather and seasonal information. Weather information are extracted from http://www.freemeteo.com.

Data description:

Source: https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset
Sample size: 198 records (**random selection using R, sample method**)
Feature details:
- **temp**: Temperature in Celsius **(x)**
- **atemp**: Feeling temperature in Celsius
- **hum**: Normalized humidity. The values are divided to 100 (max)
- **windspeed**: Normalized wind speed. The values are divided to 67 (max)
- **cnt**: count of total rental bikes aggregated on daily basis **(y)**

Type: Numerical (All features are quantitative)

Qualitative features are excluded because they were not adding any value to the model.

The statistical analysis done with the help of Excel and R.

The objective of this report is to analyze the bike sharing data, generate some insight about the factors affecting the bike rental. Finally, we will do regression analysis and predict daily bike rental count based on the environmental and seasonal settings.
In Initial Report we found out the Bike Rental count (cnt) is highly correlated to the Temperature (temp) and that's why we decide to perform Simple Linear Regression on these 2 parameters.

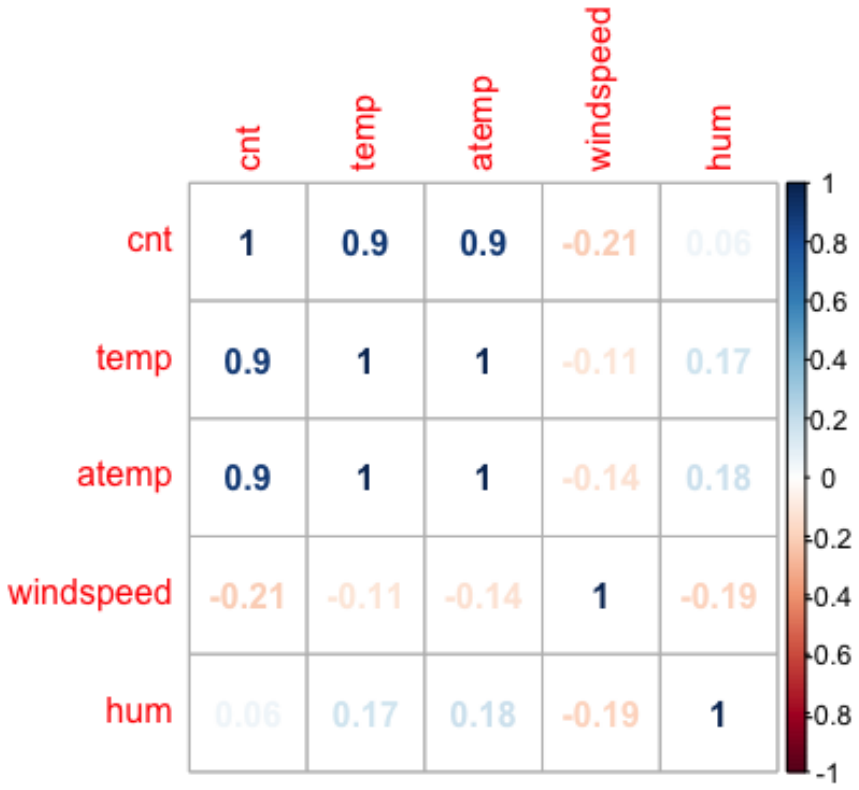
After analyzing the result, we got from Simple Linear Regression i.e. Coefficient of Determination (R-squared) equal to 0.8072 we can say that 80.72% of variance in bike counts is explained by variation in temperature.
We recommended that we should consider more features for regression analysis, so we can achieve better Adjusted R-squared value which will further help in increasing the prediction accuracy.
We managed to improve the model little by adding more features but still more improvement can be done by collecting more data.

Results and Discussions

Firstly, we performed descriptive analysis to get the idea about distribution and correlation and here are the results:
cnt – rented bike count **(response variable)**
temp – temperature **(explanatory variable)**
atemp – feeling temperature **(explanatory variable)**
hum – humidity **(explanatory variable)**
windspeed – speed of wind **(explanatory variable)**

cnt	**temp**	**atemp**	**hum**	**windspeed**
Min.: 623	Min: 3.957	Min.: 4.052	Min.: 0.00	Min.: 3.375
Median :4642	Median :22.191	Median :21.665	Median: 62.06	Median: 12.125
Mean :4710	Mean :21.253	Mean :20.333	Mean: 62.39	Mean: 12.127
Max. :8227	Max. :34.782	Max. :34.477	Max.: 96.25	Max.: 25.833

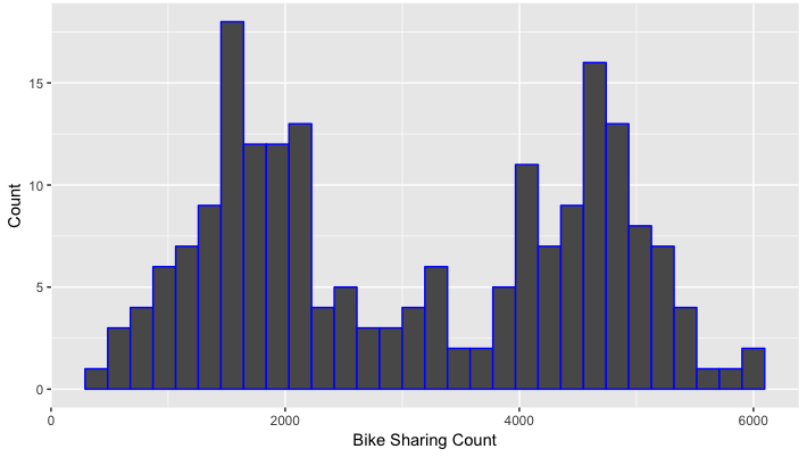
We performed **correlation analysis** to get the features which is highly correlated with our outcome variable.

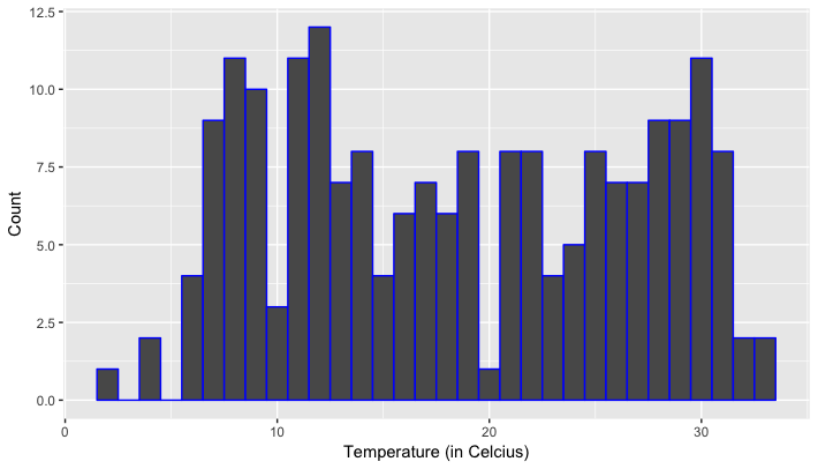


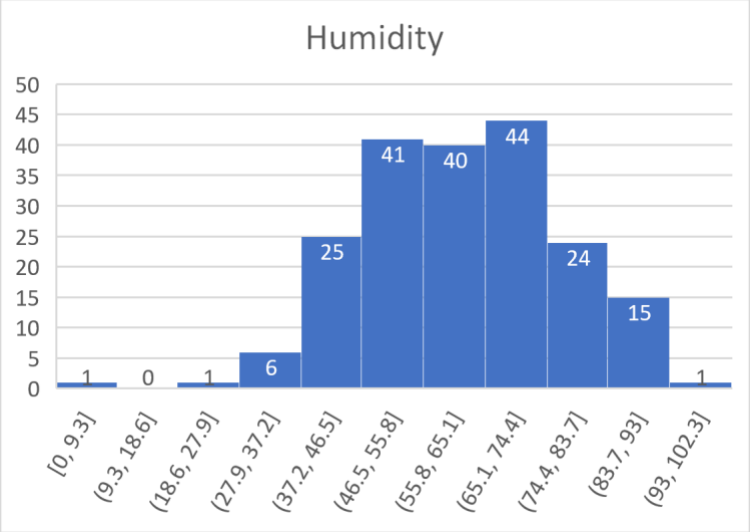
Column1	temp	atemp	hum	windspeed	cnt
temp	1				
atemp	**0.997691378**	1			
hum	0.167888447	0.18097622	1		
windspeed	-0.113731554	-0.1350151	-0.1909161	1	
cnt	**0.89845248**	**0.89630109**	0.05847274	-0.214104042	1

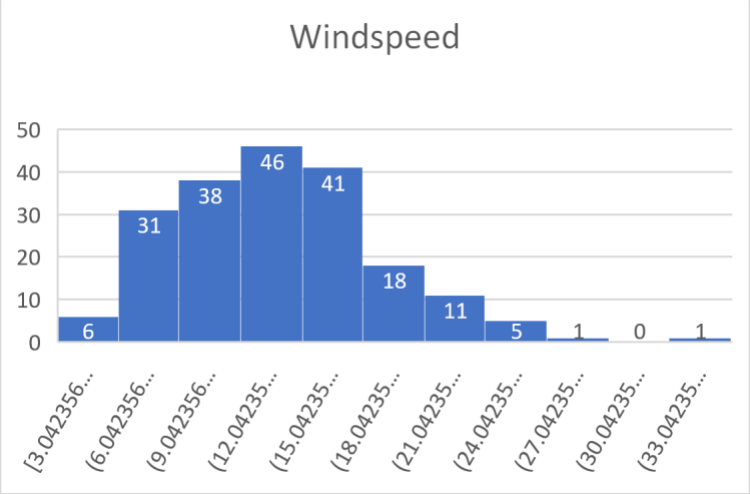
We ignored **atemp** because it is highly correlated with temp which means it will explain same variance in outcome variable as compared to temp variable (Multicollinearity).

We plotted histogram to see the distribution:

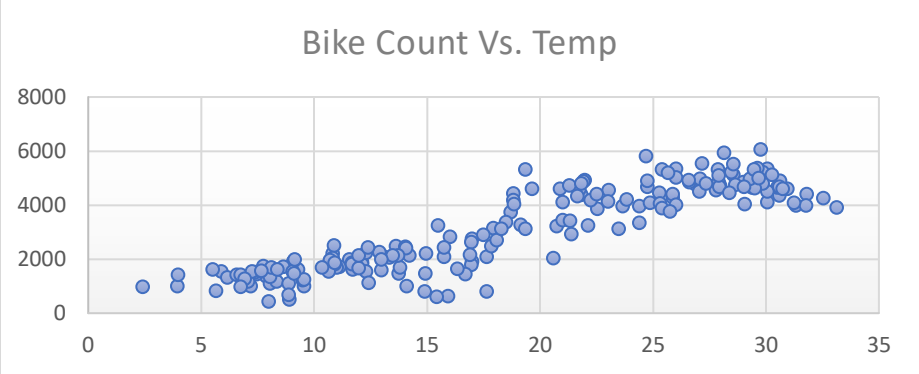


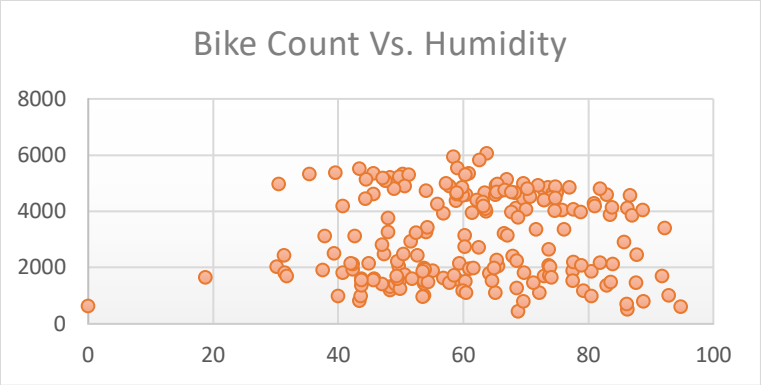


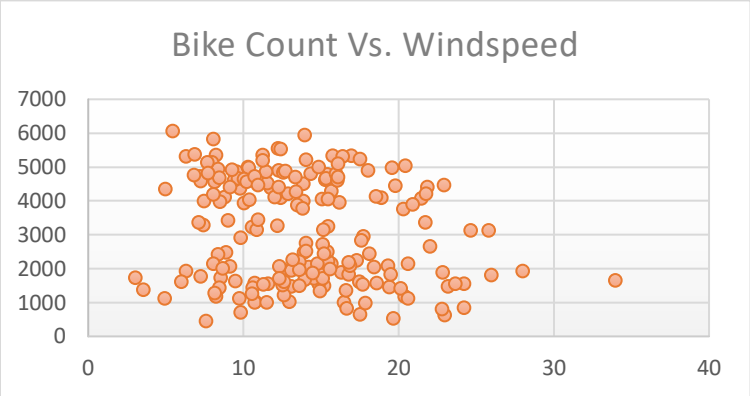


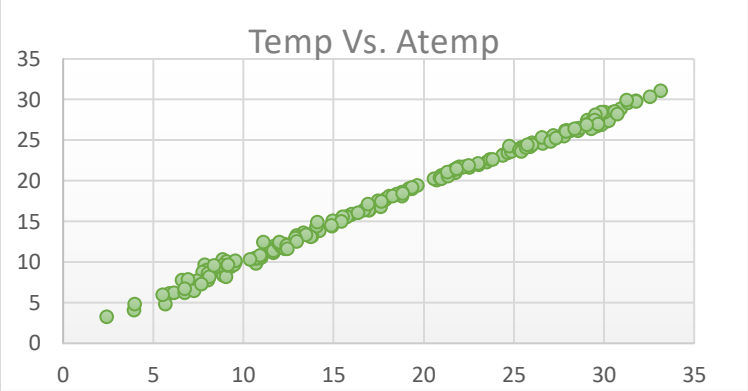


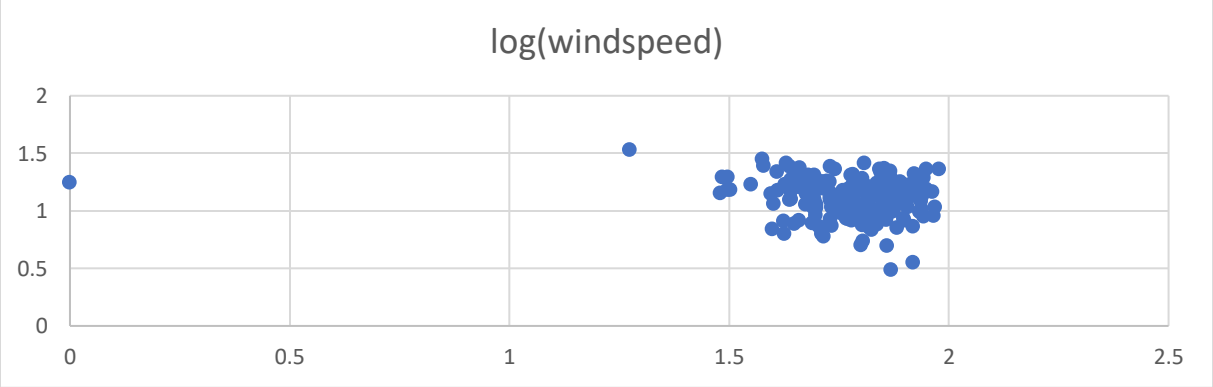
We will check for the linear relation between these two variables with the help of **Scatter plot**




Bike Count Vs. Humidity
8000
6000
4000
2000
0
0
20
40
60
80
100


Bike Count Vs. Windspeed
7000
6000
5000
4000
3000
2000
1000
0
0
10
20
30
40


Temp Vs. Atemp
35
30
25
20
15
10
5
0
0
5
10
15
20
25
30
35


log(windspeed)
2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5

Feature Extractions/Engineering:

From Correlation analysis and Linear pattern b/w Outcome and Exploratory variables we can conclude the following:

- **atemp** predictor is highly correlated with temp so we can drop it (Multicollinearity)
- **Humidity** predictor has very less correlation (0.05) and doesn't show any linear relation with outcome variable Bike Count and we can drop it too
- **Windspeed** predictor has also very less correlation which means it explains very less variance in outcome variable (bike count) and we can drop it too.

We are finally left with Temperature (temp) (explanatory) and Bike Count (outcome) for our regression analysis.

We will also try to include **Humidity** and **Windspeed** later in our model to check if we can improve our model. The reason behind it is that though these variables have less correlation with outcome variable but in combination with Temperature we might achieve better accuracy because all these explanatory variables can explain variance in outcome variable (Bike Count).

Since we have just one explanatory variable we will perform Simple Linear Regression, it is same as our Initial Report. Here are the results from it:
After performing Simple Linear Regression, we got the estimated regression equation:

y = -19.227+166.85x
y – Bike count
x – Temperature

For 1 unit increase in Temperature (x) there will be 166.85 increase in the bike count.

Intercept value tells us that -19.227 is the portion of the bike count which is not explained by the Temperature.

Coefficient of determination R^2 = 0.8072 which means **80.72%** of variance in bike rent counts is explained by variation in temperature.
Rest of the variance (19.28%) in bike counts could be explained with help of other features (f.e humidity, wind speed etc.)
SSE (Error Sum of Squares) which tells about unexplained variation is 88279347.88 which is much higher this can be decreased with the introduction of more features (Multivariate Regression)

Testing for significance:

$\alpha = 0.05$
p-value: 5.4702E-72 ($p < \alpha$) – this means we reject out null hypothesis (H0) that there is not linear relationship between Bike Count and Temperature.

Let's include **Humidity** and **Windspeed** and check Adjusted R2, p-value and residual plot to see if we can improve our model.

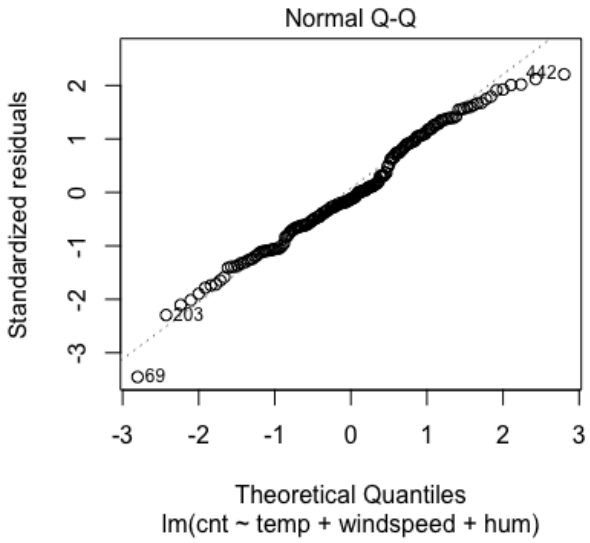
Multivariate Regression is done with the help of R and here is the summary:

```
Call:
lm(formula = cnt ~ temp + windspeed + hum, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2585.26  -318.42    16.06   351.23  2044.66

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1239.028    256.635   4.828 2.79e-06 ***
temp         167.720      5.544  30.254  < 2e-16 ***
windspeed    -40.258      9.003  -4.472 1.32e-05 ***
hum          -11.658      2.966  -3.930 0.000118 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 627.5 on 194 degrees of freedom
Multiple R-squared:  0.8332,    Adjusted R-squared:  0.8306
F-statistic:   323 on 3 and 194 DF,  p-value: < 2.2e-16
```



Equation from summary (Estimate column):

Bike Count (predicted) = 1239.028 + 167.720(Temperature) - 40.25 (Windspeed) - 11.65 (Humidity)

The equation tells us that:

b1 = 167.72, Bike count will increase on average by 167 per day with increase in 1degree Celsius increase in temperature, net of the effects of changes due to humidity and windspeed

b2 = - 40.25, Bike count will decrease on average by 40.25 per day for 1km/h increase in windspeed, net of the effects of changes due to humidity and temperature

b3 = -11.65, Bike count will decrease on average by 11.65 per day for 1% increase in humidity, net of the effects of changes due to humidity and temperature

We can see the Adjusted R-squared increased by 3% which is not much but still it is better than before.

p-value (from last line, F-statistic) for whole model is also very less which tells us out model is significant.
Also, p-value for individual feature (test statistics) is < 0.05 which means there is an evidence that Temperature, Humidity and Windspeed affect Bike Count.

Residual graph for Normality Test shows that errors are following normal distribution but not perfect at the tails (top right and bottom left). This is expected because we still haven't explained 17% of variance in outcome variable since we have Adjust R2 value as 83.06%.

Standard error is still a lot in our model and we will discuss about in our conclusion. This signifies we still don't have our best model and the reason could be that this what best we will get from our data. We will give some suggestion later to improve model.

We further tried different combinations of features to build model but none of them were better. Standard Error increased and Adjusted R2 value decreased too.

```
Call:
lm(formula = cnt ~ temp + windspeed, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1900.61  -379.12     0.25   374.71  2164.84

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  497.568    180.316   2.759 0.006342 **
temp         164.454      5.680  28.951  < 2e-16 ***
windspeed    -34.051      9.186  -3.707 0.000273 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 650.3 on 195 degrees of freedom
Multiple R-squared:  0.8199,    Adjusted R-squared:  0.8181
F-statistic: 443.9 on 2 and 195 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = cnt ~ hum + windspeed, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3050.5 -1319.0  -104.4  1308.8  2860.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3840.634    576.720   6.659 2.73e-10 ***
hum            1.793      6.995   0.256   0.7980
windspeed    -63.249     21.396  -2.956   0.0035 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1497 on 195 degrees of freedom
Multiple R-squared:  0.04616,   Adjusted R-squared:  0.03638
F-statistic: 4.719 on 2 and 195 DF,  p-value: 0.009972
```

Conclusions:

From Simple Linear Regression we found that Coefficient of Determination (R-squared) equal to 0.8072 and we can say that 80.72% of variance in bike counts is explained by variation in temperature. From Multivariate Regression we found that 83% (Adjusted R2) of variance in bike counts is explained by variation in temperature, humidity and windspeed. This is still not a perfect model, but the error got reduces little bit. We also tried different combination of features to build model but the best one was only Temperature (temp), Humidity (hum) and Windspeed with Bike Count (cnt) as outcome variable.

Our suggestion is that the increase in sample size will be helpful in analysis since we will get better insight about data and eventually mitigate overfitting issue in regression analysis.

Collecting more data will definitely help us in improving our model and it will help in decreasing the Residual errors.
Trying some non-linear model (windspeed) could be helpful but I didn’t get satisfactory result with logarithmic model may be some polynomial model will work.

Our findings about the bike rent count correlation with temperature makes sense because people tends to rent bike more in a pleasant weather and this is why we got the higher correlation between bike rent count and temperature (0.9).

This analysis will be useful to the Bike Rental companies in handling the supplies of bikes according to the demand.