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
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

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.

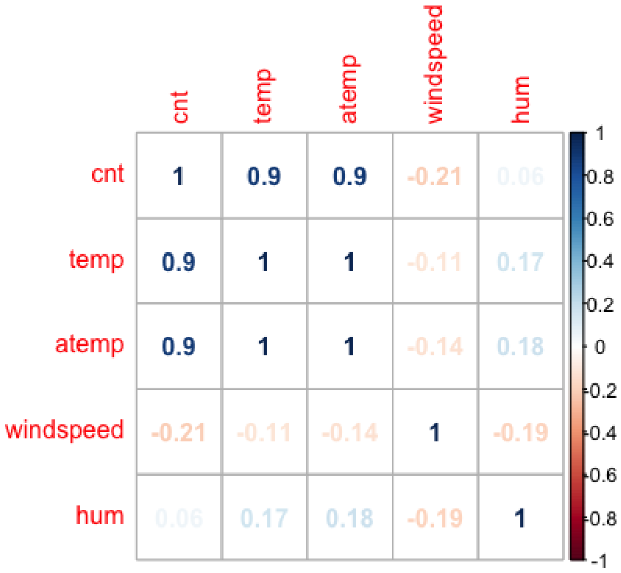
After doing feature engineering on our project we found out the Bike Rental count (cnt) is majorly correlated to the Temperature (temp) and that's why we decide to perform Simple Linear Regression for these 2 parameters.
After analyzing the result, we got from Simple Linear Regression i.e. Coefficient 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 recommend to that we should consider more features for regression analysis, so we can achieve better R-squared value which will further help in increasing the prediction accuracy. Another recommendation is to use more data because it will help in reducing the overfitting issue.

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)**

cnt		**temp**	
Min.:	431	Min.:	2.424
Median:	2911	Median:	18.023
Mean:	3073	Mean:	18.535
Max.:	6043	Max.:	33.142

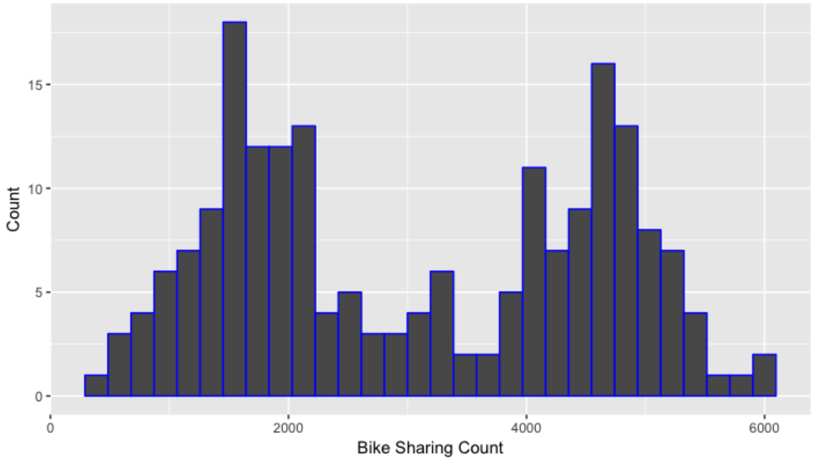
We performed **correlation analysis** to get the feature which is highly correlated with our outcome variable.

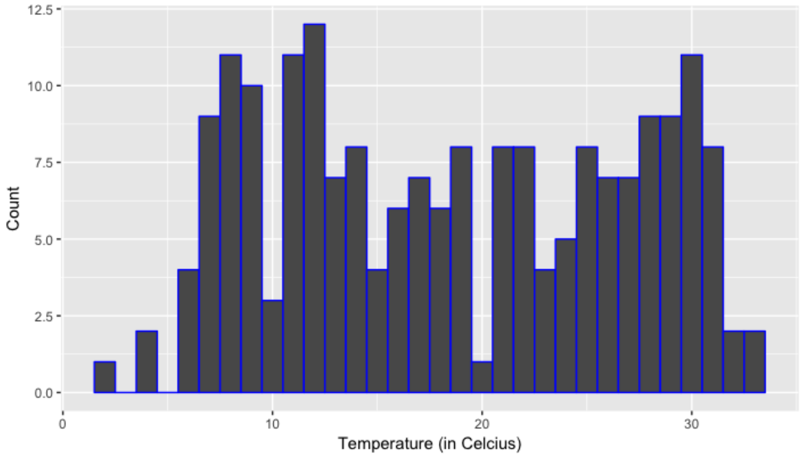


	cnt	temp	atemp	windspeed	hum
cnt	1	0.9	0.9	-0.21	0.06
temp	0.9	1	1	-0.11	0.17
atemp	0.9	1	1	-0.14	0.18
windspeed	-0.21	-0.11	-0.14	1	-0.19
hum	0.06	0.17	0.18	-0.19	1

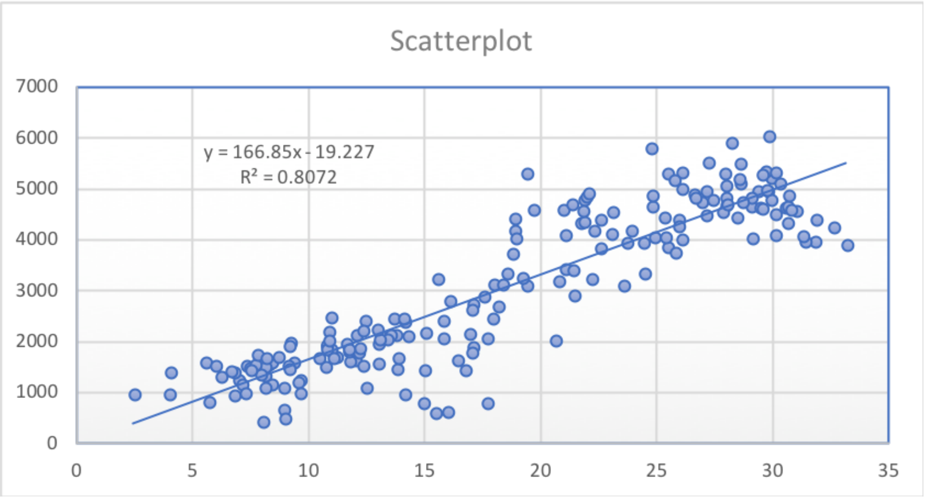
We choose temp (temperature) for our Simple Linear Regression model since it has higher correlation with our outcome variable. We ignored atemp because it is highly correlated with temp which means it will explain same variance in outcome variable as compared to temp variable (Multicollinearity).

We plotted histogram see the distribution:





Now we have our explanatory variable **temp** and response variable **cnt**. We will check for the linear relation between these two variables with the help of **Scatter plot**, Y axis Bike Count and X axis Temperature:



We can see these variable have positive linear relationship though not perfect.

After performing Simple Linear Regression, we got the estimated regression equation:

y = -19.227+166.85x

y – Bike count
x – Temperature
b_0– -19.227 (intercept)
b_1– 166.85 (slope)

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 will 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)
Standard Error of estimate is 118.015
Rest of the results from Regression analysis are here:

Statistics from Simple Linear Regression:

Slope: b_1	**166.849**
b_0	**-19.227**

SSE	**88279347.88**
SSR	**369641129.96**
SST	**457920477.84**

R^2	**0.80721686**

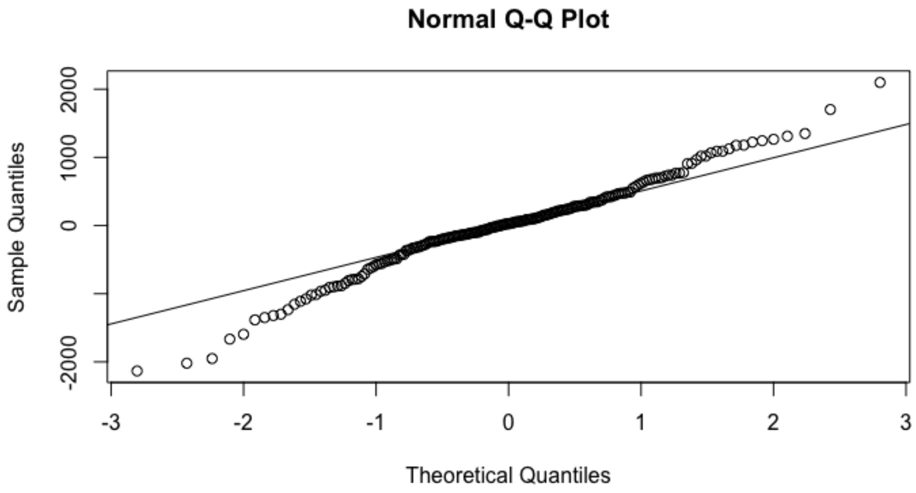
Estimate	**MSE**	**450404.836**
	S_e	**671.122**
	Sb_1	**5.824**

95% confidence level	**t stat**	**28.648**	**$t_{\alpha/2}$**	**1.664**
	F stat	**196.000**	**F_α**	**4.26**

Residuals Statistics:

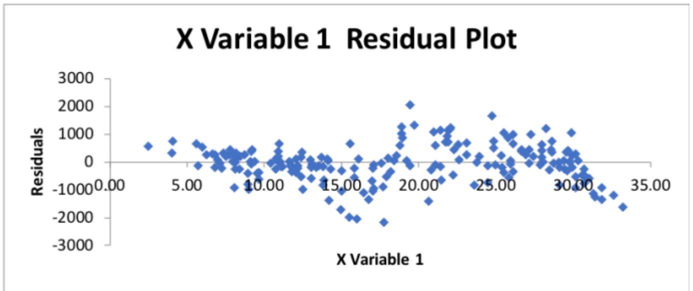
Min	Median	Max
-2133.02	26.15	2098.95

Residual Analysis - We know Residuals should follow Normal Distribution, here is the Q-Q plot (from R tool) which gives out idea about it:



We can see that residuals are not perfectly normally distributed which tells us there is some problem with our regression model, we have to perform some more analysis like outlier detection to minimize the variance in data and hence making the Residuals follow Normal Distribution.

Independence:



The graph shows the independence of errors which fulfills assumption of Linear Regression.

Testing for significance:

α = 0.05

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

Pick a random value for “x” within the range of your data and build a 95% confidence interval for E(y_p) and a 95% prediction interval for y_p.

Xp	Bike Count
22.49	3,733.20

95% C.I.	95% P.I.
M.O.E.	M.O.E.
88.13694509	1120.21974

95% C.I.	95% P.I.
3,645.07	2,612.98
3,821.34	4,853.42

This confidence interval seems fine when we compare it with the datasets.

Conclusions

We found that Coefficient of Determination (R-squared) equal to 0.8072 and we can say that 80.72% of variance in bike rent counts is explained by variation in temperature. Rest of the variance (19.28%) in bike counts will be explained with help of other features (f.e humidity, wind speed etc.).

The increase in sample size will be also helpful in analysis since we will get better insight about data and eventually mitigate overfitting issue in regression analysis.

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).

The other analysis we can do after considering more features in building regression model is to select best features which gives us highest accuracy for the model in predictions the bike rental count.

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