Project Report

**Sahil kumar**

**February 26, 2018**

**Title : Analysis of Bike Sharing Dataset**

**Name : Sahil kumar**

**Email :** [**sahilriders@gmail.com**](mailto:sahilriders@gmail.com)

**College : Institute of Engineering and Management , Kolkata**

# Introduction

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world.

The data generated by these systems makes them attractive for researchers because the duration of travel, departure location, arrival location, and time elapsed is explicitly recorded. Bike sharing systems therefore function as a sensor network, which can be used for studying mobility in a city.

# Data Description

Dataset contains these attributes :

- instant: record index  
- dteday : date  
- season : season (1:springer, 2:summer, 3:fall, 4:winter)  
- yr : year (0: 2011, 1:2012)  
- mnth : month ( 1 to 12)  
- hr : hour (0 to 23)  
- holiday : weather day is holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule)  
- weekday : day of the week  
- workingday : if day is neither weekend nor holiday is 1, otherwise is 0.  
+ weathersit :   
 - 1: Clear, Few clouds, Partly cloudy, Partly cloudy  
 - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist  
 - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds  
 - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog  
- temp : Normalized temperature in Celsius. The values are divided to 41 (max)  
- atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max)  
- hum: Normalized humidity. The values are divided to 100 (max)  
- windspeed: Normalized wind speed. The values are divided to 67 (max)  
- casual: count of casual users  
- registered: count of registered users  
- cnt: count of total rental bikes including both casual and registered

# Model Analysis

**Hypothesis:** Bike-sharing rental process is highly correlated to the environmental and seasonal settings. For instance, weather conditions,precipitation, temperature, humidity, windspeed , casual and registered users.

In order to test this hypothesis, we are proposing this model:

y=b0 + b1\*weathersit + b2\*temp + b3\*atemp + b4\*hum + b5\*windspeed + b6\*casual + b7\*registered

df <- read.csv(paste("hour.csv",sep=""))  
mod1 <-lm(cnt~weathersit+temp+atemp+hum+windspeed+casual+registered,df)  
summary(mod1)

##   
## Call:  
## lm(formula = cnt ~ weathersit + temp + atemp + hum + windspeed +   
## casual + registered, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.355e-11 -7.000e-14 -3.000e-14 2.000e-14 3.864e-10   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.928e-12 1.286e-13 2.277e+01 < 2e-16 \*\*\*  
## weathersit -1.223e-13 4.011e-14 -3.049e+00 0.00230 \*\*   
## temp 1.561e-12 7.842e-13 1.991e+00 0.04652 \*   
## atemp 7.901e-13 8.811e-13 8.970e-01 0.36991   
## hum -1.816e-13 1.474e-13 -1.231e+00 0.21819   
## windspeed -5.273e-13 2.034e-13 -2.593e+00 0.00953 \*\*   
## casual 1.000e+00 5.998e-16 1.667e+15 < 2e-16 \*\*\*  
## registered 1.000e+00 1.786e-16 5.600e+15 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.013e-12 on 17371 degrees of freedom  
## Multiple R-squared: 1, Adjusted R-squared: 1   
## F-statistic: 8.995e+30 on 7 and 17371 DF, p-value: < 2.2e-16

# Results

We can see that the Residual standard error is very low which means that overall the model is a good fit. p-value corresponding to each variable is also less than 0.5 which means that they are significant for the model. Most significant variables are Weathersit, temp, windspeed, casual and registered. F-statistic is also very high which means tht our hypothesis is compatible with the observed data.

# Conclusion

Bike sharing systems function as a sensor network, which can be used for studying mobility in a city. Through this project we found out that there are several factors which affect the bike sharing and these includes Situtation of the weather, Temperature, Humidity , Windspeed , No. of casual and registered users in the city.

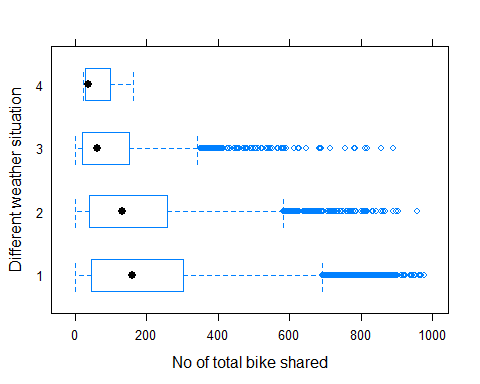
# Appendix

head(df)

## instant dteday season yr mnth hr holiday weekday workingday  
## 1 1 2011-01-01 1 0 1 0 0 6 0  
## 2 2 2011-01-01 1 0 1 1 0 6 0  
## 3 3 2011-01-01 1 0 1 2 0 6 0  
## 4 4 2011-01-01 1 0 1 3 0 6 0  
## 5 5 2011-01-01 1 0 1 4 0 6 0  
## 6 6 2011-01-01 1 0 1 5 0 6 0  
## weathersit temp atemp hum windspeed casual registered cnt  
## 1 1 0.24 0.2879 0.81 0.0000 3 13 16  
## 2 1 0.22 0.2727 0.80 0.0000 8 32 40  
## 3 1 0.22 0.2727 0.80 0.0000 5 27 32  
## 4 1 0.24 0.2879 0.75 0.0000 3 10 13  
## 5 1 0.24 0.2879 0.75 0.0000 0 1 1  
## 6 2 0.24 0.2576 0.75 0.0896 0 1 1

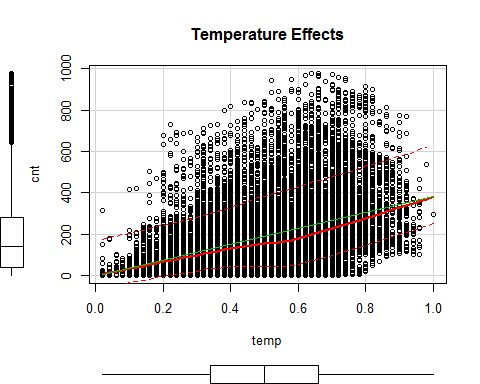
# EFFECT OF WEATHER SITUATION ON TOTAL BIKE SHARING

library(car)  
library(ggplot2)  
library("lattice", lib.loc="C:/Program Files/R/R-3.4.3/library")  
bwplot(weathersit~cnt,data=df,xlab ="No of total bike shared",ylab="Different weather situation",main="")

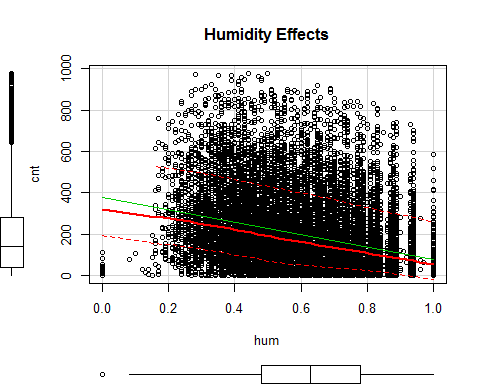


# Scatterplot of temp,hum,windspeed vs total bike sharing

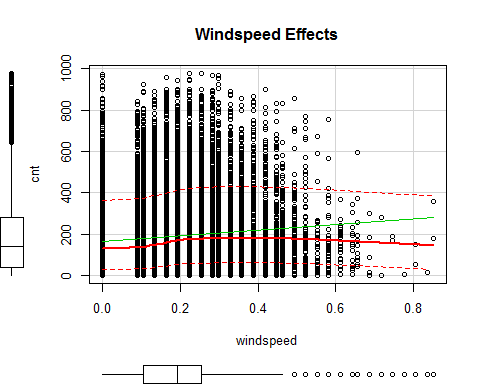
scatterplot(cnt~temp,data=df,main="Temperature Effects")



scatterplot(cnt~hum,data=df,main="Humidity Effects")



scatterplot(cnt~windspeed,data=df,main="Windspeed Effects")



df\_subset <- df[,10:17]  
df\_cor <- cor(df\_subset)  
library(corrplot)

## corrplot 0.84 loaded

corrplot(df\_cor, method = 'color', addCoef.col="black")

