Bike Sharing

Sahil kumar

February 18, 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

# Setting Working Directory

setwd("G:/Intern - IIM Lucknow/Bike-Sharing-Dataset")

# Loading necessary packages

library(psych)  
library(car)

##   
## Attaching package: 'car'

## The following object is masked from 'package:psych':  
##   
## logit

library(corrgram)  
library(ggplot2)

##   
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':  
##   
## %+%, alpha

library("lattice", lib.loc="C:/Program Files/R/R-3.4.3/library")

# Reading Dataset

df <- read.csv(paste("hour.csv",sep=""))  
View(df)

# Describing Dataset

describe(df)

## vars n mean sd median trimmed mad min  
## instant 1 17379 8690.00 5017.03 8690.00 8690.00 6441.90 1.00  
## dteday\* 2 17379 367.17 210.07 367.00 367.24 269.83 1.00  
## season 3 17379 2.50 1.11 3.00 2.50 1.48 1.00  
## yr 4 17379 0.50 0.50 1.00 0.50 0.00 0.00  
## mnth 5 17379 6.54 3.44 7.00 6.54 4.45 1.00  
## hr 6 17379 11.55 6.91 12.00 11.56 8.90 0.00  
## holiday 7 17379 0.03 0.17 0.00 0.00 0.00 0.00  
## weekday 8 17379 3.00 2.01 3.00 3.00 2.97 0.00  
## workingday 9 17379 0.68 0.47 1.00 0.73 0.00 0.00  
## weathersit 10 17379 1.43 0.64 1.00 1.30 0.00 1.00  
## temp 11 17379 0.50 0.19 0.50 0.50 0.24 0.02  
## atemp 12 17379 0.48 0.17 0.48 0.48 0.20 0.00  
## hum 13 17379 0.63 0.19 0.63 0.63 0.22 0.00  
## windspeed 14 17379 0.19 0.12 0.19 0.18 0.13 0.00  
## casual 15 17379 35.68 49.31 17.00 25.13 23.72 0.00  
## registered 16 17379 153.79 151.36 115.00 129.26 131.95 0.00  
## cnt 17 17379 189.46 181.39 142.00 162.04 166.05 1.00  
## max range skew kurtosis se  
## instant 17379.00 17378.00 0.00 -1.20 38.06  
## dteday\* 731.00 730.00 0.00 -1.19 1.59  
## season 4.00 3.00 -0.01 -1.33 0.01  
## yr 1.00 1.00 -0.01 -2.00 0.00  
## mnth 12.00 11.00 -0.01 -1.20 0.03  
## hr 23.00 23.00 -0.01 -1.20 0.05  
## holiday 1.00 1.00 5.64 29.78 0.00  
## weekday 6.00 6.00 0.00 -1.26 0.02  
## workingday 1.00 1.00 -0.79 -1.38 0.00  
## weathersit 4.00 3.00 1.23 0.35 0.00  
## temp 1.00 0.98 -0.01 -0.94 0.00  
## atemp 1.00 1.00 -0.09 -0.85 0.00  
## hum 1.00 1.00 -0.11 -0.83 0.00  
## windspeed 0.85 0.85 0.57 0.59 0.00  
## casual 367.00 367.00 2.50 7.57 0.37  
## registered 886.00 886.00 1.56 2.75 1.15  
## cnt 977.00 976.00 1.28 1.42 1.38

str(df)

## 'data.frame': 17379 obs. of 17 variables:  
## $ instant : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ dteday : Factor w/ 731 levels "2011-01-01","2011-01-02",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ season : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ yr : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ mnth : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ hr : int 0 1 2 3 4 5 6 7 8 9 ...  
## $ holiday : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ weekday : int 6 6 6 6 6 6 6 6 6 6 ...  
## $ workingday: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ weathersit: int 1 1 1 1 1 2 1 1 1 1 ...  
## $ temp : num 0.24 0.22 0.22 0.24 0.24 0.24 0.22 0.2 0.24 0.32 ...  
## $ atemp : num 0.288 0.273 0.273 0.288 0.288 ...  
## $ hum : num 0.81 0.8 0.8 0.75 0.75 0.75 0.8 0.86 0.75 0.76 ...  
## $ windspeed : num 0 0 0 0 0 0.0896 0 0 0 0 ...  
## $ casual : int 3 8 5 3 0 0 2 1 1 8 ...  
## $ registered: int 13 32 27 10 1 1 0 2 7 6 ...  
## $ cnt : int 16 40 32 13 1 1 2 3 8 14 ...

\*There are two types of customers :- 1. Casual 2. Registered

\*Last 3 columns represents bike sharing among casual, registered and both customers.

# EDA

# Season vs Bike Shared

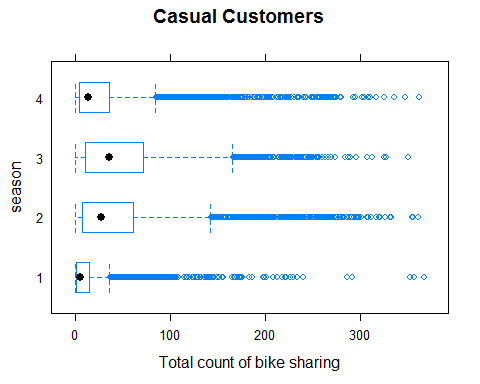
TABLE

attach(df)  
aggregate(cbind(casual,registered,cnt)~season,data=df,mean)

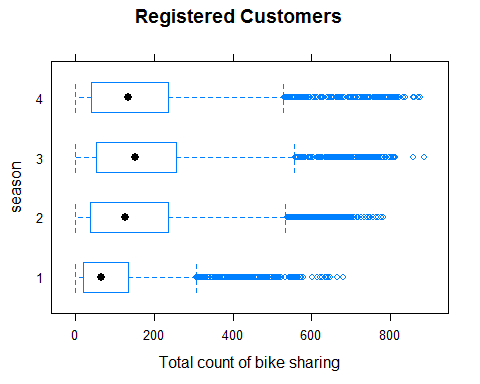
## season casual registered cnt  
## 1 1 14.29090 96.82367 111.1146  
## 2 2 46.16058 162.18349 208.3441  
## 3 3 50.28714 185.72909 236.0162  
## 4 4 30.66682 168.20203 198.8689

BOXPLOTS

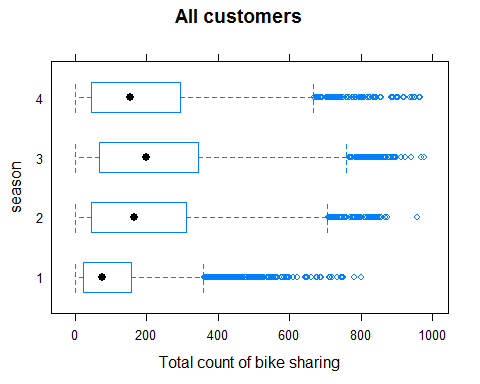
par(mfrow=c(3,1))  
bwplot(season~casual ,data=df,xlab = "Total count of bike sharing",main='Casual Customers')



bwplot(season~registered ,data=df,xlab = "Total count of bike sharing",main='Registered Customers')



bwplot(season~cnt ,data=df,xlab = "Total count of bike sharing",main='All customers')



Inference :

Maximum bike sharing is done in Season 3.

1. Bike shared among registered users is more than casual users.

# Month vs Bike shared

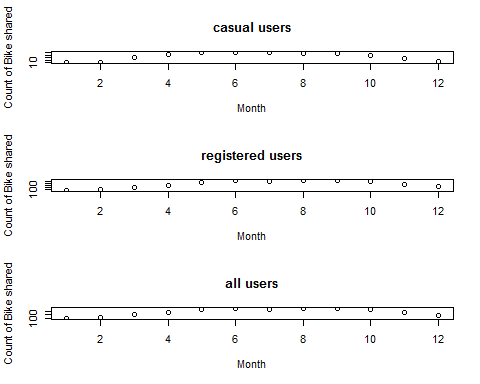
TABLE

aggregate(cbind(casual,registered,cnt)~mnth,data=df,mean)

## mnth casual registered cnt  
## 1 1 8.426872 85.9979 94.42477  
## 2 2 11.158091 101.7069 112.86503  
## 3 3 30.172437 125.2383 155.41073  
## 4 4 42.311761 144.9492 187.26096  
## 5 5 50.594758 172.3125 222.90726  
## 6 6 51.323611 189.1917 240.51528  
## 7 7 52.524866 179.2950 231.81989  
## 8 8 48.840000 189.2576 238.09763  
## 9 9 48.937370 191.8358 240.77314  
## 10 10 41.185389 180.9731 222.15851  
## 11 11 25.471816 151.8636 177.33542  
## 12 12 14.627782 127.6757 142.30344

BOXPLOTS

par(mfrow=c(3,1))  
plot(aggregate(casual~mnth ,data=df,mean),xlab='Month',ylab='Count of Bike shared',main='casual users')  
plot(aggregate(registered~mnth ,data=df,mean),xlab='Month',ylab='Count of Bike shared',main='registered users')  
plot(aggregate(cnt~mnth ,data=df,mean),xlab='Month',ylab='Count of Bike shared',main='all users')



Inference:- On an Average , most bike sharing happens between March to November month. Hypothesis:- Reason of low bike sharing in other months can be LOW TEMPERATURE.

# Hour vs Bike shared

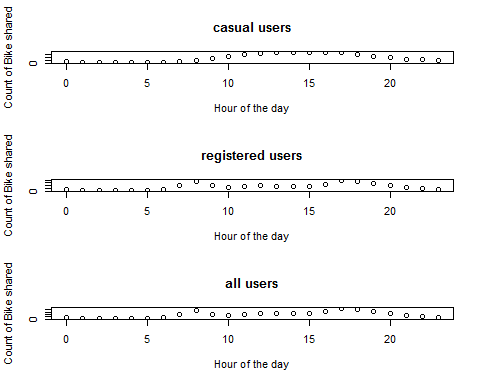
TABLE

aggregate(cbind(casual,registered,cnt)~hr,data=df,mean)

## hr casual registered cnt  
## 1 0 10.158402 43.739669 53.898072  
## 2 1 6.504144 26.871547 33.375691  
## 3 2 4.772028 18.097902 22.869930  
## 4 3 2.715925 9.011478 11.727403  
## 5 4 1.253945 5.098996 6.352941  
## 6 5 1.411437 18.478382 19.889819  
## 7 6 4.161379 71.882759 76.044138  
## 8 7 11.055021 201.009629 212.064649  
## 9 8 21.679505 337.331499 359.011004  
## 10 9 30.891334 188.418157 219.309491  
## 11 10 46.477304 127.191197 173.668501  
## 12 11 59.540578 148.602476 208.143054  
## 13 12 68.293956 185.021978 253.315934  
## 14 13 72.308642 181.352538 253.661180  
## 15 14 75.567901 165.381344 240.949246  
## 16 15 74.905350 176.327846 251.233196  
## 17 16 73.745205 238.238356 311.983562  
## 18 17 74.273973 387.178082 461.452055  
## 19 18 61.120879 364.390110 425.510989  
## 20 19 48.770604 262.752747 311.523352  
## 21 20 36.233516 189.796703 226.030220  
## 22 21 28.255495 144.059066 172.314560  
## 23 22 22.252747 109.082418 131.335165  
## 24 23 15.199176 72.631868 87.831044

BOXPLOTS

par(mfrow=c(3,1))  
plot(aggregate(casual~hr ,data=df,mean),xlab='Hour of the day',ylab='Count of Bike shared',main='casual users')  
plot(aggregate(registered~hr ,data=df,mean),xlab='Hour of the day',ylab='Count of Bike shared',main='registered users')  
plot(aggregate(cnt~hr ,data=df,mean),xlab='Hour of the day',ylab='Count of Bike shared',main='all users')



Inference :- 1. For casual customers , bike sharing starts around 7am ,reaches its peak around 2pm then further decreases. 2. In General , bike sharing reaches its peak around 8 am ,declines and again rises to reach its peak around 5pm.

Hypothesis: Bike sharing is more in morning and evening. Not in afternoon , maybe due to excess heat.

# EFFECT OF HOLIDAYS ON BIKE SHARING

TABLE

mytable<-aggregate(cbind(casual,registered,cnt)~holiday ,data=df,mean)  
mytable

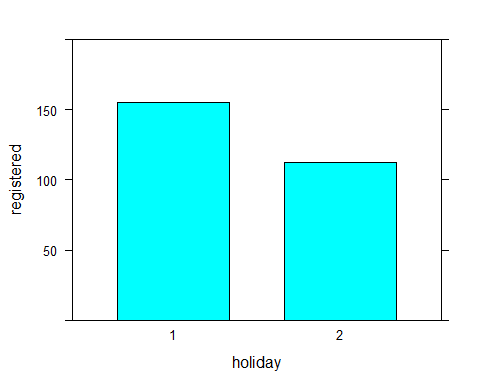
## holiday casual registered cnt  
## 1 0 35.40838 155.0202 190.4286  
## 2 1 44.71800 112.1520 156.8700

BARCHARTS

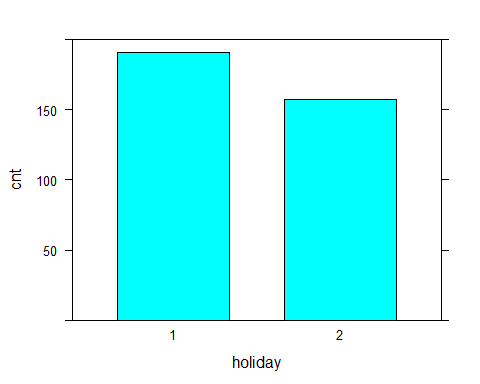
par(mfrow=c(1,3))  
barchart(casual~holiday,data=mytable,horizontal=FALSE,ylim=c(0,50))



barchart(registered~holiday,data=mytable,horizontal=FALSE,ylim=c(0,200))



barchart(cnt~holiday,data=mytable,horizontal=FALSE,ylim=c(0,200))



# EFFECT OF WEATHER SITUATION ON TOTAL BIKE SHARING

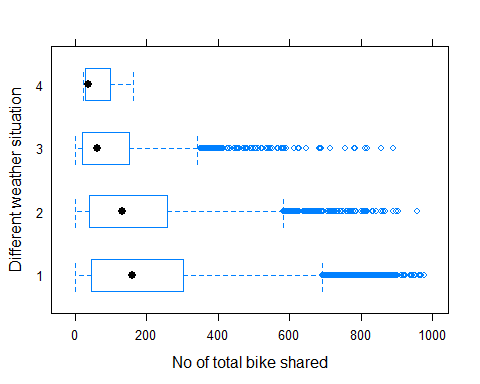
TABLE

wsittable<-aggregate(cnt~weathersit ,data=df,mean)  
wsittable

## weathersit cnt  
## 1 1 204.86927  
## 2 2 175.16549  
## 3 3 111.57928  
## 4 4 74.33333

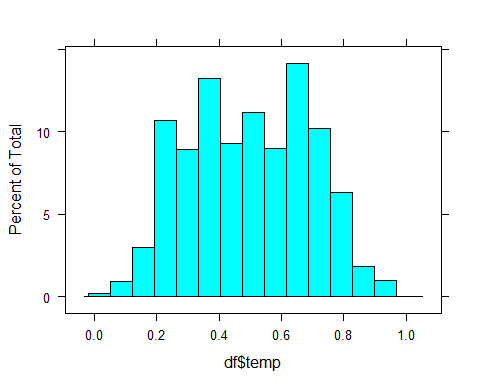
BOXPLOTS

bwplot(weathersit~cnt,data=df,xlab ="No of total bike shared",ylab="Different weather situation",main="")

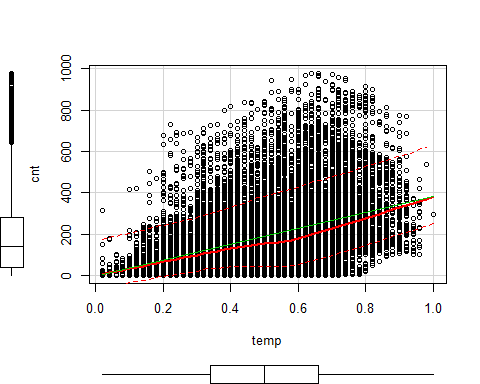
 Weather Situtation 1 is highly suitable for bike sharing.

# Scatterplot of temp vs total bike sharing

histogram(df$temp)

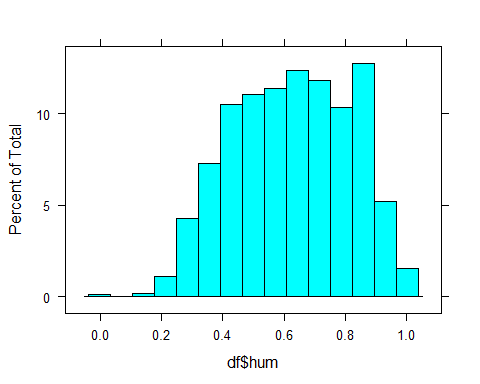


scatterplot(cnt~temp,data=df)

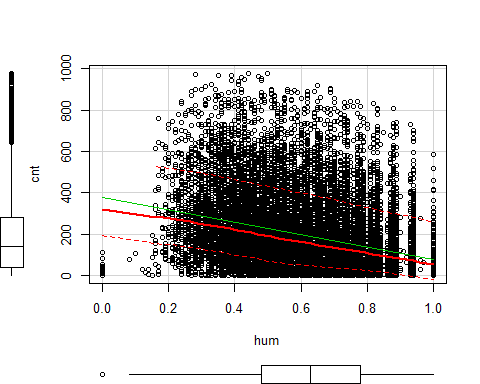


# Humidity

histogram(df$hum)

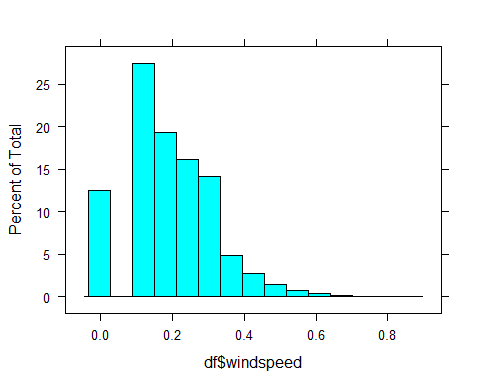


scatterplot(cnt~hum,data=df)

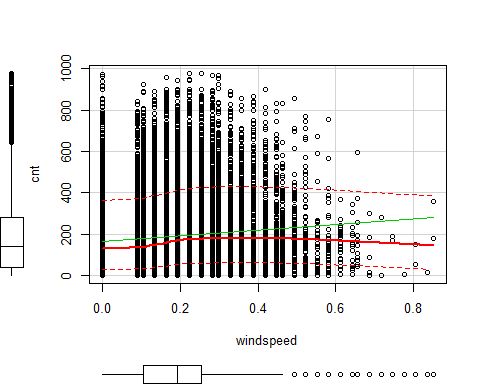


# Windspeed

histogram(df$windspeed)

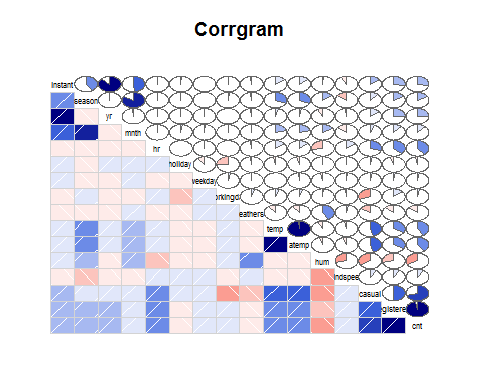


scatterplot(cnt~windspeed,data=df)



# Correlation Analysis

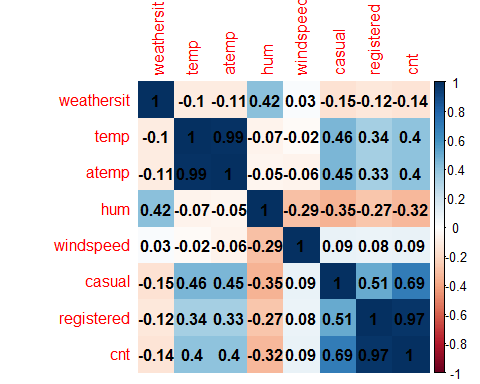
corrgram(df,lower.panel=panel.shade,  
 upper.panel=panel.pie, text.panel=panel.txt,  
 main="Corrgram")



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

## corrplot 0.84 loaded

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



# T-TEST

t.test(cnt~holiday,data=df)

##   
## Welch Two Sample t-test  
##   
## data: cnt by holiday  
## t = 4.6942, df = 539.61, p-value = 3.398e-06  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 19.51535 47.60181  
## sample estimates:  
## mean in group 0 mean in group 1   
## 190.4286 156.8700

It means that there is significance difference in bike sharing if the day is holiday or not

# Regression Analysis

We are going to fit ‘cnt’ which is count of bike shared wrt to the factors: weathersit,temp,atemp,hum,windspeed,casual,registered. 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

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

**Inference :** 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.