Bike_Sharing

Nabil Momin

2024-06-09

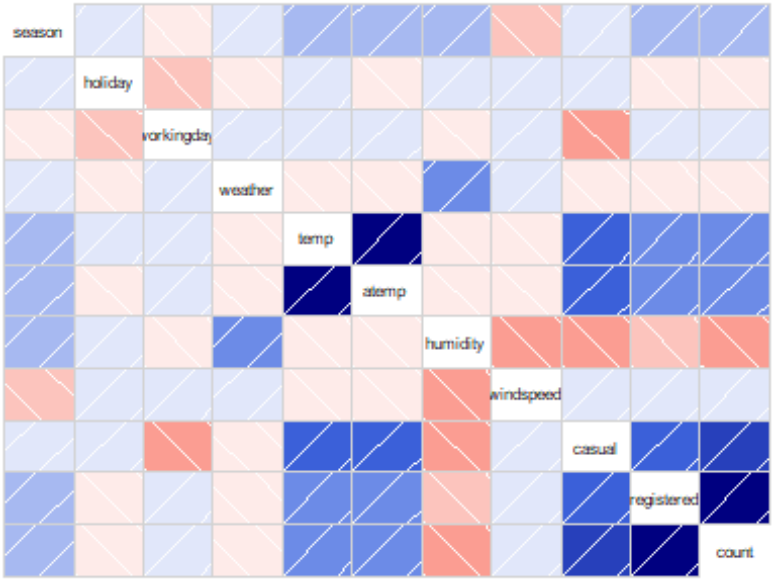
Setting up my environment

```
library(corrgram)
library(corrplot)

## corrplot 0.92 loaded

library(caTools)
library(ggplot2)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(tidyr)
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ forcats   1.0.0     ✓ readr     2.1.5
## ✓ lubridate 1.9.3     ✓ stringr   1.5.1
## ✓ purrr     1.0.2     ✓ tibble    3.2.1

## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors

library(plotly)

##
## Attaching package: 'plotly'
##
## The following object is masked from 'package:ggplot2':
```

```
##
##     last_plot
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following object is masked from 'package:graphics':
##
##     layout
```

Making bike as the data frame by fetching bikeshare.csv to bike

```
bike <- read.csv('bikeshare.csv')

corrgram(bike)
```

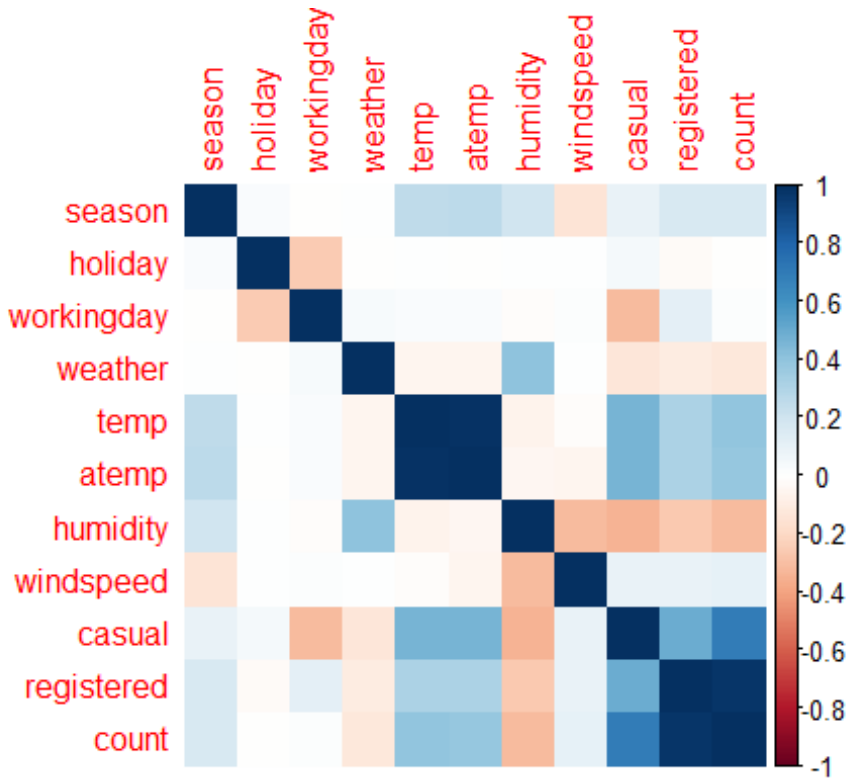


Making corrplot

```
any(is.na(bike))

## [1] FALSE

all.numeric <- sapply(bike,is.numeric)

cor.data <- cor(bike[,all.numeric])

cor.data
```

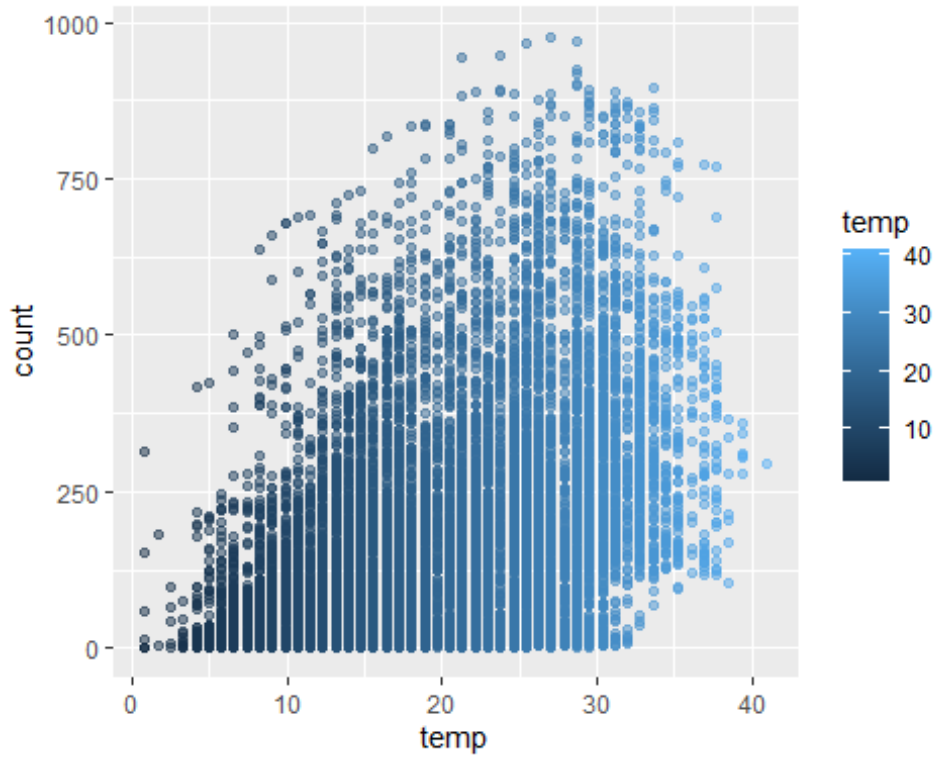
```
##                  season       holiday   workingday      weather
temp
## season      1.000000000  0.0293676097 -0.008126058  0.008878651
0.2586885508
## holiday     0.029367610  1.0000000000 -0.250491391 -0.007073912
0.0002946034
## workingday -0.008126058 -0.2504913912  1.000000000  0.033771842
0.0299655472
## weather     0.008878651 -0.0070739115  0.033771842  1.000000000 -
0.0550354182
## temp        0.258688551  0.0002946034  0.029965547 -0.055035418
1.0000000000
## atemp       0.264744326 -0.0052147782  0.024660329 -0.055375973
0.9849481105
## humidity    0.190610020  0.0019287112 -0.010879845  0.406243651 -
0.0649487709
## windspeed  -0.147121209  0.0084087378  0.013373313  0.007261124 -
0.0178520099
## casual      0.096758063  0.0437989287 -0.319110963 -0.135917680
0.4670970641
## registered  0.164010534 -0.0209556729  0.119459851 -0.109340372
0.3185712803
## count       0.163439017 -0.0053929845  0.011593866 -0.128655201
0.3944536450
##                   atemp     humidity    windspeed      casual  registered
## season      0.264744326  0.190610020 -0.147121209  0.09675806  0.16401053
## holiday    -0.005214778  0.001928711  0.008408738  0.04379893 -0.02095567
## workingday  0.024660329 -0.010879845  0.013373313 -0.31911096  0.11945985
## weather    -0.055375973  0.406243651  0.007261124 -0.13591768 -0.10934037
## temp        0.984948110 -0.064948771 -0.017852010  0.46709706  0.31857128
## atemp       1.000000000 -0.043535709 -0.057473002  0.46206654  0.31463539
## humidity   -0.043535709  1.000000000 -0.318606992 -0.34818690 -0.26545787
## windspeed  -0.057473002 -0.318606992  1.000000000  0.09227619  0.09105166
## casual      0.462066536 -0.348186899  0.092276189  1.00000000  0.49724969
## registered  0.314635386 -0.265457868  0.091051662  0.49724969  1.00000000
## count       0.389784437 -0.317371479  0.101369470  0.69041357  0.97094811
##                   count
## season      0.163439017
## holiday    -0.005392984
## workingday  0.011593866
## weather    -0.128655201
## temp        0.394453645
## atemp       0.389784437
## humidity   -0.317371479
## windspeed   0.101369470
## casual      0.690413565
## registered  0.970948106
## count       1.000000000
```

```r
library(corrplot)

corrplot(cor.data,method = 'color')
```



Making the ggplot to really see how is count of bikes rented related to other factors. From the plot we see that the count of bike rented is higher when the temp is higher

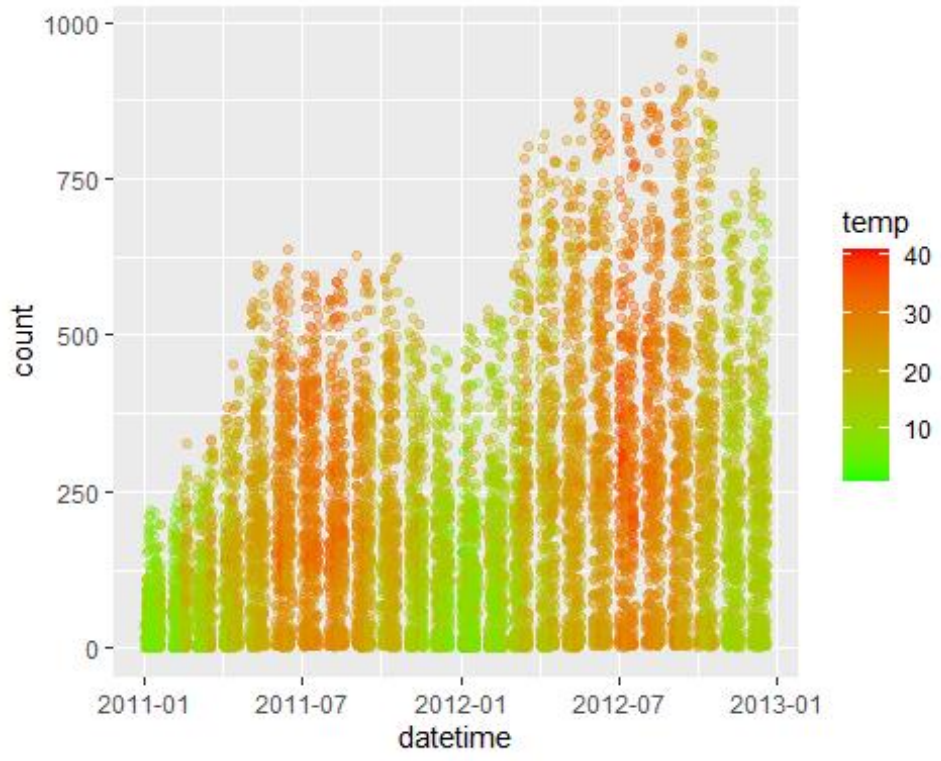
```r
ggplot(bike,aes(temp,count)) + geom_point(aes(color=temp),alpha=0.5)
```




```
bike$datetime <- as.POSIXct(bike$datetime)
```

Here we see that the bike rented is higher when its summer months

```
ggplot(bike,aes(datetime,count)) + geom_point(aes(color=temp),alpha=0.3) +
scale_color_continuous(low='green',high='red')
```

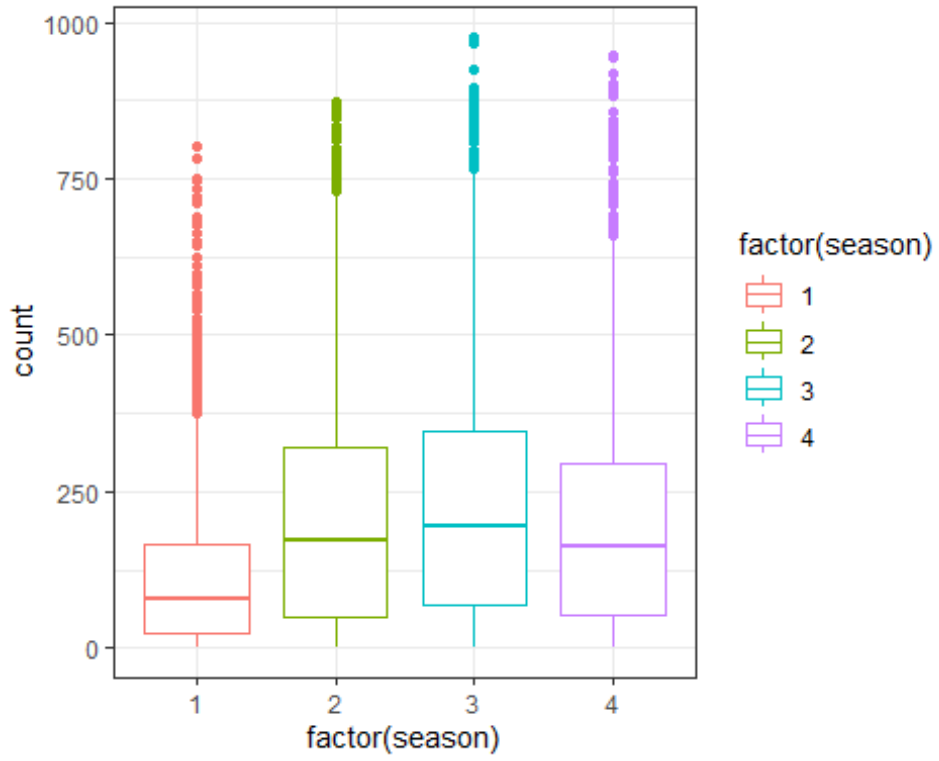


Now we using the correlation function to see how are temp and count related

```
cor(bike[,c('temp','count')])
```

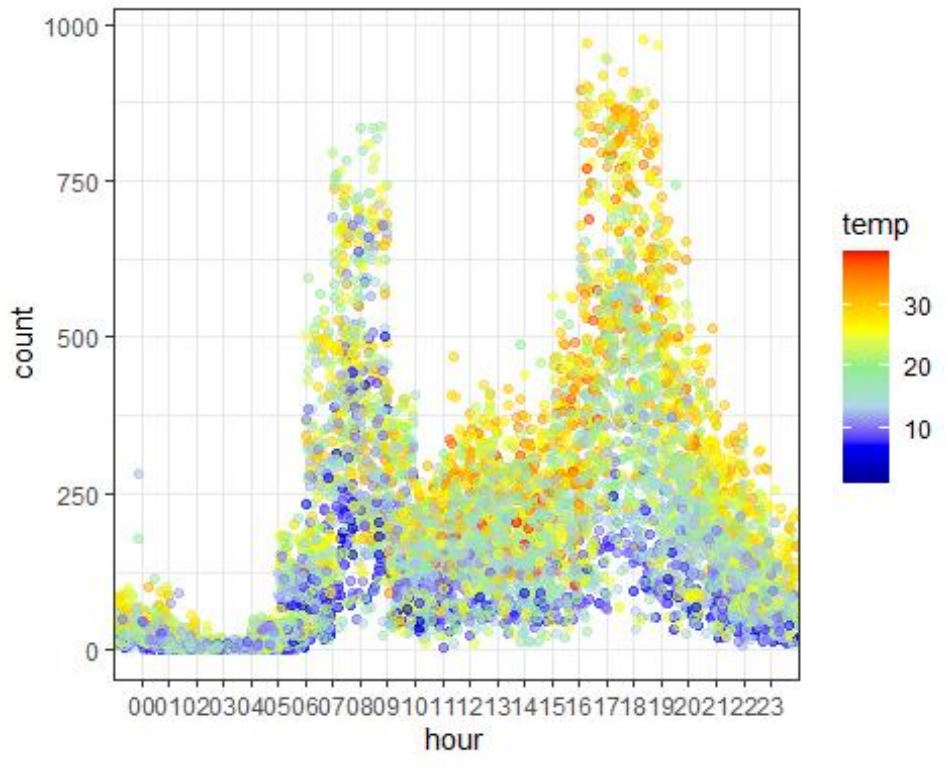
```
##            temp     count
## temp  1.0000000 0.3944536
## count 0.3944536 1.0000000
```

```
pl <- ggplot(bike,aes(factor(season),count)) +
geom_boxplot(aes(color=factor(season)))

pl + theme_bw()
```



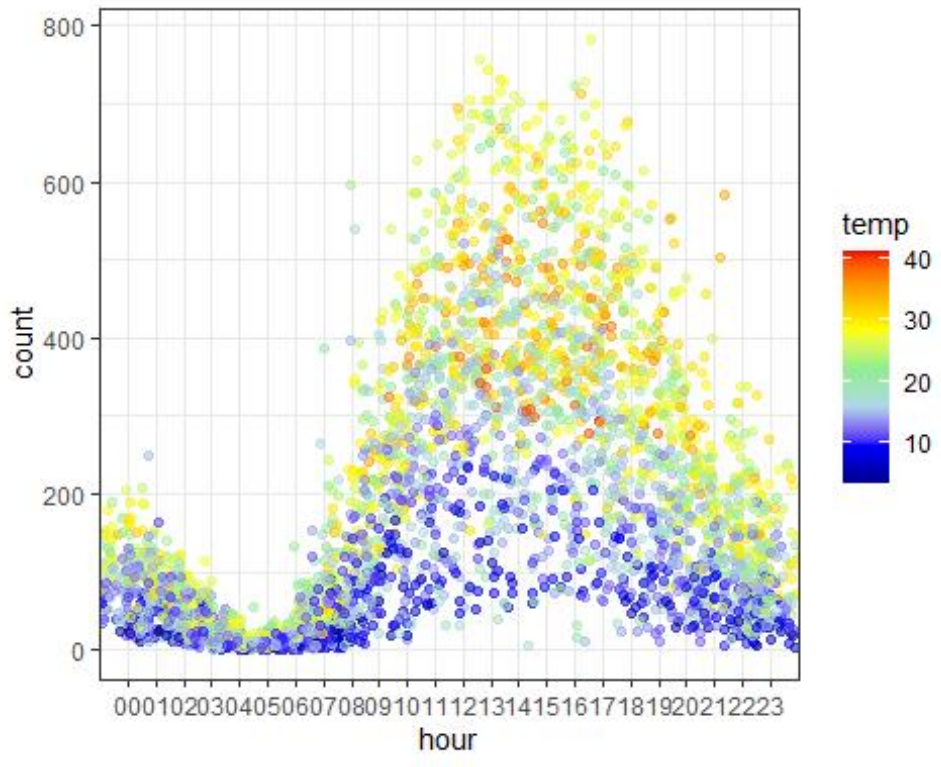
we see from the above ggplot that counts of bike rented are higher when its summer and fall

```
bike$hour <- sapply(bike$datetime,function(x){format(x,'%H')})

pl <- ggplot(filter(bike,workingday==1),aes(hour,count)) +
geom_point(position=position_jitter(w=1, h=0),aes(color=temp),alpha=0.5)

pl <- pl + scale_color_gradientn(colours = c('dark blue','blue','light
blue','light green','yellow','orange','red'))
pl + theme_bw()
```



from the ggplot we can deduce that rented bikes are higher during rush hour in the weekdays. Now we will see how is the rent of bike during the day in the weekends

```
pl2 <- ggplot(filter(bike,workingday==0),aes(hour,count)) +
geom_point(position=position_jitter(w=1, h=0),aes(color=temp),alpha=0.5)

pl2 <- pl2 + scale_color_gradientn(colours = c('dark blue','blue','light
blue','light green','yellow','orange','red'))

print(pl2 + theme_bw())
```



Enough with EDA and now we know for sure temp is major player in the bike rent. Lets go ahead and make the model based on the count and temp

```
model <- lm(count ~ temp,bike)

summary(model)

##
## Call:
## lm(formula = count ~ temp, data = bike)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -293.32 -112.36  -33.36   78.98  741.44
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.0462     4.4394   1.362    0.173
## temp          9.1705     0.2048  44.783   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 166.5 on 10884 degrees of freedom
## Multiple R-squared:  0.1556, Adjusted R-squared:  0.1555
## F-statistic:  2006 on 1 and 10884 DF,  p-value: < 2.2e-16
```

so we have our working model and if someone asks us how many bike will be rented when the temp is 25.We have two ways to answer that, one is by using the residual data and the other one is using predict function. I like more the predict way so we will do that here,know that our model is only made with two things in mind, count being affected by temp so if someone asks how many bike will be rented during weekends, then our model cant answer that, just to remember few things

```
temp.test <- data.frame(temp=c(25))

predict(model,temp.test)

##        1
## 235.3097
```

we have our answer, if the temp is 25 then the bikes rented are 235