Kaggle

Gaurav Bansal

Saturday, August 23, 2014

Data Fields

datetime - hourly date + timestamp
season - 1 = spring, 2 = summer, 3 = fall, 4 = winter holiday - whether the day is considered a holiday workingday - whether the day is neither a weekend nor holiday weather - 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 - temperature in Celsius atemp - "feels like" temperature in Celsius humidity - relative humidity windspeed - wind speed casual - number of non-registered user rentals initiated registered - number of registered user rentals initiated count - number of total rentals

```
bikejan <- read.csv("bikejan.csv")
bikejan$datetime <- as.POSIXlt(as.character(bikejan$datetime))
str(bikejan)
```

```
## 'data.frame':    456 obs. of  18 variables:
##  $ datetime  : POSIXlt, format: "0001-01-11 00:00:00" "0001-01-11 01:00:00" ...
##  $ season    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ holiday   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ workingday: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ weather   : int  1 1 1 1 1 2 1 1 1 1 ...
##  $ temp      : num  9.84 9.02 9.02 9.84 9.84 ...
##  $ atemp     : num  14.4 13.6 13.6 14.4 14.4 ...
##  $ humidity  : num  81 80 80 75 75 75 80 86 75 76 ...
##  $ windspeed : num  0 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 ...
##  $ count     : int  16 40 32 13 1 1 2 3 8 14 ...
##  $ year      : int  2011 2011 2011 2011 2011 2011 2011 2011 2011 2011 ...
##  $ month     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ date      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ hour      : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ day       : Factor w/ 7 levels "#N/A","1","2",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ flag      : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
summary(bikejan)
```

```
##     datetime                       season     holiday
##  Min.   :0001-01-11 00:00:00   Min.   :1   Min.   :0.0000
##  1st Qu.:0001-05-11 17:45:00   1st Qu.:1   1st Qu.:0.0000
##  Median :0001-10-11 11:30:00   Median :1   Median :0.0000
##  Mean   :0006-11-05 00:07:54   Mean   :1   Mean   :0.0526
##  3rd Qu.:0015-01-11 05:15:00   3rd Qu.:1   3rd Qu.:0.0000
##  Max.   :0019-01-11 23:00:00   Max.   :1   Max.   :1.0000
##
##    workingday       weather          temp           atemp
```

```
##  Min.   :0.000   Min.   :1.00   Min.   : 3.28   Min.   : 3.03
##  1st Qu.:0.000   1st Qu.:1.00   1st Qu.: 6.56   1st Qu.: 7.96
##  Median :1.000   Median :1.00   Median : 8.20   Median : 9.85
##  Mean   :0.632   Mean   :1.47   Mean   : 8.57   Mean   :10.66
##  3rd Qu.:1.000   3rd Qu.:2.00   3rd Qu.: 9.84   3rd Qu.:12.88
##  Max.   :1.000   Max.   :3.00   Max.   :18.86   Max.   :22.73
##
##     humidity       windspeed        casual        registered
##  Min.   : 28.0   Min.   : 0.0   Min.   : 0.00   Min.   :  0
##  1st Qu.: 44.0   1st Qu.: 9.0   1st Qu.: 0.00   1st Qu.: 13
##  Median : 53.0   Median :13.0   Median : 2.00   Median : 43
##  Mean   : 57.4   Mean   :13.9   Mean   : 4.66   Mean   : 50
##  3rd Qu.: 69.0   3rd Qu.:19.0   3rd Qu.: 6.00   3rd Qu.: 70
##  Max.   :100.0   Max.   :39.0   Max.   :47.00   Max.   :216
##                                 NA's   :25      NA's   :25
##      count            year          month         date          hour
##  Min.   :  1.0   Min.   :2011   Min.   :1   Min.   : 1   Min.   : 0.00
##  1st Qu.: 12.0   1st Qu.:2011   1st Qu.:1   1st Qu.: 5   1st Qu.: 5.75
##  Median : 44.0   Median :2011   Median :1   Median :10   Median :11.50
##  Mean   : 52.7   Mean   :2011   Mean   :1   Mean   :10   Mean   :11.50
##  3rd Qu.: 77.2   3rd Qu.:2011   3rd Qu.:1   3rd Qu.:15   3rd Qu.:17.25
##  Max.   :219.0   Max.   :2011   Max.   :1   Max.   :19   Max.   :23.00
##
##    day          flag
##  #N/A:72   Min.   :-23
##  1   :72   1st Qu.:  1
##  2   :72   Median :  1
##  4   :48   Mean   :  0
##  5   :48   3rd Qu.:  1
##  6   :72   Max.   :  1
##  7   :72
```

```
x <- 1:10
y <- 990:999
```

Univariate Analysis of Categorical Variables

1. Season

```
table(bikejan$season)/24
```

```
##
##  1
## 19
```

2. Holiday

```
table(bikejan$holiday)/24
```

```
##
##  0  1
## 18  1
```

```
bikejan[(bikejan$holiday==1),c(15,17)]
```

```
##     date day
## 385   17   1
## 386   17   1
## 387   17   1
## 388   17   1
## 389   17   1
## 390   17   1
## 391   17   1
## 392   17   1
## 393   17   1
## 394   17   1
## 395   17   1
## 396   17   1
## 397   17   1
## 398   17   1
## 399   17   1
## 400   17   1
## 401   17   1
## 402   17   1
## 403   17   1
## 404   17   1
## 405   17   1
## 406   17   1
## 407   17   1
## 408   17   1
```

3. Working Day

```
table(bikejan$workingday)/24
```

```
##
##  0  1
##  7 12
```

```
table(bikejan$day)/24
```

```
##
## #N/A    1    2    4    5    6    7
##    3    3    3    2    2    3    3
```

4. Weather

```
table(bikejan$weather)/24
```

```
##
##      1      2      3
## 11.458  6.167  1.375
```

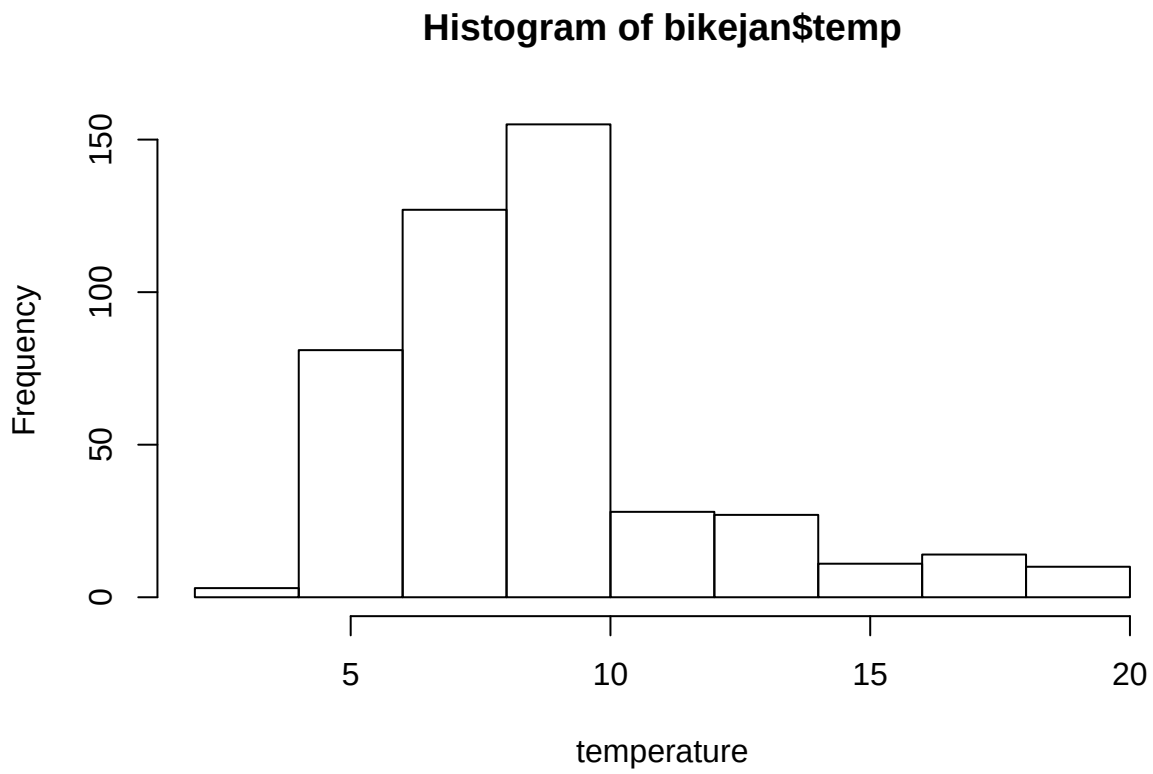
```
table(bikejan$weather)
```

```
##
##   1   2   3
## 275 148  33
```

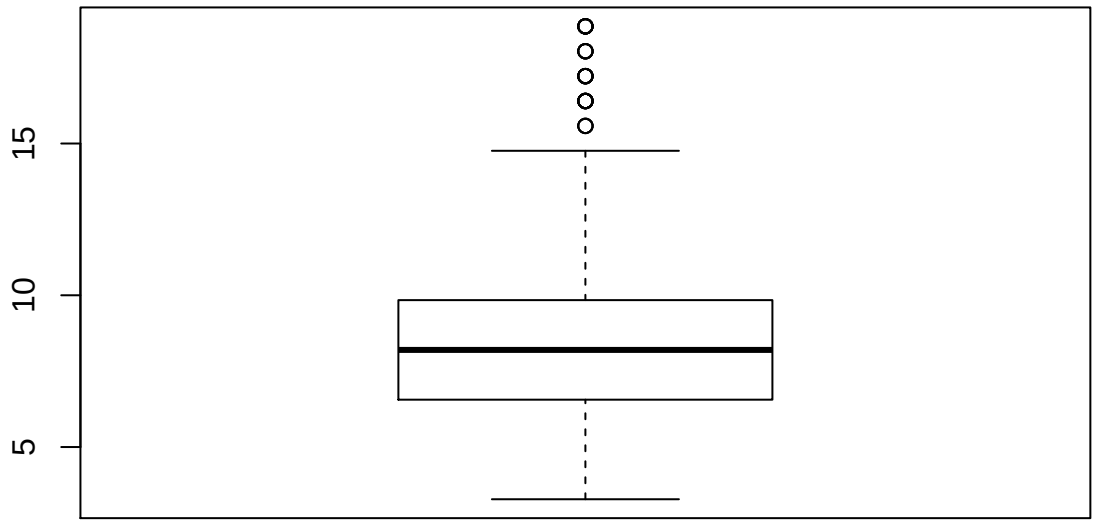
Univariate Analysis for continuous variables

1. temp

```
hist(bikejan$temp,xlab="temperature")
```




```
boxplot(bikejan$temp,xlab="temperature")
```



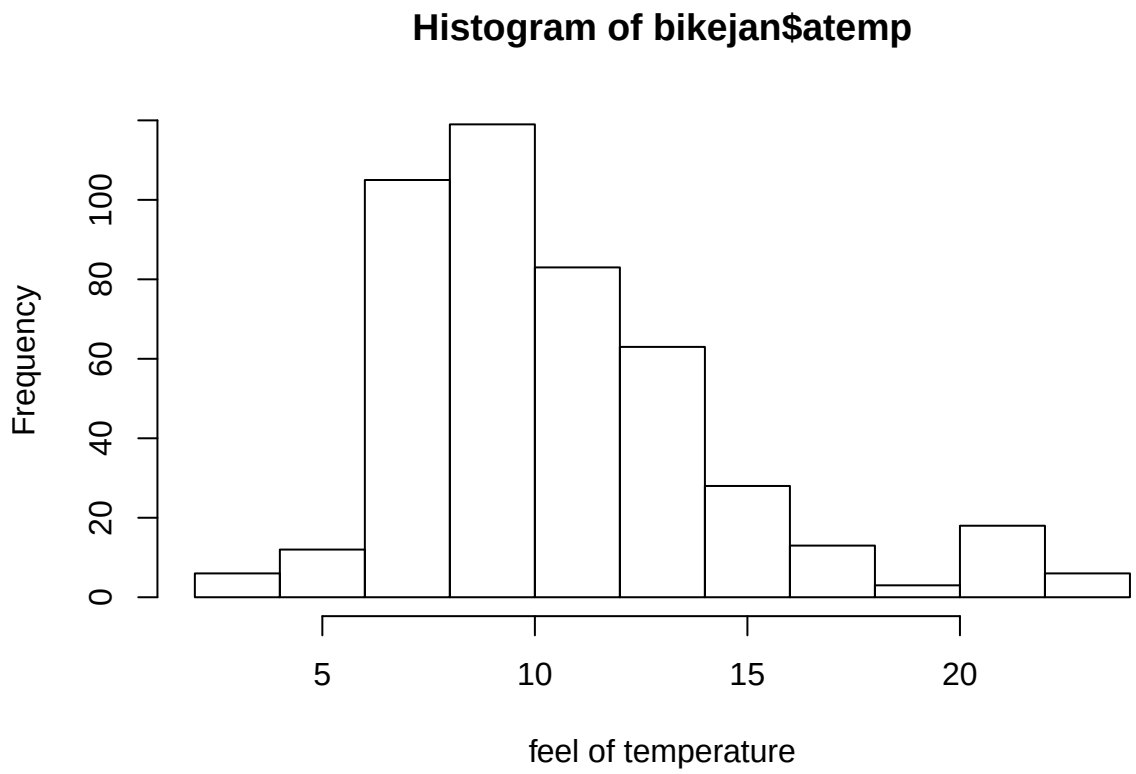
temperature

```
quantile(bikejan$temp,c(x/1000,0.05,0.1,0.2,0.25,0.3,0.4,0.5,0.6,0.7,0.75,0.8,0.9,0.95,0.96,0.97,0.98,0
```

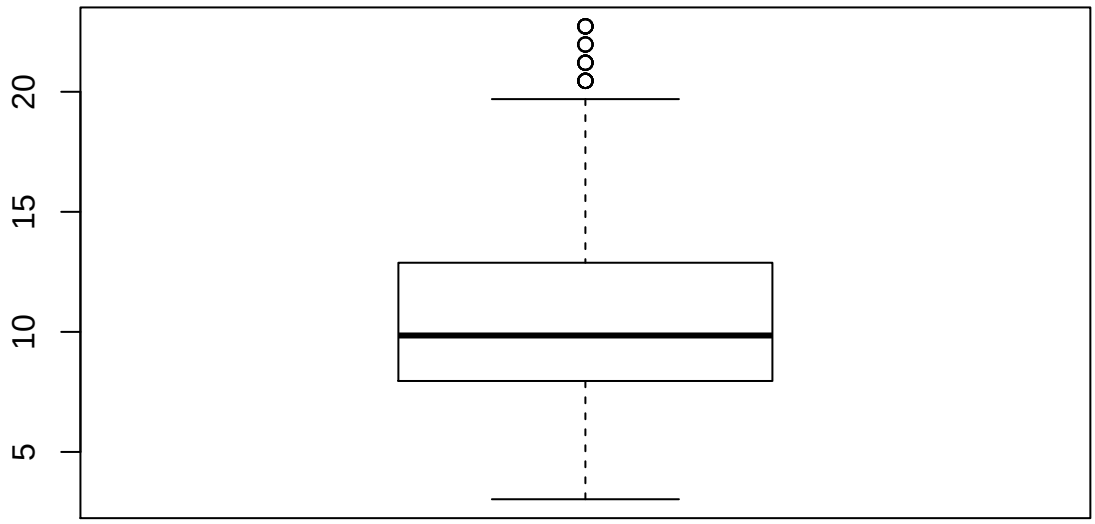
```
##   0.1%   0.2%   0.3%   0.4%   0.5%   0.6%   0.7%   0.8%   0.9%     1%
##  3.280  3.280  3.280  3.280  3.505  3.879  4.100  4.100  4.100  4.100
##     5%    10%    20%    25%    30%    40%    50%    60%    70%    75%
##  4.920  5.740  6.560  6.560  6.560  7.380  8.200  8.200  9.020  9.840
##    80%    90%    95%    96%    97%    98%    99%    99%  99.1%  99.2%
##  9.840 13.120 16.400 16.400 17.220 17.958 18.860 18.860 18.860 18.860
##  99.3%  99.4%  99.5%  99.6%  99.7%  99.8%  99.9%
## 18.860 18.860 18.860 18.860 18.860 18.860 18.860
```

2. atemp

```
hist(bikejan$atemp,xlab="feel of temperature")
```




```
boxplot(bikejan$atemp,xlab="feel of temperature")
```



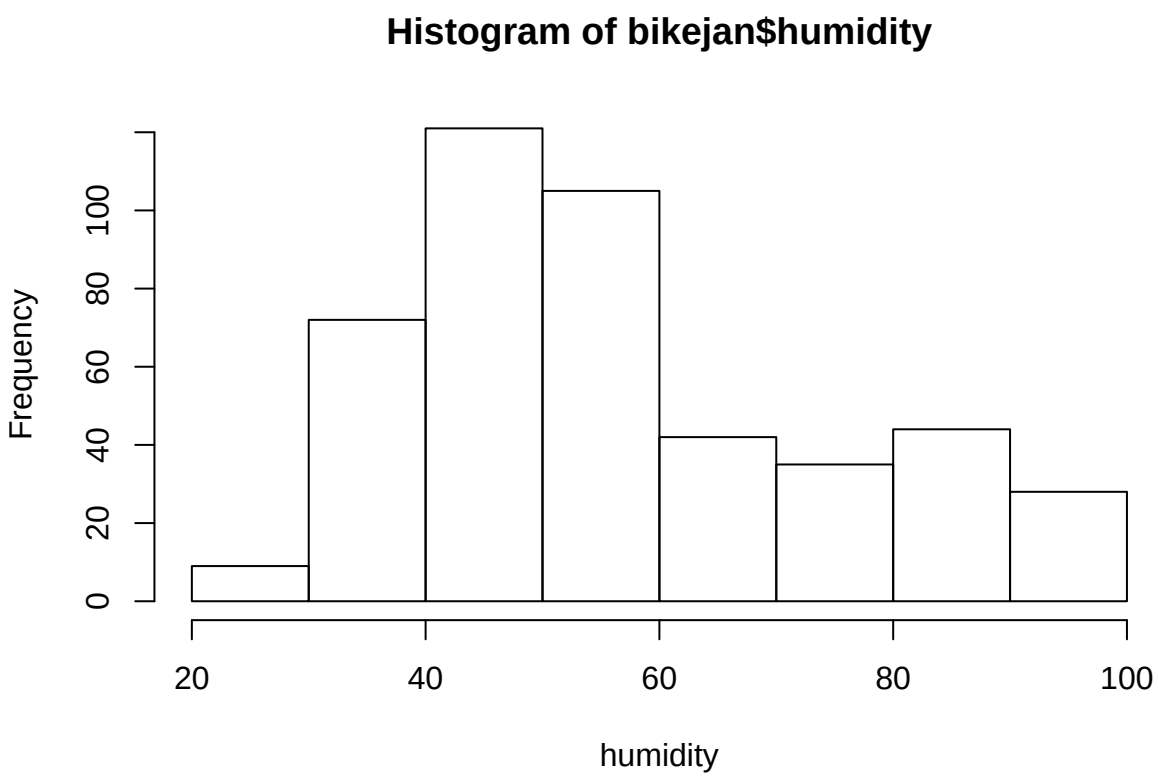
feel of temperature

```
quantile(bikejan$atemp,c(x/1000,0.05,0.1,0.2,0.25,0.3,0.4,0.5,0.6,0.7,0.75,0.8,0.9,0.95,0.96,0.97,0.98,
```

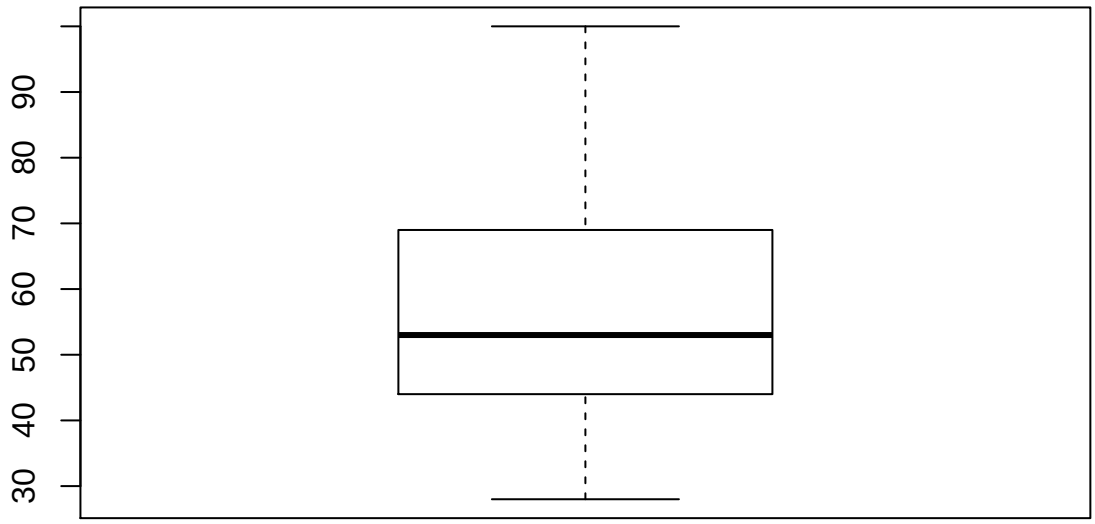
```
##   0.1%   0.2%   0.3%   0.4%   0.5%   0.6%   0.7%   0.8%   0.9%     1%
##  3.030  3.030  3.030  3.030  3.239  3.585  3.790  3.790  3.790  3.790
##     5%    10%    20%    25%    30%    40%    50%    60%    70%    75%
##  6.060  6.060  7.575  7.955  8.335  9.090  9.850 10.605 11.365 12.880
##    80%    90%    95%    96%    97%    98%    99%    99% 99.1% 99.2%
## 12.880 15.150 20.455 20.455 21.210 21.892 22.725 22.725 22.725 22.725
##  99.3%  99.4%  99.5%  99.6%  99.7%  99.8%  99.9%
## 22.725 22.725 22.725 22.725 22.725 22.725 22.725
```

3. humidity

```
hist(bikejan$humidity,xlab="humidity")
```




```
boxplot(bikejan$humidity,xlab="humidity")
```

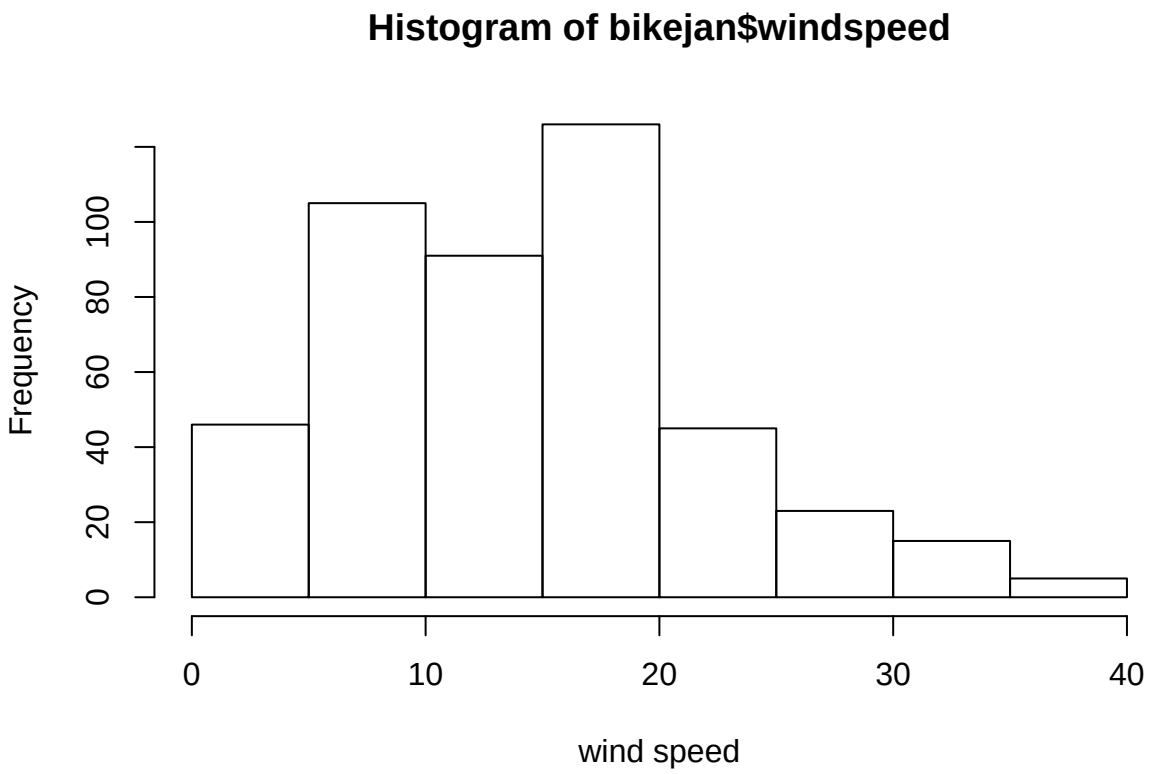



```
quantile(bikejan$humidity,c(x/1000,0.05,0.1,0.2,0.25,0.3,0.4,0.5,0.6,0.7,0.75,0.8,0.9,0.95,0.96,0.97,0.9
```

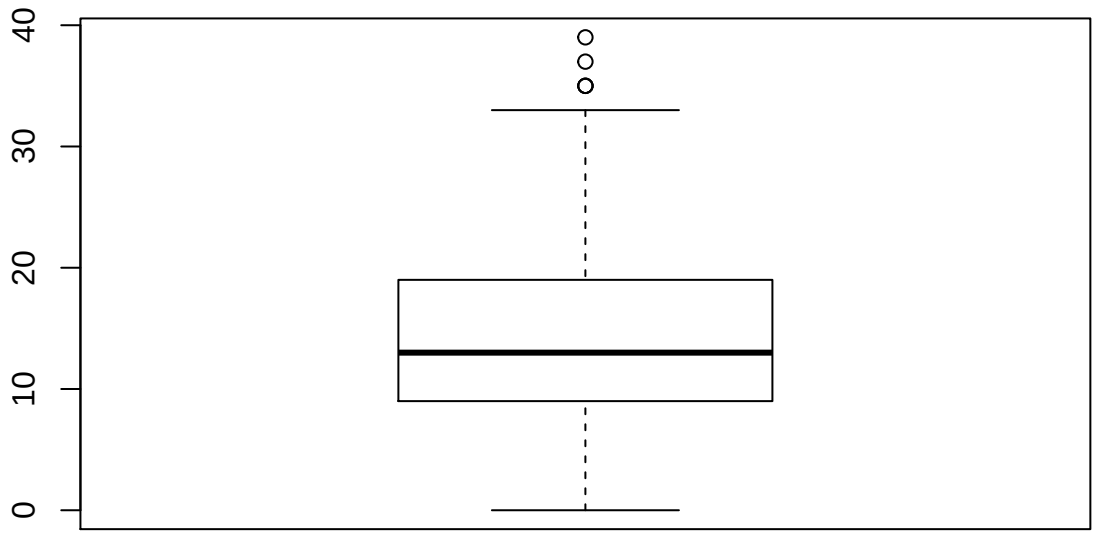
```
##  0.1%  0.2%  0.3%  0.4%  0.5%  0.6%  0.7%  0.8%  0.9%    1%    5%   10%
## 28.00 28.00 28.00 28.00 28.28 28.73 29.18 29.64 30.00 30.00 35.00 38.00
##   20%   25%   30%   40%   50%   60%   70%   75%   80%   90%   95%   96%
## 43.00 44.00 47.00 50.00 53.00 56.00 64.00 69.00 75.00 86.00 93.00 93.00
##   97%   98%   99%   99% 99.1% 99.2% 99.3% 99.4% 99.5% 99.6% 99.7% 99.8%
## 93.00 93.00 93.45 93.45 93.90 94.00 94.00 94.00 94.00 94.00 94.00 94.54
## 99.9%
## 97.27
```

4. Wind Speed

```
hist(bikejan$windspeed,xlab="wind speed")
```




```
boxplot(bikejan$windspeed,xlab="wind speed")
```



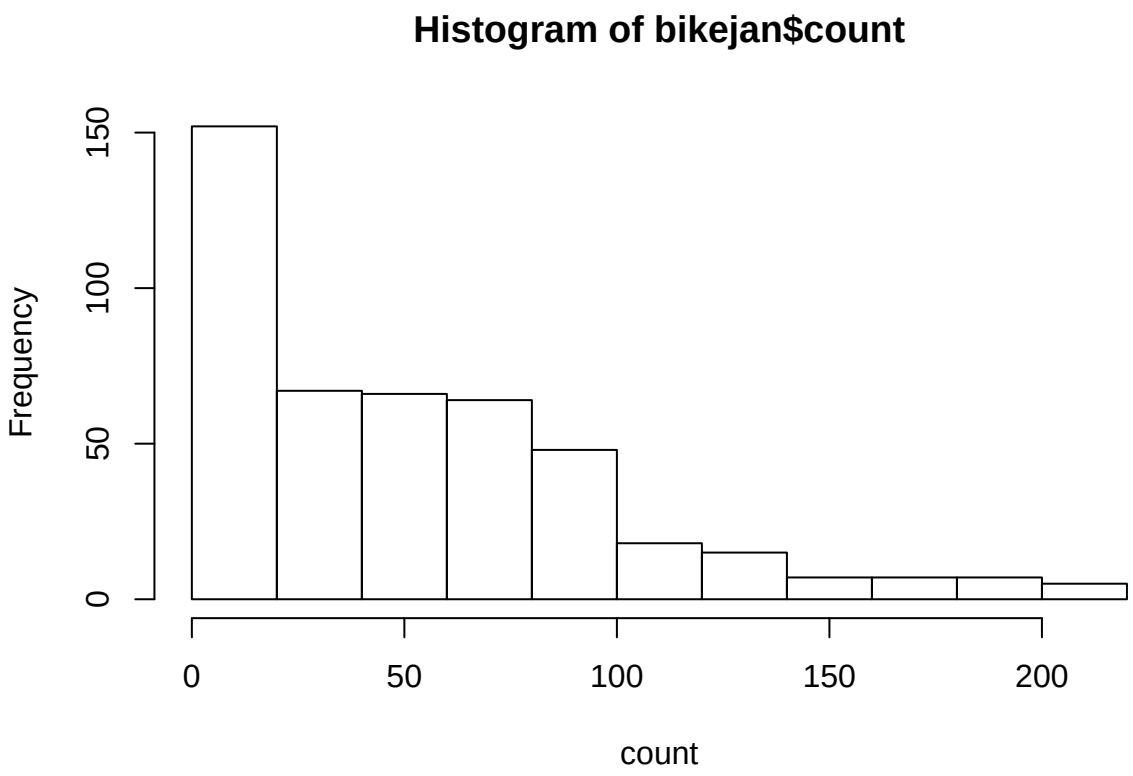
wind speed

```
quantile(bikejan$windspeed,c(x/1000,0.05,0.1,0.2,0.25,0.3,0.4,0.5,0.6,0.7,0.75,0.8,0.9,0.95,0.96,0.97,0
```

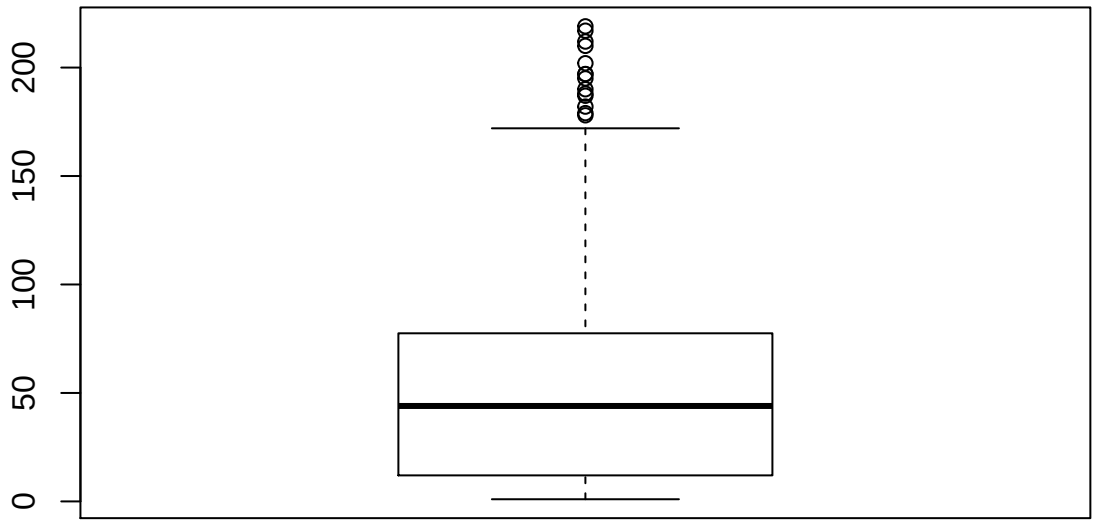
```
##   0.1%   0.2%   0.3%   0.4%   0.5%   0.6%   0.7%   0.8%   0.9%     1%
##  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000
##     5%    10%    20%    25%    30%    40%    50%    60%    70%    75%
##  0.000  4.502  7.002  8.998  8.998 11.001 12.998 15.001 19.001 19.001
##    80%    90%    95%    96%    97%    98%    99%    99%  99.1%  99.2%
## 20.000 23.999 27.999 30.003 30.003 30.901 33.899 33.899 34.810 35.001
##  99.3%  99.4%  99.5%  99.6%  99.7%  99.8%  99.9%
## 35.001 35.001 35.001 35.360 36.269 37.178 38.089
```

5. Count

```
hist(bikejan$count,xlab="count")
```




```
boxplot(bikejan$count,xlab="count")
```



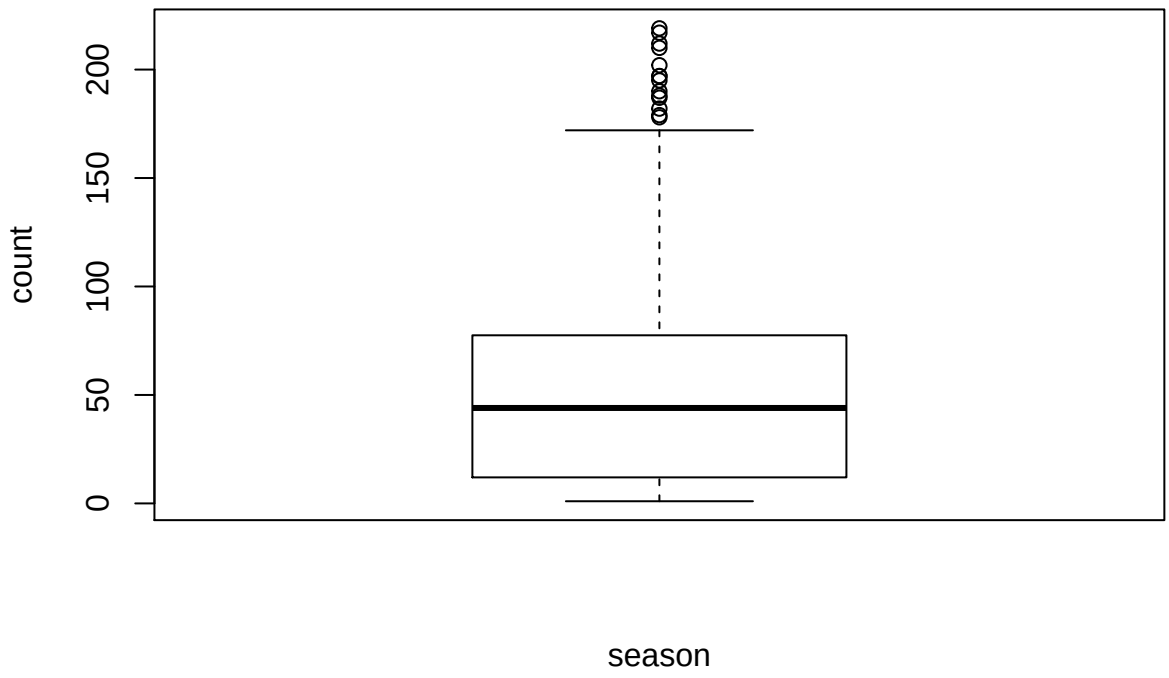
count

```
quantile(bikejan$count,c(x/1000,0.05,0.1,0.2,0.25,0.3,0.4,0.5,0.6,0.7,0.75,0.8,0.9,0.95,0.96,0.97,0.98,
```

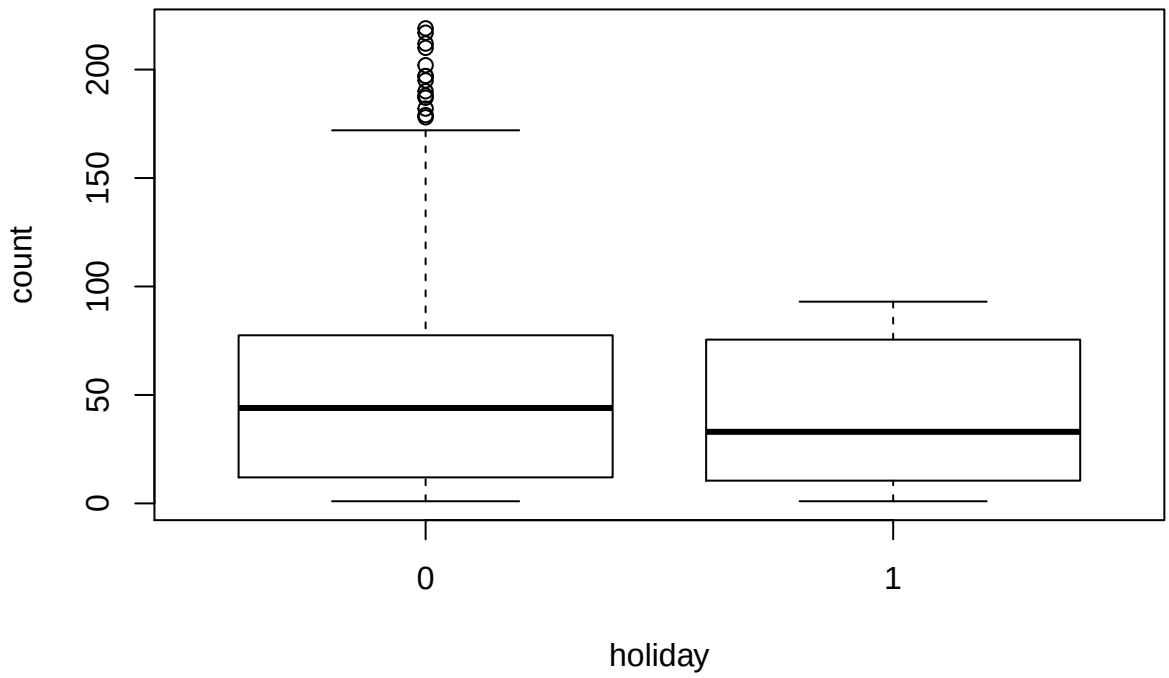
```
##   0.1%   0.2%   0.3%   0.4%   0.5%   0.6%   0.7%   0.8%   0.9%     1%
##   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00
##     5%    10%    20%    25%    30%    40%    50%    60%    70%    75%
##   1.00   3.00   6.00  12.00  17.00  32.00  44.00  57.00  71.00  77.25
##    80%    90%    95%    96%    97%    98%    99%    99%  99.1%  99.2%
##  86.00 114.00 155.50 160.60 174.10 187.90 199.25 199.25 201.52 204.88
##  99.3%  99.4%  99.5%  99.6%  99.7%  99.8%  99.9%
## 208.52 210.54 211.45 212.90 215.17 217.18 218.09
```

Bivariate Analysis with categorical variables

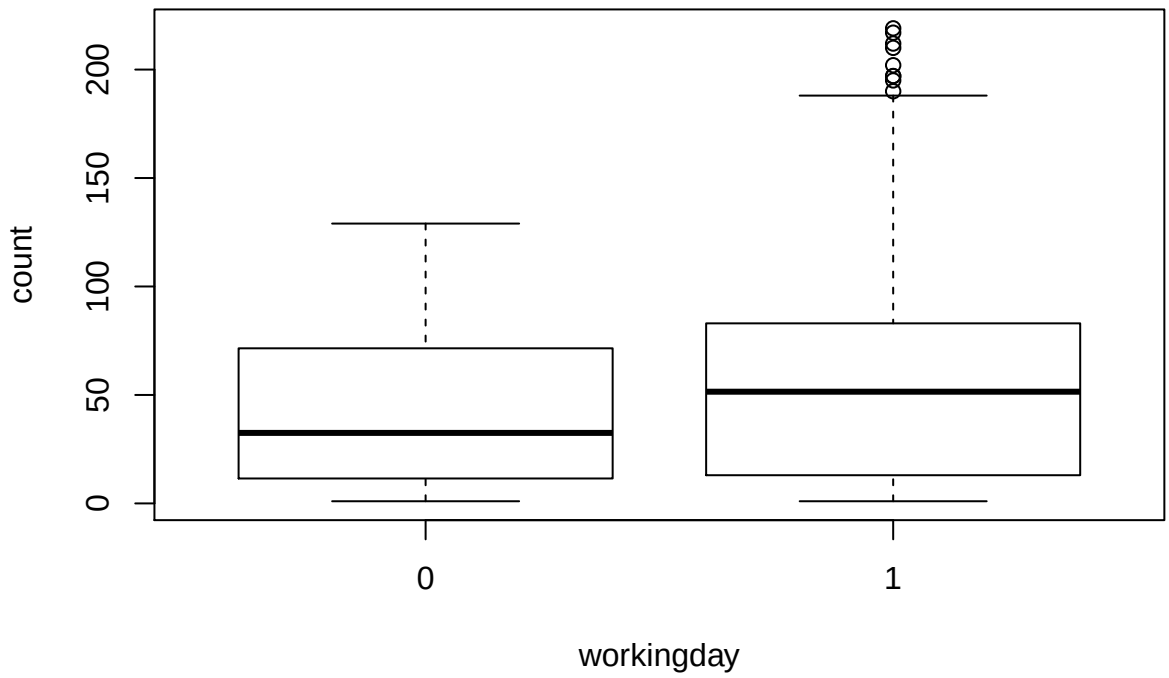
```
boxplot(count~season,bikejan,xlab="season",ylab="count")
```



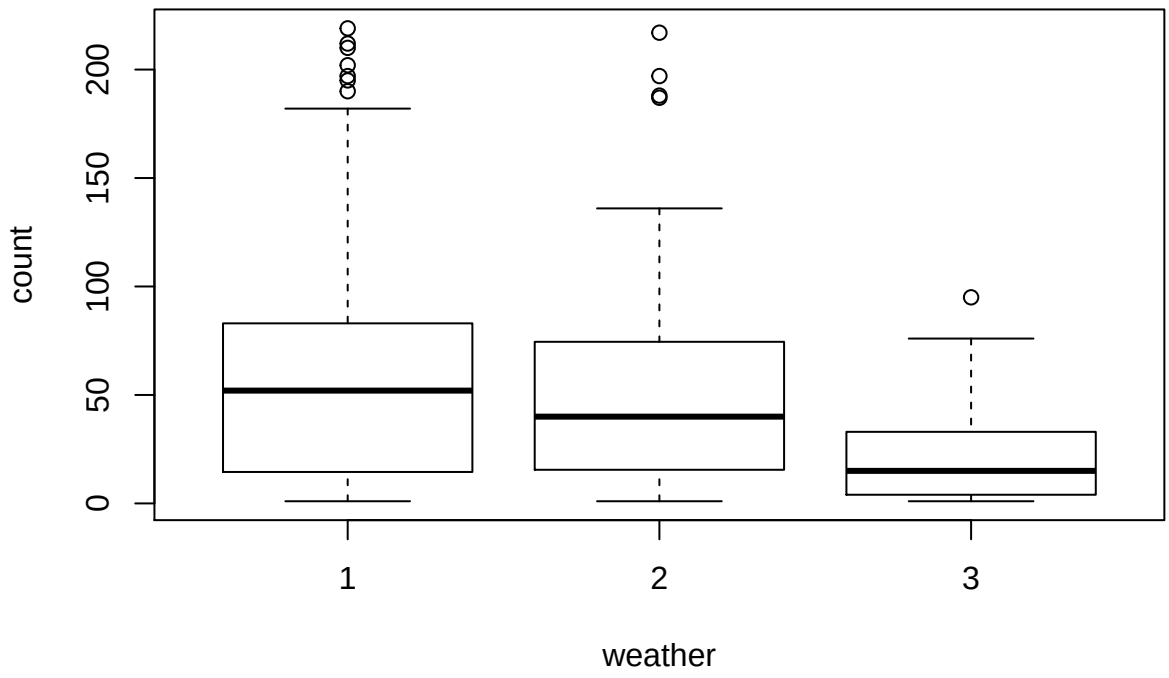

```
boxplot(count~holiday,bikejan,xlab="holiday",ylab="count")
```




```
boxplot(count~workingday,bikejan,xlab="workingday",ylab="count")
```

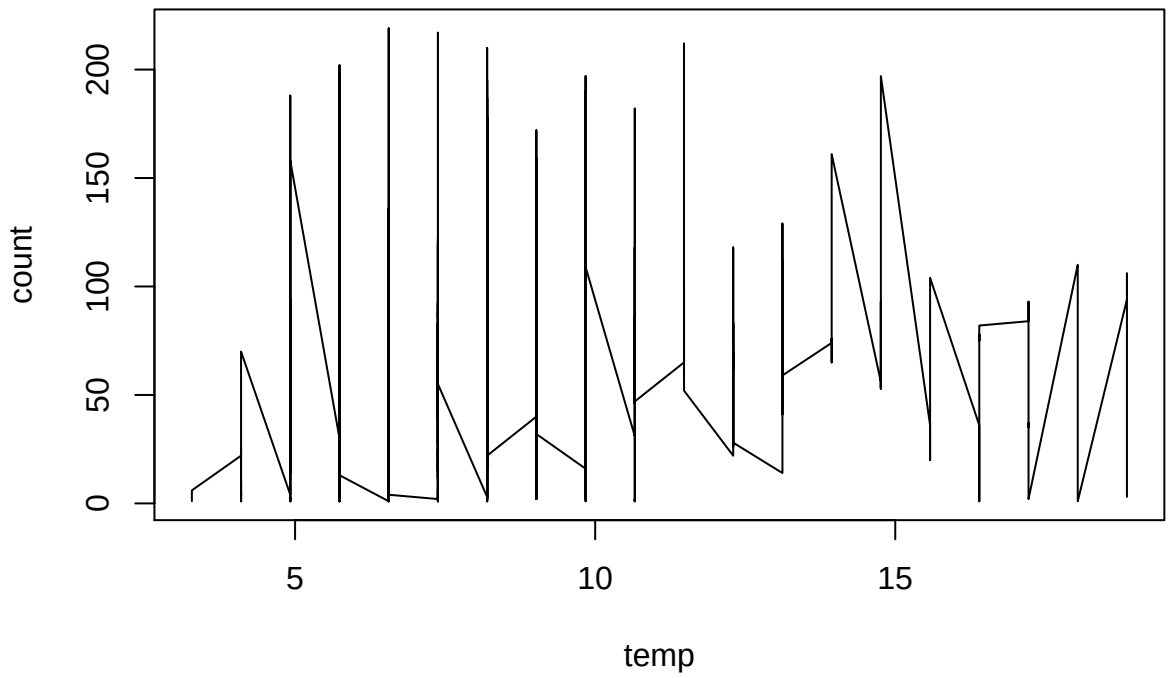



```
boxplot(count~weather,bikejan,xlab="weather",ylab="count")
```

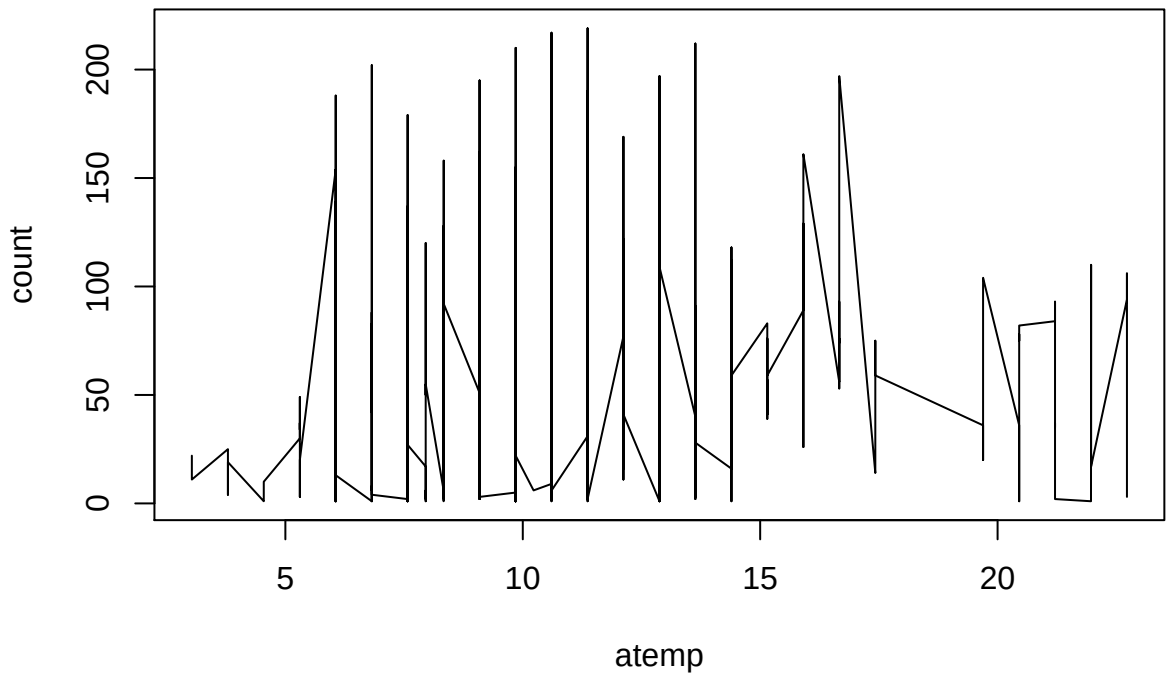


Bivariate Analysis with continuous variables

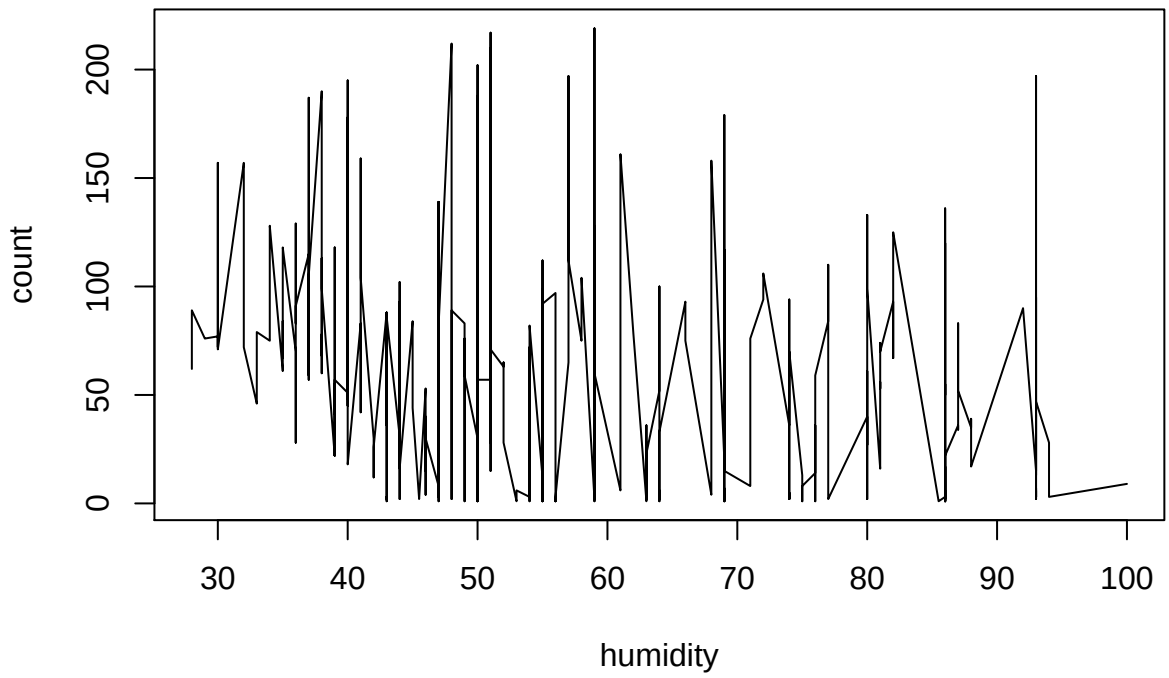
```
plot(count~temp,bikejan[order(bikejan$temp),],type="l",xlab="temp",ylab="count")
```



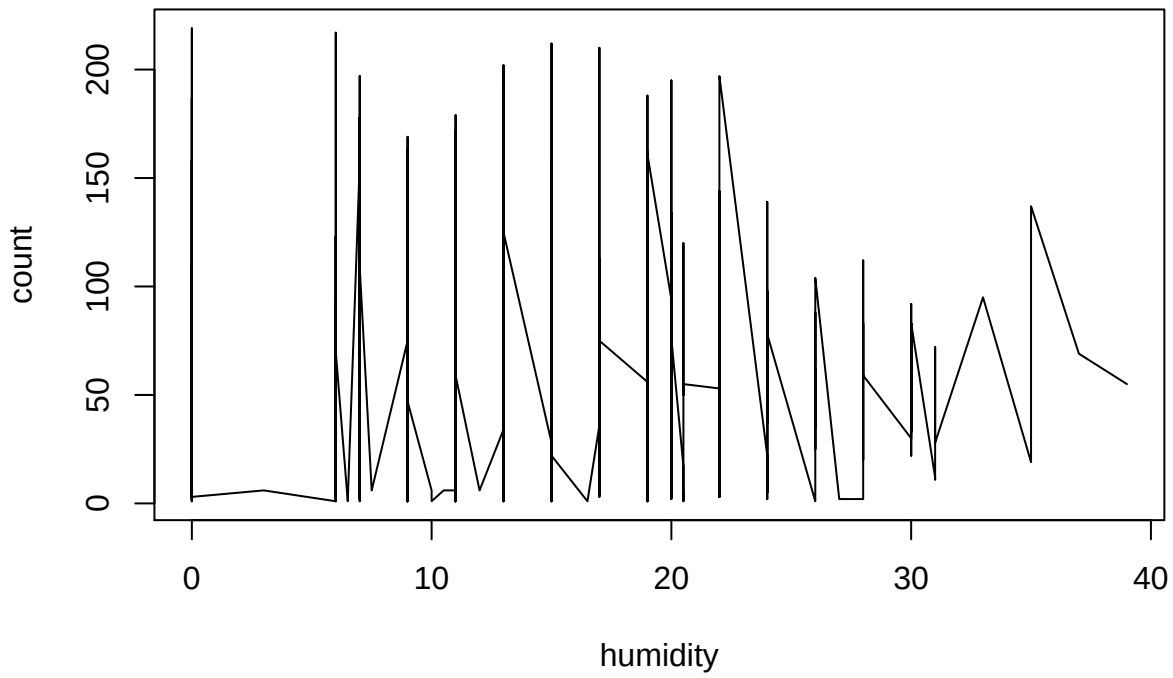

```
plot(count~atemp,bikejan[order(bikejan$atemp),],type="l",xlab="atemp",ylab="count")
```




```
plot(count~humidity,bikejan[order(bikejan$humidity),],type="l",xlab="humidity",ylab="count")
```




```
plot(count~windspeed,bikejan[order(bikejan$windspeed),],type="l",xlab="humidity",ylab="count")
```



Time series Analysis

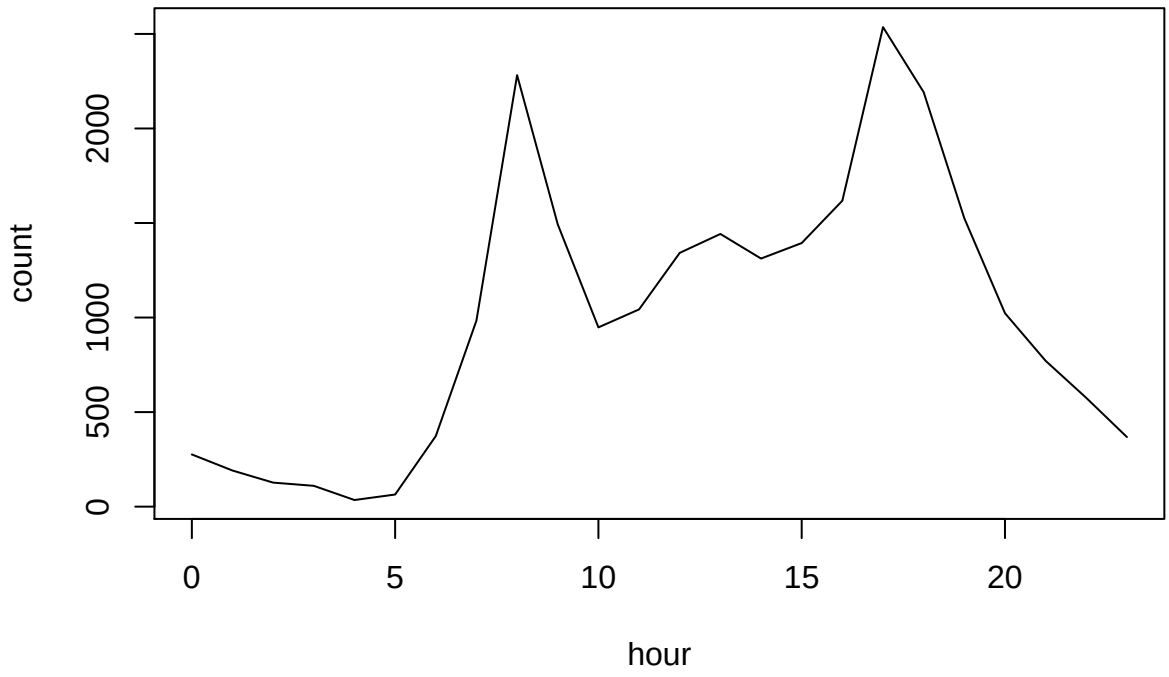
---

1. Hour

```
bikejan_hour <- aggregate(count~hour,bikejan,sum)
head(bikejan_hour[order(-bikejan_hour$count),],4)
```

```
##    hour count
## 18   17  2536
## 9     8  2282
## 19   18  2192
## 17   16  1618
```

```
plot(count~hour,bikejan_hour,type="l")
```

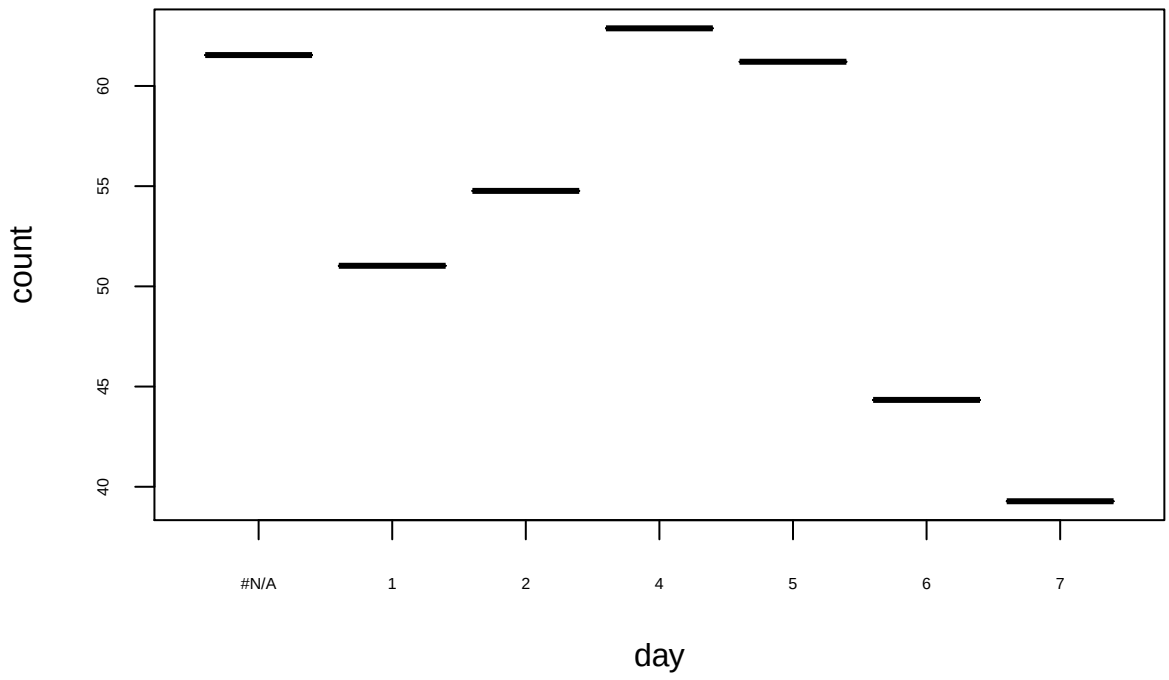


2. Day

```
bikejan_day <- aggregate(count~day,bikejan,mean)
bikejan_day[order(-bikejan_day$count),]
```

```
##    day count
## 4    4 62.88
## 1 #N/A 61.54
## 5    5 61.21
## 3    2 54.76
## 2    1 51.03
## 6    6 44.33
## 7    7 39.28
```

```
plot(count~day,bikejan_day,cex.axis=0.50)
```



3. Date

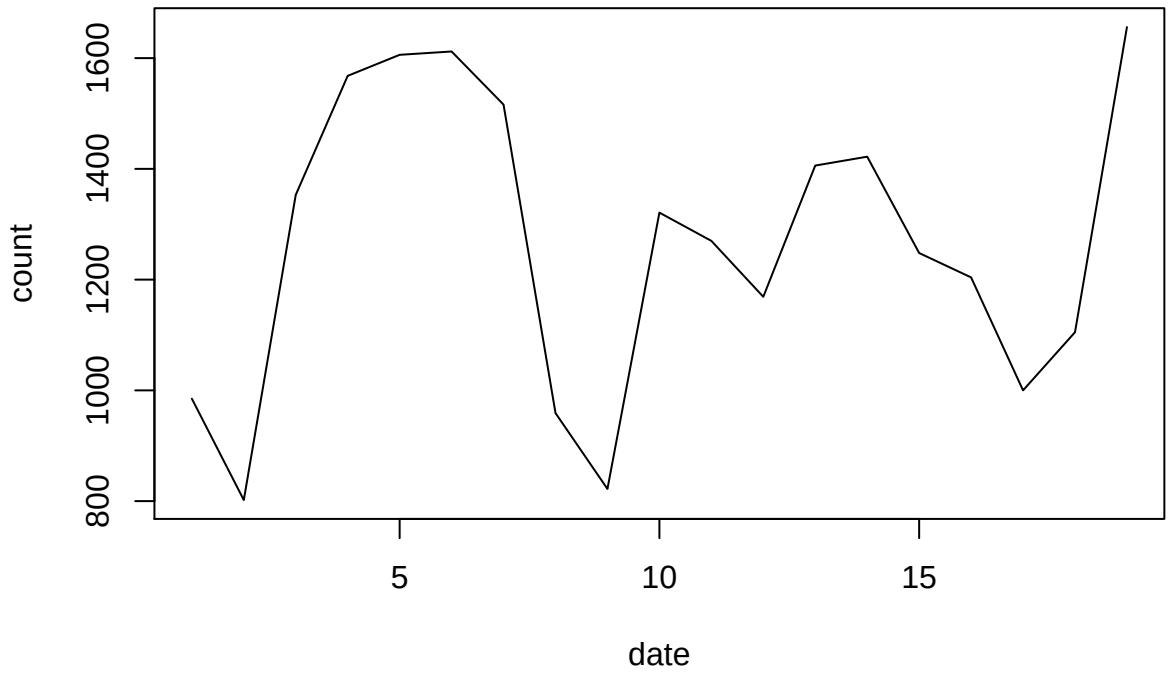
```
bikejan_date <- aggregate(count~date,bikejan,sum)
head(bikejan_date[order(-bikejan_date$count),])
```

```
##    date count
## 19   19  1656
## 6     6  1612
## 5     5  1606
## 4     4  1568
## 7     7  1516
## 14   14  1422
```

```
tail(bikejan_date[order(-bikejan_date$count),])
```

```
##    date count
## 18   18  1105
## 17   17  1000
## 1     1   985
## 8     8   959
## 9     9   822
## 2     2   802
```

```
plot(count~date,bikejan_date,type="l")
```



Correlation

```
cor(bikejan[,-c(1,18,17,6,15,14,13,11,10)])
```

```
## Warning: the standard deviation is zero
```

```
##            season  holiday workingday  weather   atemp humidity windspeed
## season          1       NA         NA       NA      NA       NA        NA
## holiday        NA  1.00000   -0.30861  0.26193 -0.1108 -0.04875  -0.02607
## workingday     NA -0.30861    1.00000 -0.14604 -0.2322  0.01107  -0.11545
## weather        NA  0.26193   -0.14604  1.00000  0.2118  0.53104  -0.14539
## atemp          NA -0.11085   -0.23221  0.21185  1.0000  0.27018  -0.21568
## humidity       NA -0.04875    0.01107  0.53104  0.2702  1.00000  -0.32051
## windspeed      NA -0.02607   -0.11545 -0.14539 -0.2157 -0.32051   1.00000
## count          NA -0.05473    0.17542 -0.17627  0.1408 -0.26894   0.08240
## hour           NA  0.00000    0.00000 -0.05503  0.1437 -0.20945   0.14173
##               count     hour
## season           NA       NA
## holiday    -0.05473  0.00000
## workingday  0.17542  0.00000
```

```
## weather    -0.17627 -0.05503
## atemp       0.14076  0.14369
## humidity   -0.26894 -0.20945
## windspeed   0.08240  0.14173
## count       1.00000  0.37426
## hour        0.37426  1.00000
```

Summary of EDA

1. Spring season whole of january
2. 17th of January was a holiday and a Monday
3. 3 saturdays, 3 sundays and 1 Monday were holidays
4. No extreme weathers, even light rains are found only in 33 observations
5. feel of temeprature is greater than actual temp.
6. only one season throughout january
7. people rent more bikes when their are no holidays but there was only 1 day of holiday so this may not be correct metric to show
8. people rent more bikes on working days than on holidays/saturday/sundays#clearly, weather has a role to play for people to rent bike. '3' depics rainy weather hence less bikes, '1' depicts clear weather hence more bikes, '2' is misty.#more bikes are rented when temp between 5-10, humidity between 40-60 and windspeed between 8-22# more bikes are in hours 8 AM and 5,6 PM # clearly there is a dip in the values on holidays

Model

```
#converting all categorical variables as factor variables
bikejan$season <- as.factor(bikejan$season)
bikejan$holiday <- as.factor(bikejan$holiday)
bikejan$workingday <- as.factor(bikejan$workingday)
bikejan$hour <- as.numeric(bikejan$hour)
bikejan$day <- as.factor(bikejan$day)
#weather is taken as numeric variable, as there is a value in the prediction data set with weather=4, t
bikejan$weather <- as.numeric(bikejan$weather)
```

```
#let us put all the variables in the model
library(car)
```

```
## Warning: package 'car' was built under R version 3.1.1
```

```
library(MASS)
model <- lm(count~ holiday+workingday+weather+temp+atemp+humidity+windspeed+hour+day, bikejan)
#vif(model)
# using this as model, it shows an ERROR :"Error in vif.default(model) : there are aliased coefficients
#hence, i drop day from the model, as it may have a perfect multicollinearity with the workingday
model <- lm(count~ holiday+workingday+weather+temp+atemp+humidity+windspeed+hour, bikejan)
vif(model)
```

```
##    holiday workingday    weather       temp      atemp   humidity
##      1.314      1.268      1.651     55.984     58.508      1.739
##  windspeed       hour
##      6.621      1.129
```

```r
#VIF of temp and atemp is very high, they are collinear variables, so we drop atemp, as it has high val
model <- lm(count~ holiday+workingday+weather+temp+humidity+windspeed+hour, bikejan)
vif(model)
```

```
##    holiday workingday    weather       temp   humidity  windspeed
##      1.297      1.258      1.622      1.274      1.700      1.158
##       hour
##      1.127
```

```r
#The VIF of all variable is less than 2, hence we consider al the variables for the model, let us ty to
model_bike <- stepAIC(model,direction= "both")
```

```
## Start:  AIC=3397
## count ~ holiday + workingday + weather + temp + humidity + windspeed +
##     hour
##
##              Df Sum of Sq    RSS  AIC
## - windspeed   1      1017 758524 3396
## - holiday     1      2443 759950 3397
## - weather     1      3235 760742 3397
## <none>                    757507 3397
## - humidity    1     30752 788259 3414
## - workingday  1     51185 808692 3425
## - temp        1     52956 810463 3426
## - hour        1     69248 826755 3435
##
## Step:  AIC=3396
## count ~ holiday + workingday + weather + temp + humidity + hour
##
##              Df Sum of Sq    RSS  AIC
## - holiday     1      2666 761190 3396
## - weather     1      3307 761831 3396
## <none>                    758524 3396
## + windspeed   1      1017 757507 3397
## - humidity    1     30066 788590 3412
## - temp        1     52027 810551 3424
## - workingday  1     53113 811637 3425
## - hour        1     68569 827093 3433
##
## Step:  AIC=3396
## count ~ workingday + weather + temp + humidity + hour
##
##              Df Sum of Sq    RSS  AIC
## - weather     1      1856 763046 3395
## <none>                    761190 3396
## + holiday     1      2666 758524 3396
## + windspeed   1      1240 759950 3397
## - humidity    1     34430 795620 3414
## - temp        1     49362 810552 3422
## - workingday  1     51005 812195 3423
## - hour        1     69359 830550 3433
##
## Step:  AIC=3395
```

```
## count ~ workingday + temp + humidity + hour
##
##              Df Sum of Sq    RSS  AIC
## <none>                     763046 3395
## + weather     1      1856 761190 3396
## + windspeed   1      1234 761813 3396
## + holiday     1      1216 761831 3396
## - temp        1     48443 811490 3421
## - workingday  1     55484 818531 3425
## - humidity    1     59537 822583 3427
## - hour        1     68358 831404 3432
```

```
model_bike$anova
```

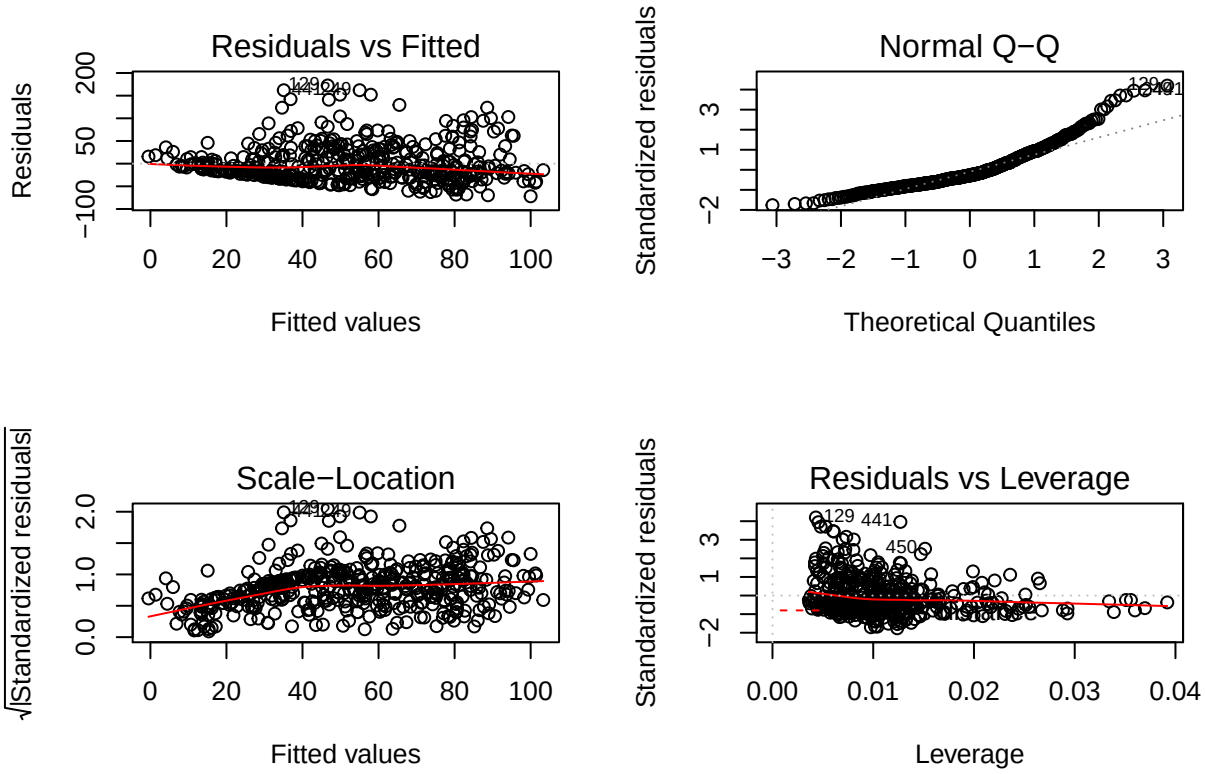
```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## count ~ holiday + workingday + weather + temp + humidity + windspeed +
##     hour
##
## Final Model:
## count ~ workingday + temp + humidity + hour
##
##
##          Step Df Deviance Resid. Df Resid. Dev  AIC
## 1                                448     757507 3397
## 2 - windspeed  1     1017       449     758524 3396
## 3   - holiday  1     2666       450     761190 3396
## 4   - weather  1     1856       451     763046 3395
```

```
#using stepwise regression it has dropped the 'holiday' variable. When cheked with the initial model, h
summary(model_bike)
```

```
##
## Call:
## lm(formula = count ~ workingday + temp + humidity + hour, data = bikejan)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -72.0  -27.6  -11.3   19.7  172.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   24.732      9.365    2.64   0.0086 **
## workingday1   23.871      4.168    5.73  1.9e-08 ***
## temp           3.571      0.667    5.35  1.4e-07 ***
## humidity      -0.684      0.115   -5.93  6.0e-09 ***
## hour           1.872      0.294    6.36  5.1e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.1 on 451 degrees of freedom
```

```
## Multiple R-squared:  0.257,  Adjusted R-squared:  0.25
## F-statistic:   39 on 4 and 451 DF,  p-value: <2e-16
```

```
par(mfrow=c(2,2))
plot(model_bike)
```




```
#input test data set
testjan <- read.csv("testjan.csv")
#changing variables
testjan$season <- as.factor(testjan$season)
testjan$holiday <- as.factor(testjan$holiday)
testjan$workingday <- as.factor(testjan$workingday)
testjan$hour <- as.numeric(testjan$hour)
testjan$weather <- as.numeric(testjan$weather)

#predicting count values from the model
predict_test <- predict(model,testjan)

predict_values <- cbind(testjan,ceiling(predict_test))
write.csv(predict_values,"output.csv")
```