Widget Use Report

1. Summarize and Interpret Data

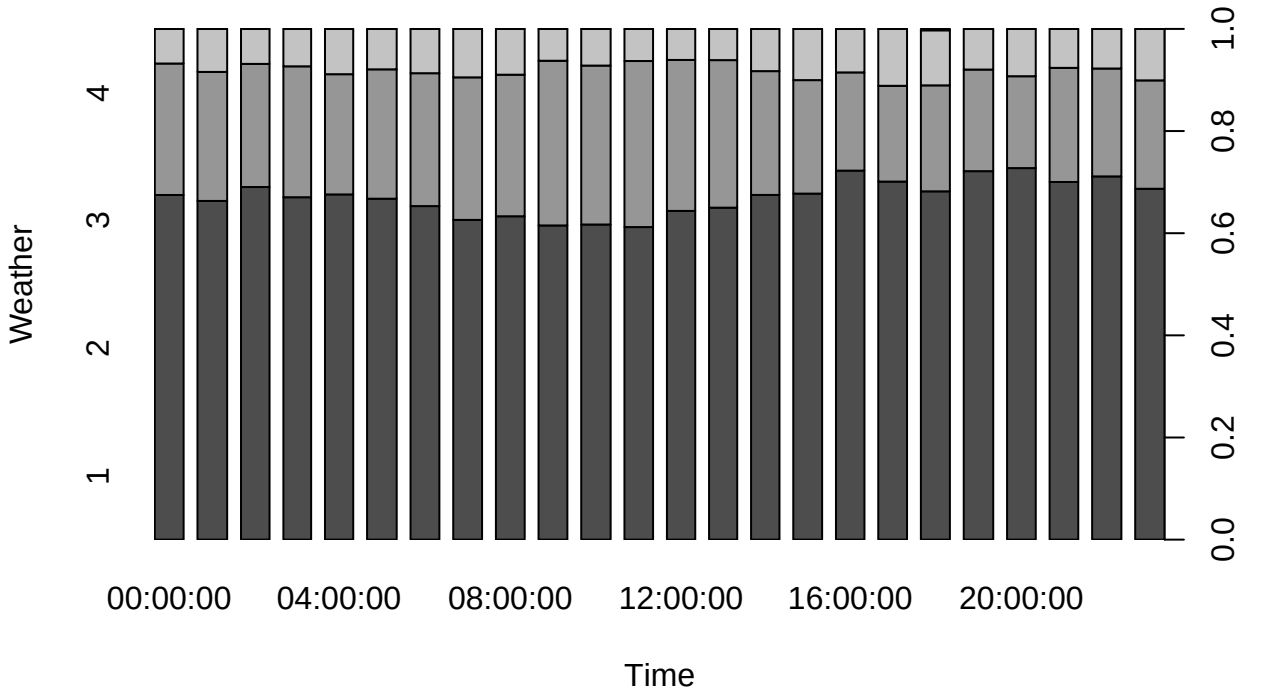
a

```
##   index           time_date quarter vacation workday weather temp_c
## 1     1 2011-01-01 00:00:00       1        0       0       1   9.84
## 2     2 2011-01-01 01:00:00       1        0       0       1   9.02
## 3     5 2011-01-01 04:00:00       1        0       0       1   9.84
## 4     6 2011-01-01 05:00:00       1        0       0       2   9.84
## 5     7 2011-01-01 06:00:00       1        0       0       1   9.02
## 6     8 2011-01-01 07:00:00       1        0       0       1   8.20
##   humidity count year   month      day     hour
## 1       81    16 2011 January Saturday 00:00:00
## 2       80    40 2011 January Saturday 01:00:00
## 3       75     1 2011 January Saturday 04:00:00
## 4       75     1 2011 January Saturday 05:00:00
## 5       80     2 2011 January Saturday 06:00:00
## 6       86     3 2011 January Saturday 07:00:00
```

b

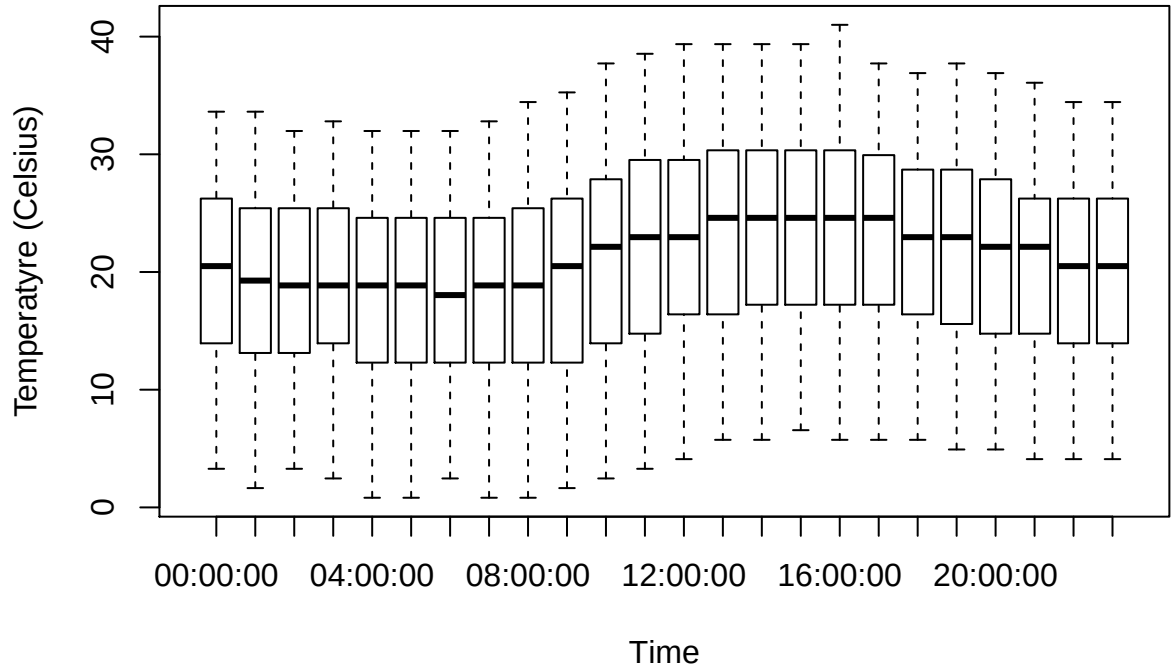
There is no indication of a relationship between the time and weather type. Weather types seem to be uniformly distributed.

```
plot(df$hour,df$weather, xlab= "Time", ylab="Weather")
```



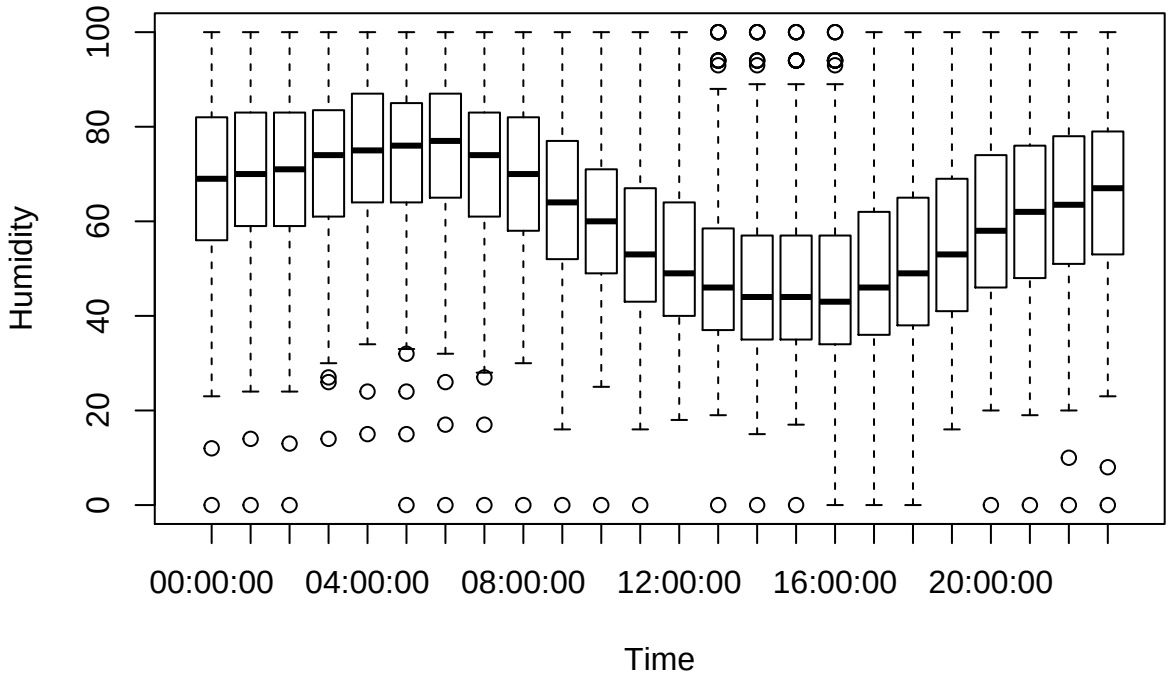
There is a point that corresponds to a temperature of -200. We will omit it from the graphical analysis, but requires further investigation in subsection **e**. The temperature fluctuates from low to high values in a non-linear fashion. The relationship is sinusoidal.

```
plot(df$hour, df$temp_c, xlab= "Time", ylab="Temperatyre (Celsius)")
```



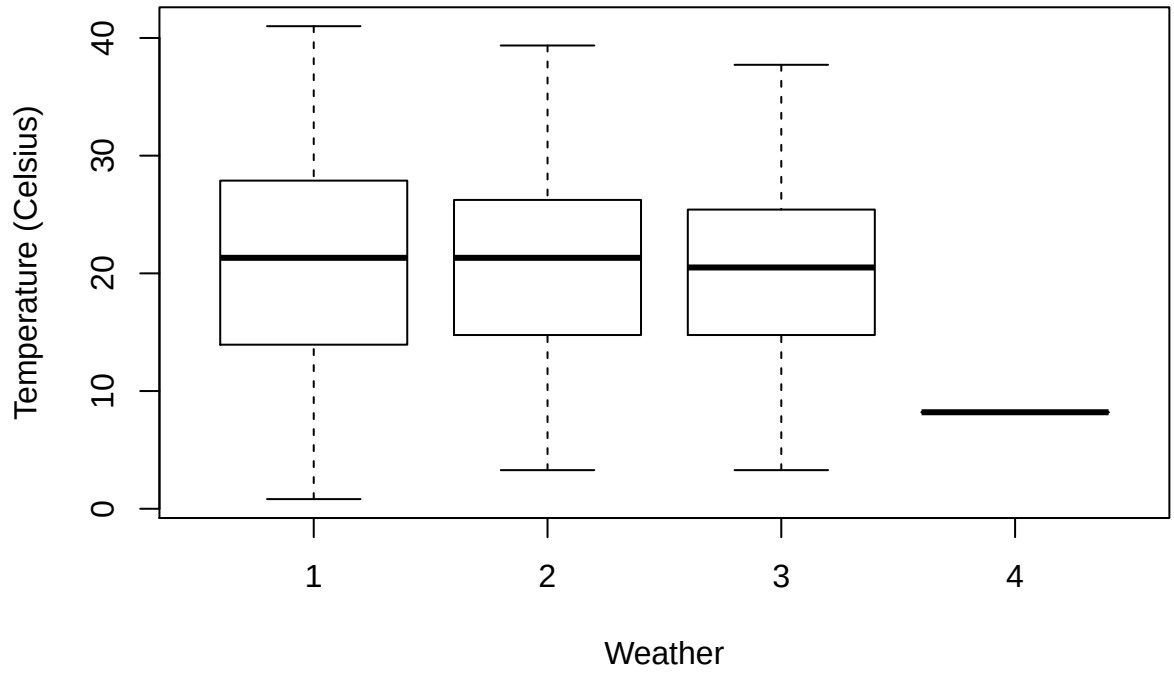
The relationship between time and humidity is also sinusoidal.

```
plot(df$hour, df$humidity, xlab= "Time", ylab="Humidity")
```



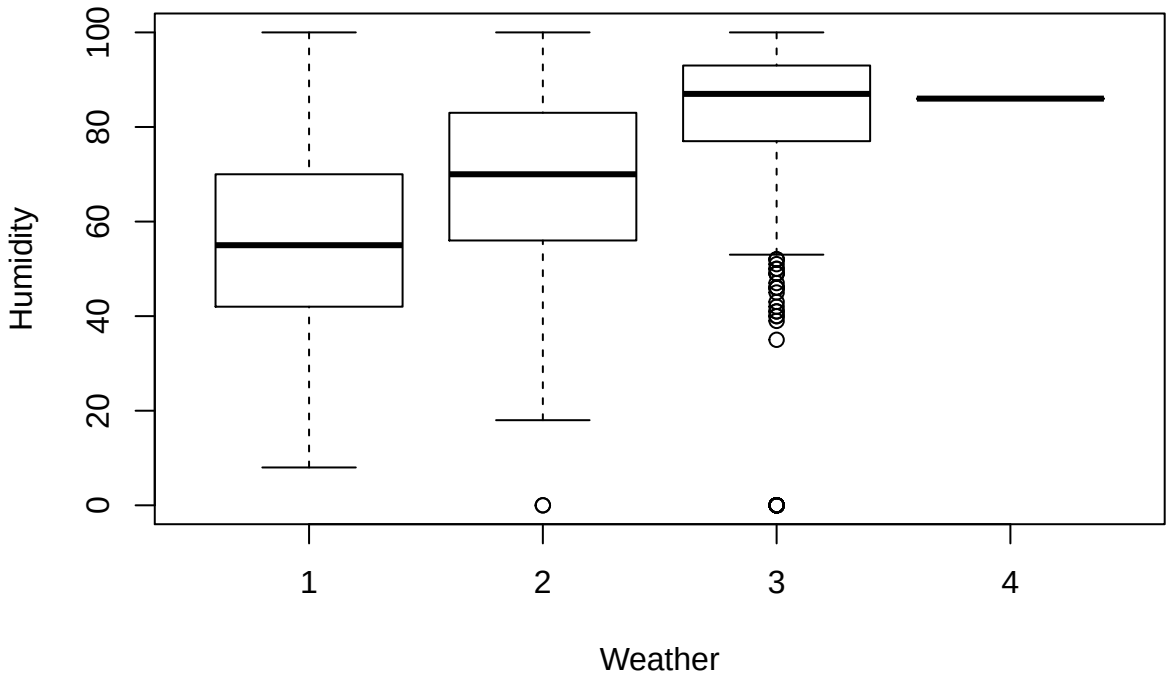
There is no indication of a relationship between weather and temp_c.

```
plot(df$weather, df$temp_c, xlab= "Weather", ylab="Temperature (Celsius)")
```



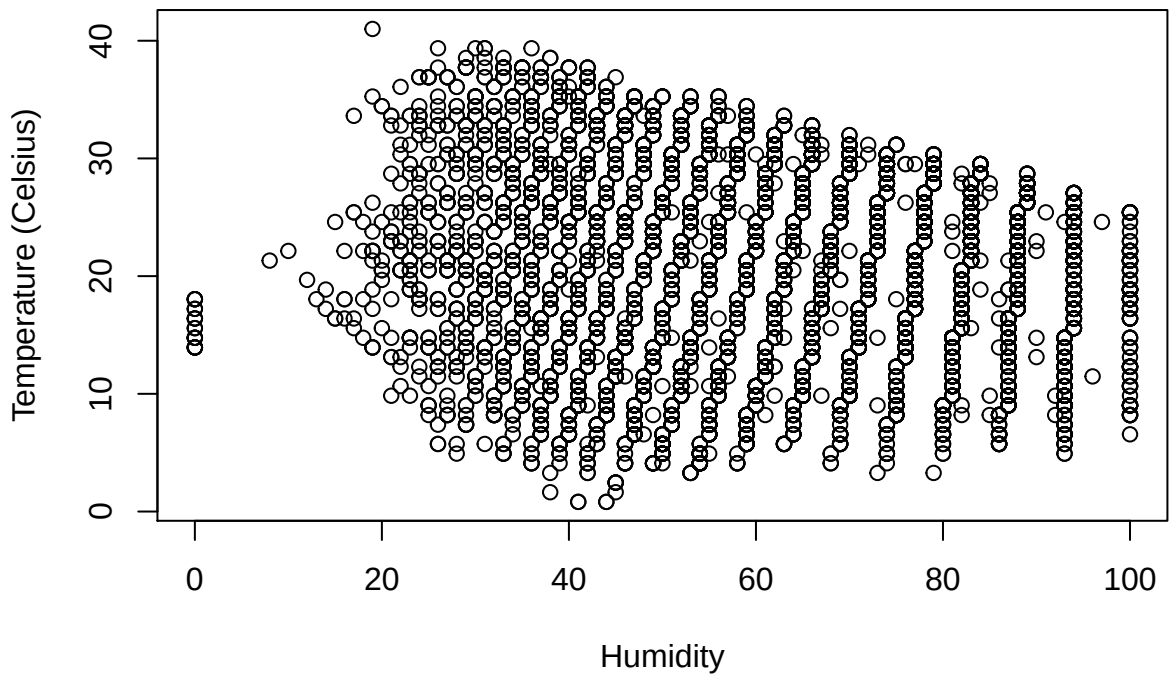
As the weather begins to change from 1-3 the median humidity also increases. There is only 1 observation for weather 4 so we cannot determine what the accurate relationship at this point.

```
plot(df$weather, df$humidity, xlab= "Weather", ylab="Humidity")
```



There is a relationship between humidity and temperature. From the graph we can see that as the humidity increases the temperature lowers.

```
plot(df$humidity, df$temp_c, xlab= "Humidity", ylab="Temperature (Celsius)")
```



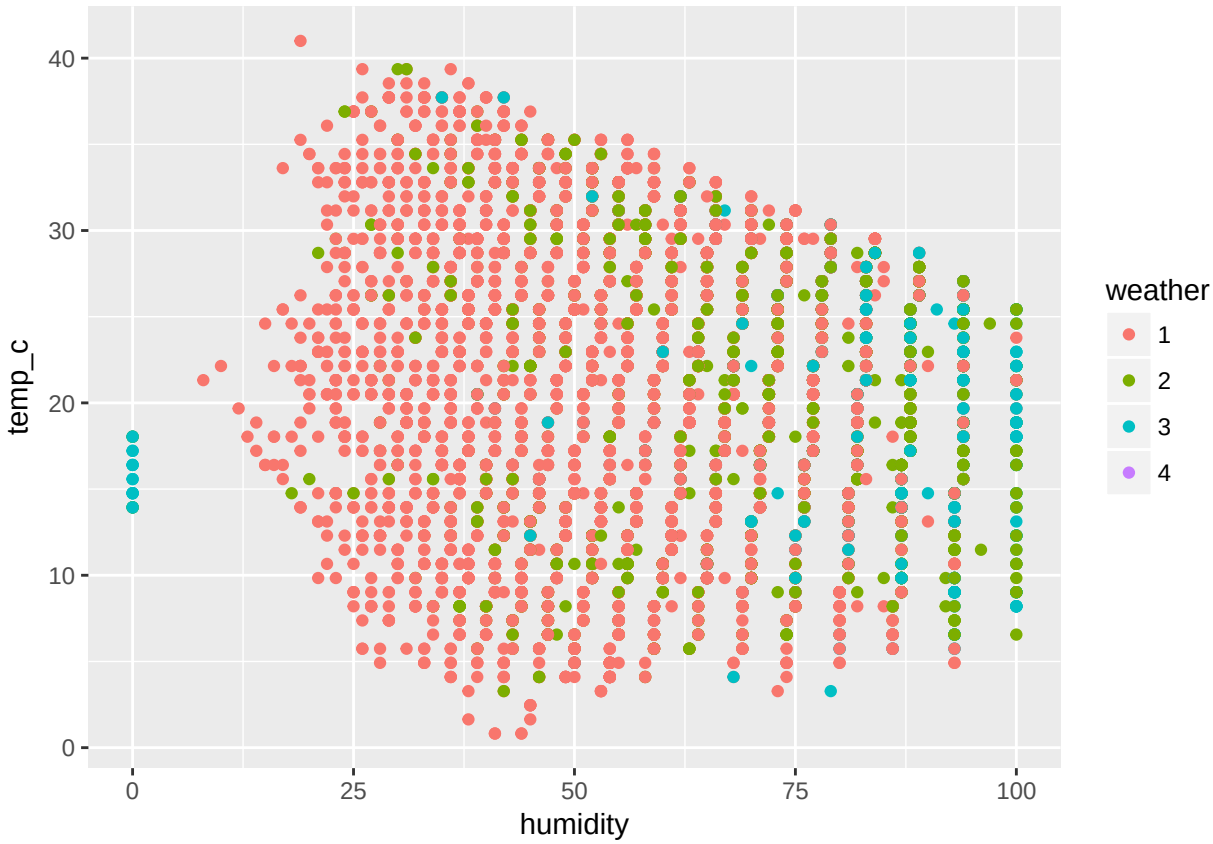
c

Numerical: Temperature, Humidity, Year
Categorical: Month, Day, Hour, Count, Weather, Vacation, Quarter, Workday

d

The following plot has temperature vs humidity and the observations are filled with the weather type. Analysis of the graph shows that there is diminishing variation and that the max temp also decreases as humidity increases. From this we can hypothesize the weather 1 corresponds to humidity between (7%, 62%) and temperature that can be as high as (34,40+) degrees Celsius. Weather 2 is dense near the (63%, 82%) humidity region and the temperature can be as reach temperatures around (28,33) degrees Celsius. Weather 3 is centered around the (83%, 100%) humidity region. Temperatures can be as high as (26,29) degrees Celsius. Weather 4 only has 1 observation and it corresponds 86% humidity and 8.2 degrees Celsius.

```
pl <- ggplot(data=df,aes(x = humidity,y=temp_c))
pl + geom_point(aes( color=weather))
```



e

The data contained na values which I discarded to do the analysis above. In addition there was one reading of -200 degrees Celsius which is most likely an error and should be discarded.

2. Analyze Widget usage rates and create predictive model

I will be performing linear regression to create my model. I decided to create a preliminary model which includes most the factors in the dataframe and their two factor-interactions. The Anova output suggest that the main effects and some of the two-factor interactions are significant.

```
model1 <- lm(count~ (hour + day + month + humidity +
                     weather + temp_c +workday + vacation+ quarter + year)^2, train)
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: count
##                    Df    Sum Sq  Mean Sq   F value    Pr(>F)
## hour               23 105865138  4602832 1987.0841 < 2.2e-16 ***
## day                 6    334183    55697   24.0450 < 2.2e-16 ***
## month              11  18071812  1642892  709.2513 < 2.2e-16 ***
## humidity            1   3962532  3962532 1710.6606 < 2.2e-16 ***
## weather             3   1625685   541895  233.9409 < 2.2e-16 ***
## temp_c              1   3423478  3423478 1477.9463 < 2.2e-16 ***
```

```
## workday             1     12121    12121    5.2327 0.0222040 *
## year                1  11457083 11457083 4946.1260 < 2.2e-16 ***
## hour:day          138  26105541   189171   81.6666 < 2.2e-16 ***
## hour:month        253   8640109    34151   14.7431 < 2.2e-16 ***
## hour:humidity      23   1808974    78651   33.9544 < 2.2e-16 ***
## hour:weather       46    901440    19597    8.4600 < 2.2e-16 ***
## hour:temp_c        23   1218742    52989   22.8757 < 2.2e-16 ***
## hour:workday       23   1263227    54923   23.7107 < 2.2e-16 ***
## hour:year          23   5918586   257330  111.0916 < 2.2e-16 ***
## day:month          66    529991     8030    3.4667 < 2.2e-16 ***
## day:humidity        6     19988     3331    1.4382 0.1957640
## day:weather        12     89218     7435    3.2097 0.0001313 ***
## day:temp_c          6     72149    12025    5.1912 2.458e-05 ***
## day:workday         2     46479    23240   10.0327 4.475e-05 ***
## day:year            6    268923    44821   19.3494 < 2.2e-16 ***
## month:humidity     11    876150    79650   34.3856 < 2.2e-16 ***
## month:weather      22    127843     5811    2.5087 0.0001182 ***
## month:temp_c       11    711530    64685   27.9249 < 2.2e-16 ***
## month:workday       5     94849    18970    8.1895 1.027e-07 ***
## month:year          9    346745    38527   16.6326 < 2.2e-16 ***
## humidity:weather    2    140135    70068   30.2488 8.615e-14 ***
## humidity:temp_c     1     19724    19724    8.5152 0.0035363 **
## humidity:workday    1      1301     1301    0.5619 0.4535430
## humidity:year       1     66327    66327   28.6342 9.099e-08 ***
## weather:temp_c      2     32172    16086    6.9445 0.0009724 ***
## weather:workday     2     47986    23993   10.3579 3.237e-05 ***
## weather:year        2    126709    63355   27.3508 1.516e-12 ***
## temp_c:workday      1     13551    13551    5.8502 0.0156073 *
## temp_c:year         1     68618    68618   29.6231 5.476e-08 ***
## workday:year        1      5958     5958    2.5719 0.1088321
## Residuals        5467  12663623     2316
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I've reduced the model to the most significant effects and determine the validity of this model.

```
library(alr3)
model2 <-lm(count ~ year*(temp_c+hour+day+month) +
              hour*(humidity+weather+temp_c+workday+day+month) +
              month*(humidity + temp_c) + humidity*weather+day*month, train)
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: count
##                    Df   Sum Sq  Mean Sq    F value    Pr(>F)
## year                1 14510584 14510584  5976.4992 < 2.2e-16 ***
## temp_c              1 32727242 32727242 13479.4253 < 2.2e-16 ***
## hour               23 89806701  3904639  1608.2104 < 2.2e-16 ***
## day                 6   471006    78501    32.3324 < 2.2e-16 ***
## month              11  3374348   306759   126.3453 < 2.2e-16 ***
## humidity            1  2101195  2101195   865.4229 < 2.2e-16 ***
## weather             3  1724892   574964   236.8114 < 2.2e-16 ***
## workday             1    36063    36063    14.8535 0.0001175 ***
## year:temp_c         1   638330   638330   262.9099 < 2.2e-16 ***
```

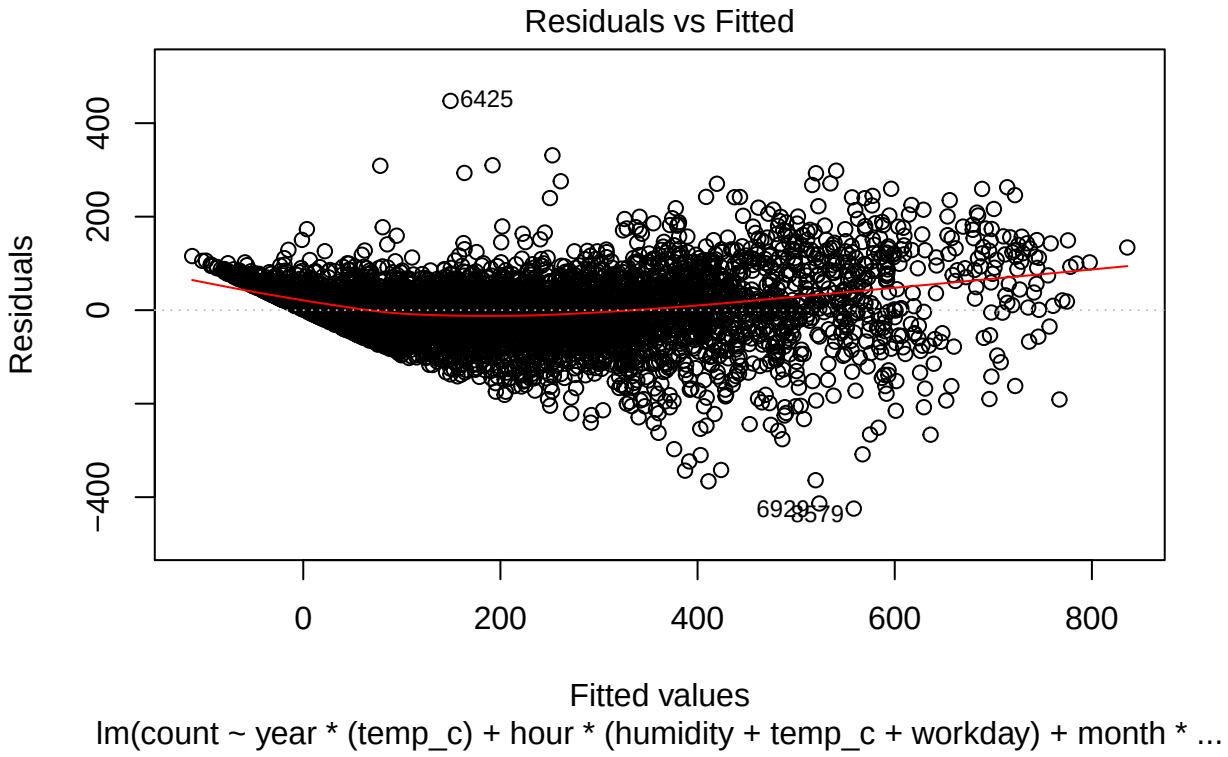
```
## year:hour           23  7466050   324611   133.6980 < 2.2e-16 ***
## year:day             6   382449    63742    26.2533 < 2.2e-16 ***
## year:month           9   737426    81936    33.7472 < 2.2e-16 ***
## hour:humidity       23   247590    10765     4.4337 8.600e-12 ***
## hour:weather        46  1969097    42806    17.6308 < 2.2e-16 ***
## temp_c:hour         23  5987755   260337   107.2255 < 2.2e-16 ***
## hour:workday        23 25024012  1088001   448.1166 < 2.2e-16 ***
## hour:day           138  1601978    11609     4.7812 < 2.2e-16 ***
## hour:month         253  2807133    11095     4.5699 < 2.2e-16 ***
## month:humidity      11   813424    73948    30.4569 < 2.2e-16 ***
## temp_c:month        11   523307    47573    19.5941 < 2.2e-16 ***
## humidity:weather     2   127992    63996    26.3581 4.046e-12 ***
## day:month           66   471110     7138     2.9400 3.317e-14 ***
## Residuals         5531 13428939     2428
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

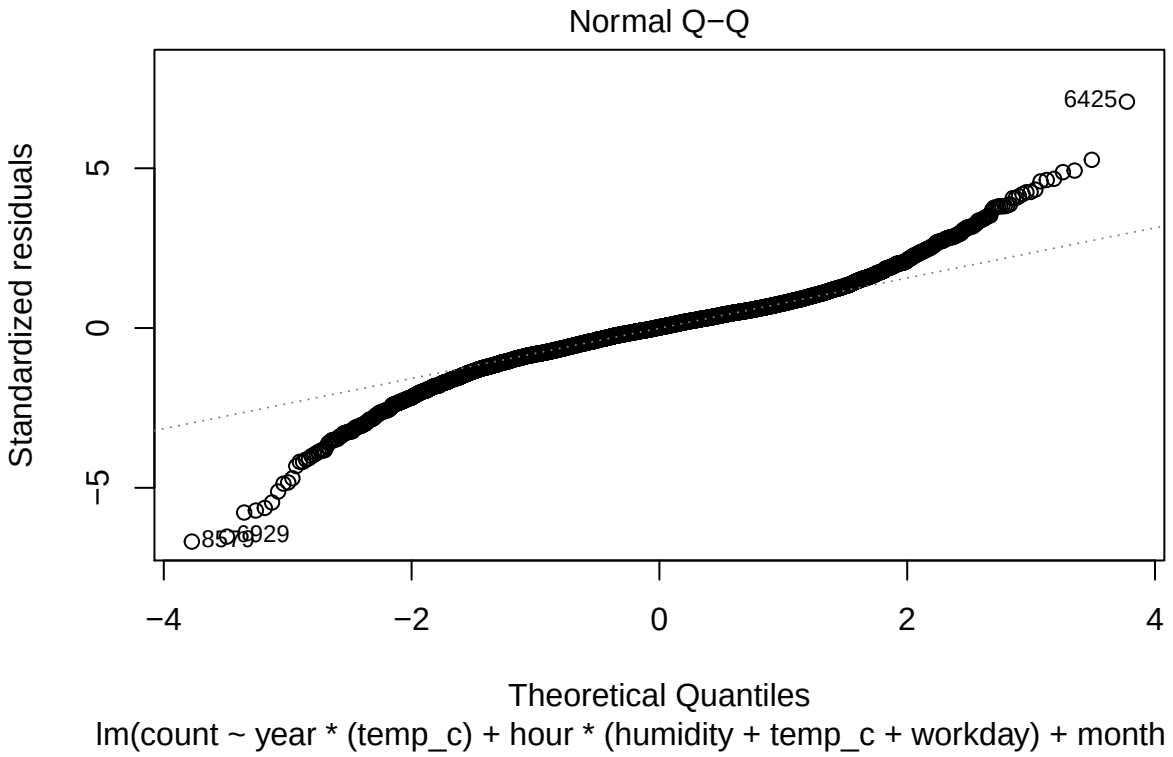
Since I am also modeling interaction terms any interactions and their main effects are going to be highly correlated. The variance inflation factor will not work due to aliasing in the terms so I will reduce the model by removing interactions between measures of time. Additionally the output from summary of the model shows a few NA's. I read up on this and found it may be due to something called "Dummy Variable Trap" and the main suggestion was to drop one of the categorical variables.

```
model2 <-lm(count ~ year*(temp_c) +
              hour*(humidity+temp_c+workday) +
               month*(humidity + temp_c) + day, train)
```

The residual plot has a trend line and signs of heteroskedasticity.

```
plot(model2, 1:2)
```





I use the powerTransform function to transform the response variable. The function suggests 0.273, so I used $\frac{1}{4}$ to for my transform.

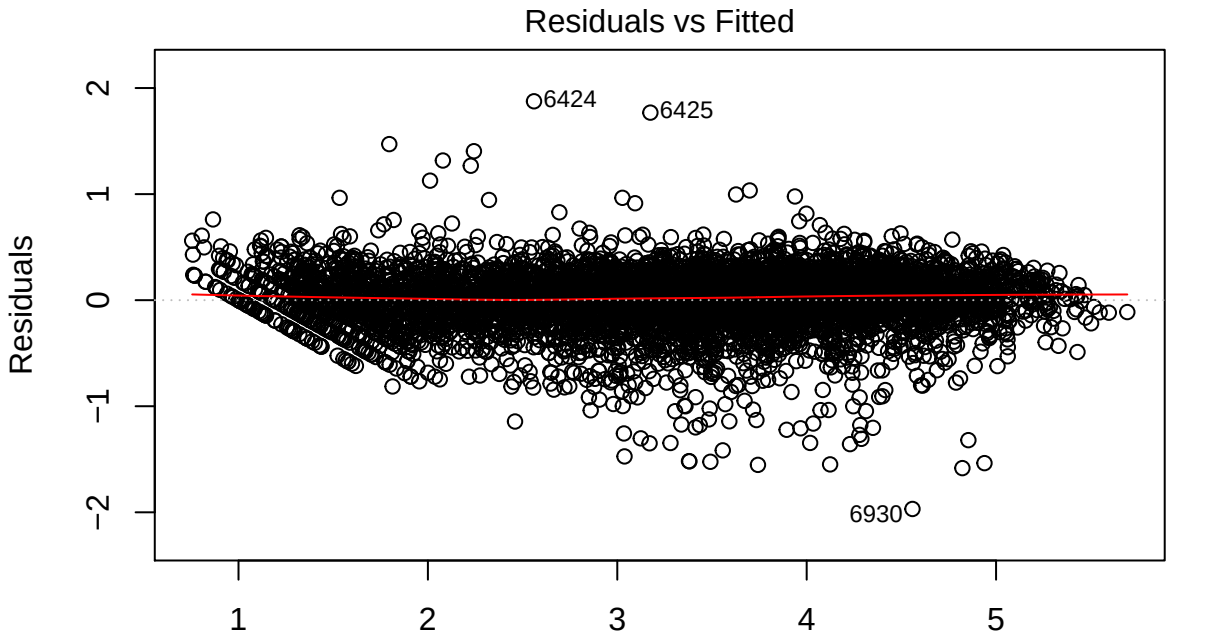
```
powerTransform(model2)
```

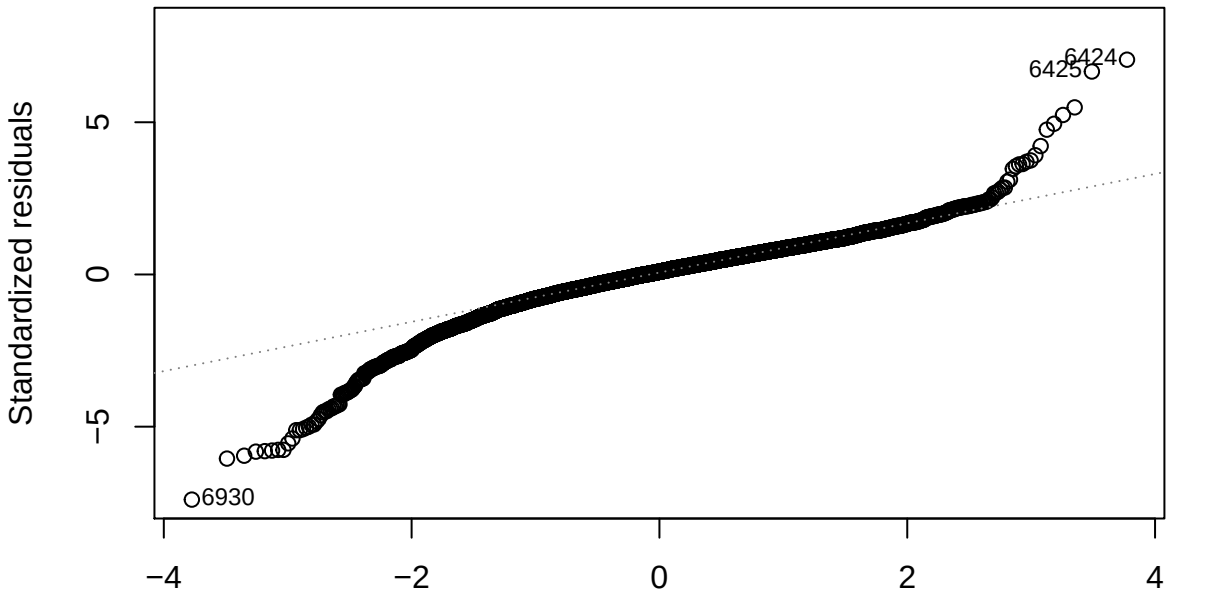
```
## Estimated transformation parameters
##        Y1
## 0.2264431
```

```
model3 <-lm(count^(.25) ~ year*(temp_c) +
              hour*(humidity+temp_c+workday) +
               month*(humidity + temp_c)  + day , train)
```

The trend line in the residual plot is no longer there and the variance is approximately constant. The marginal model plots are approximately fitted as well.

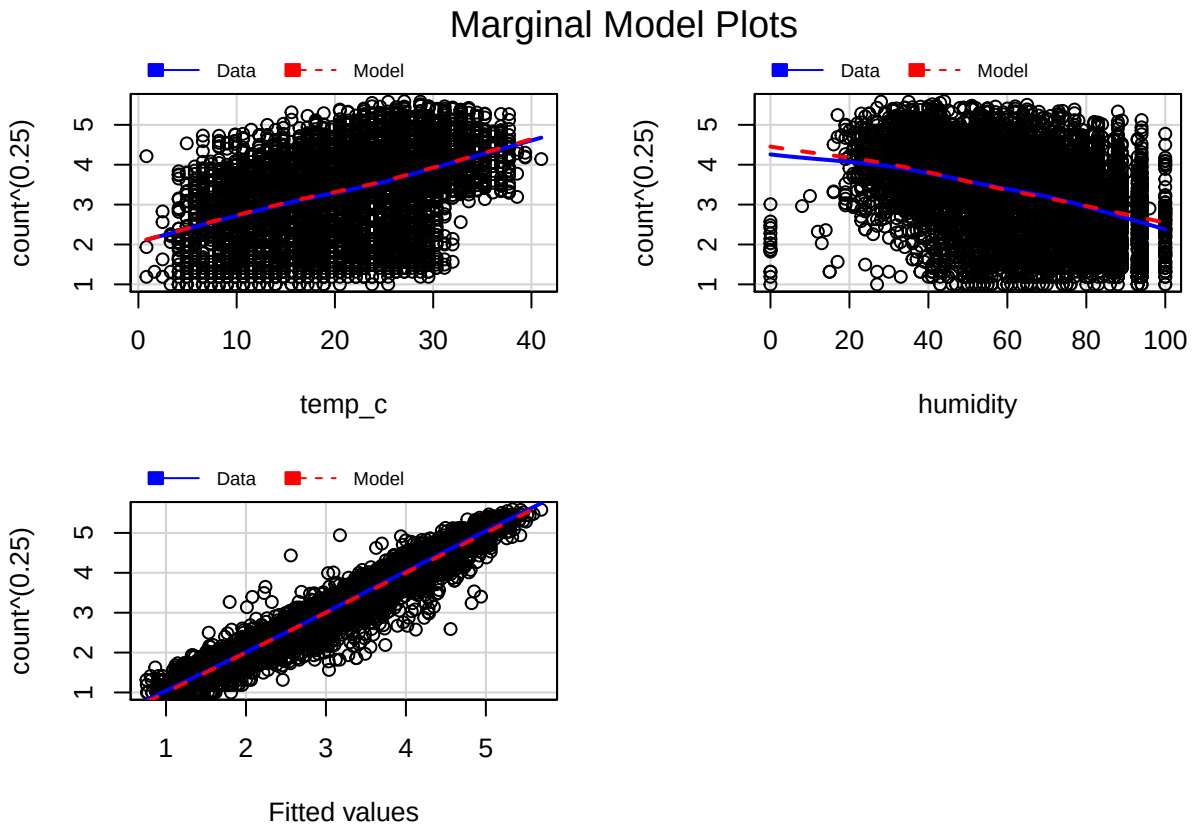
```
plot(model3,1:2)
```






```
mmps(model3)
```

```
## Warning in mmps(model3): Interactions and/or factors skipped
```



The final mode consists of the main terms and some two factor interactions. I transformed the response variable according to the powerTransform to the power of $\frac{1}{4}$. According to the anova output; time, temperature, hour:workday, year, month are all highly significant and have high sum of squares.

Prediction

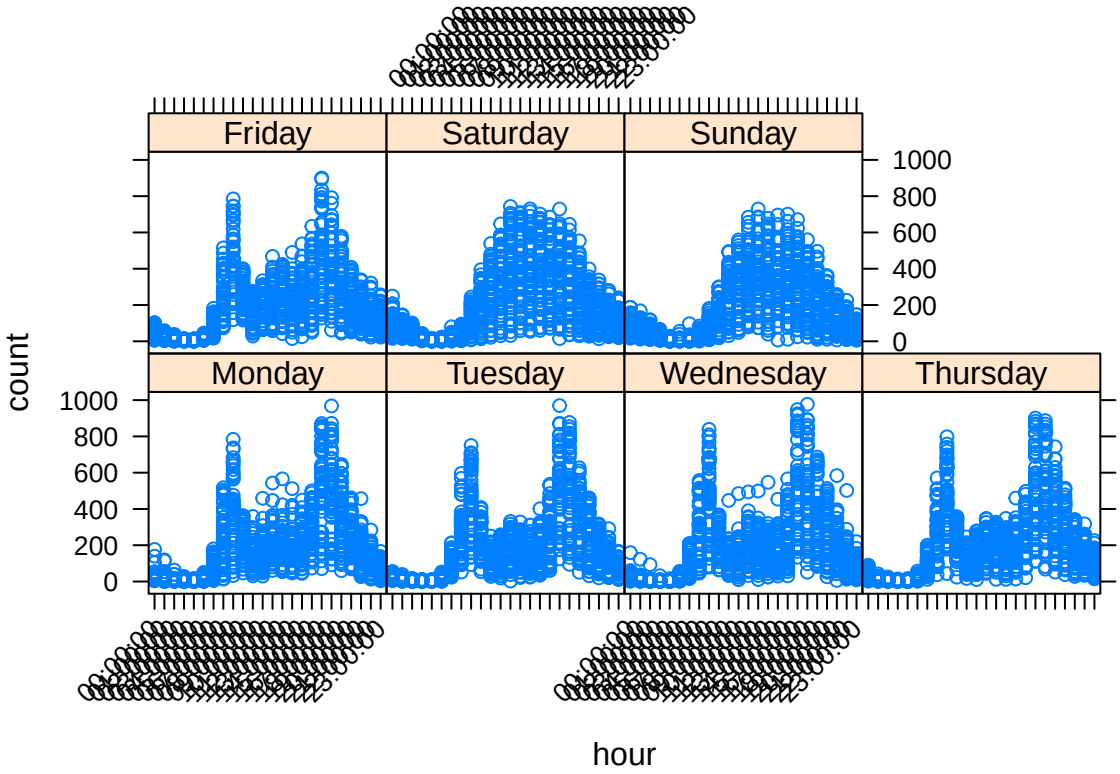
I predict the rest of the data and calculate the root mean square error:

```
## [1] 0.004491936
```

a

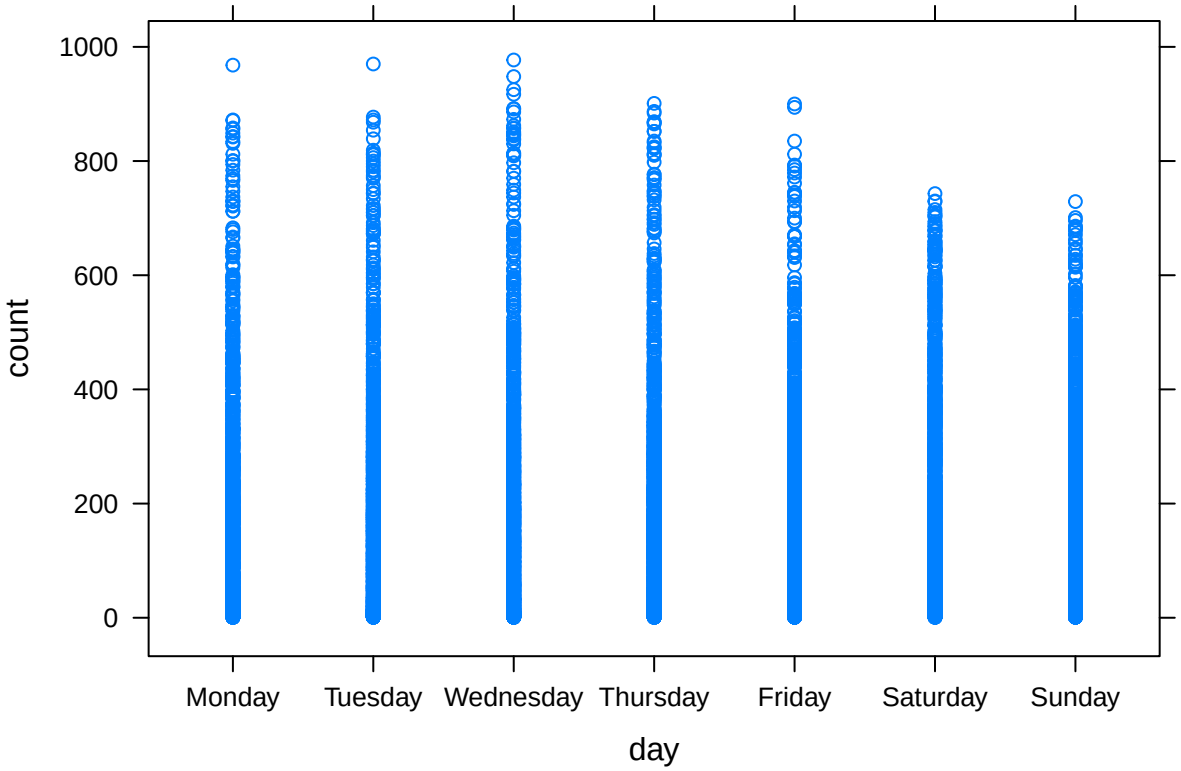
The following plot shows the widget usage per hour given a day. From the graph we can see that on weekdays the peak times are 7:00:00-8:00:00 and 17:00:00-18:00:00. On the weekends the peak times are 12:00:00-17:00:00.

```
library(lattice)
library(car)
xyplot(count~hour|day,data = df,scales=list(x=list(rot=45)))
```



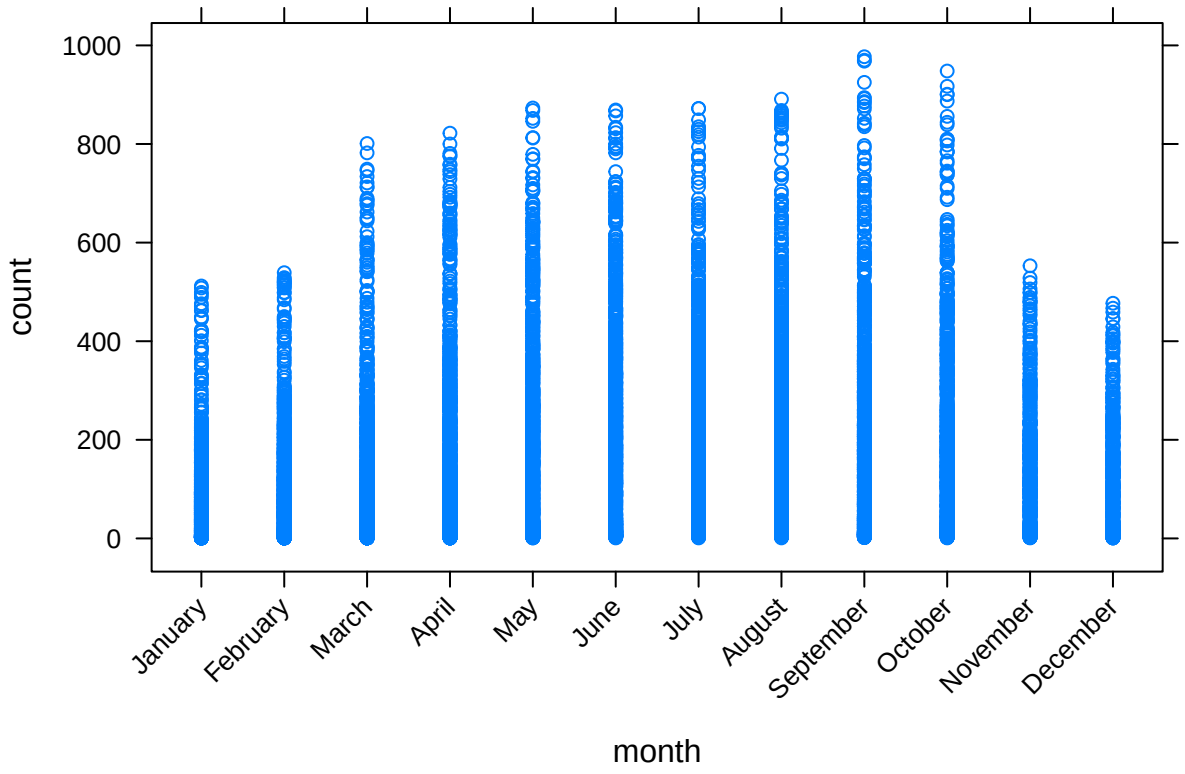
The plot shows that Monday, Tuesday, and Wednesday are the most popular days of the week.

```
xyplot(count~day,data=df)
```



The plot shows that September and October are the most popular months of the year.

```
xyplot(count~month,data=df, scales=list(x=list(rot=45)))
```

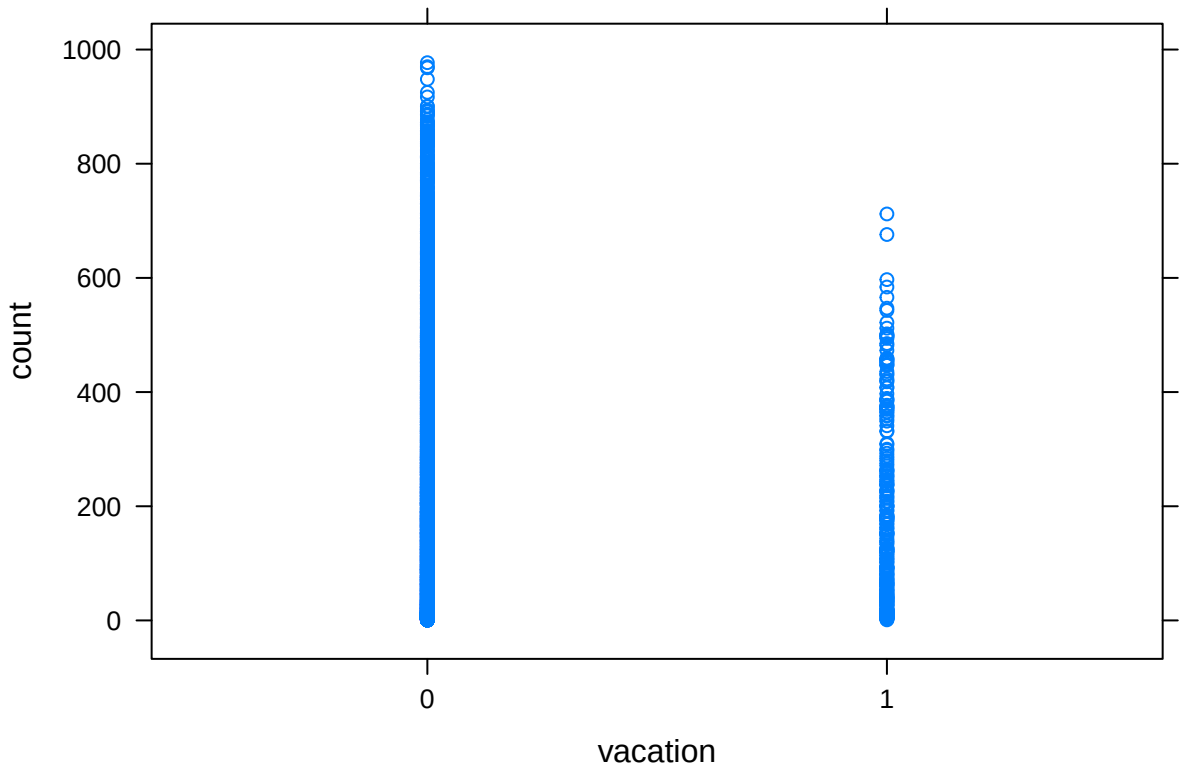


b

According to the model which suggested that the interaction between time and workday was significant there is less usage on the weekends and this is also evident in the graph of usage as a function of days.

The following graph shows that usage is lower on holidays.

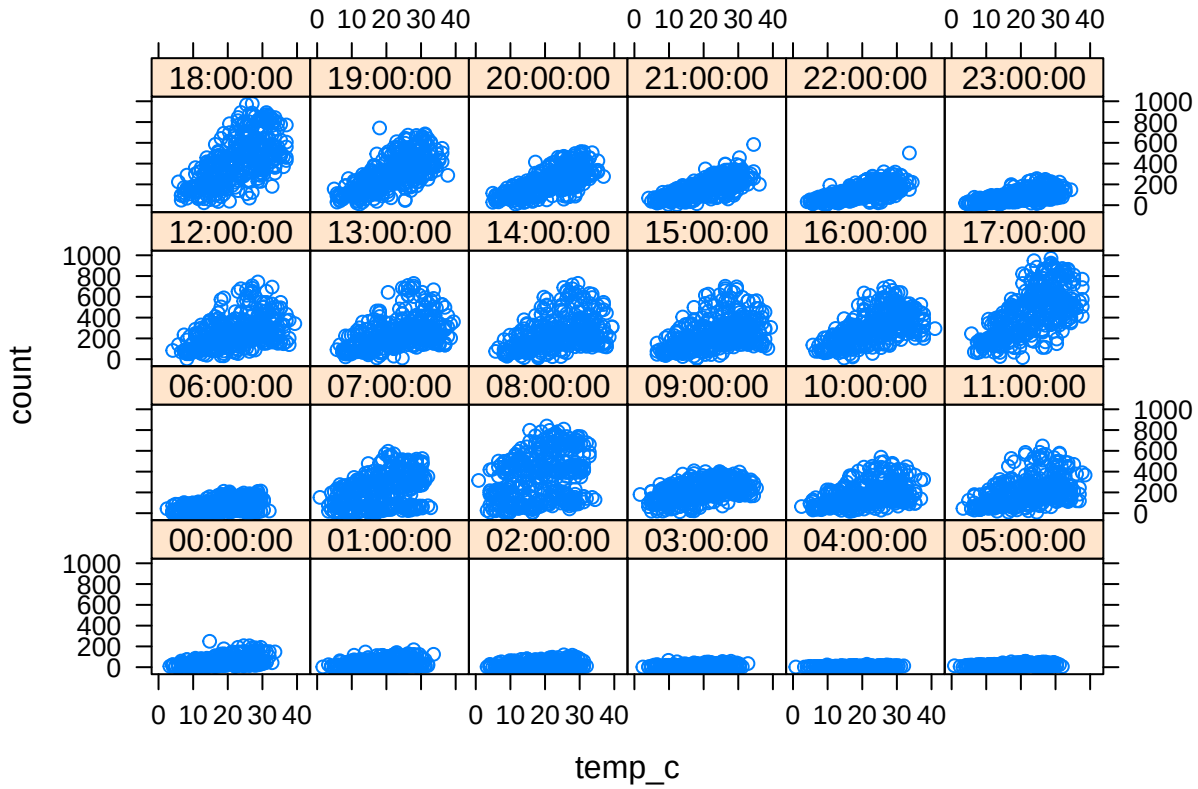
```
xyplot(count~vacation,data=df)
```



c

According to the model temperature also has an effect on usage. The model suggest that the interaction between temperature and months and days is significant. The following graph shows that widget usage increases as temperature increases on most days.

```
xyplot(count~ temp_c| hour,data=df)
```



The same is true when we plot for months usage increases as temperature increases.

```
xyplot(count~ temp_c| month,data=df)
```