EDA

Jiashen Wang

2025-02-03

```
##
## Attaching package: 'lubridate'

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

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

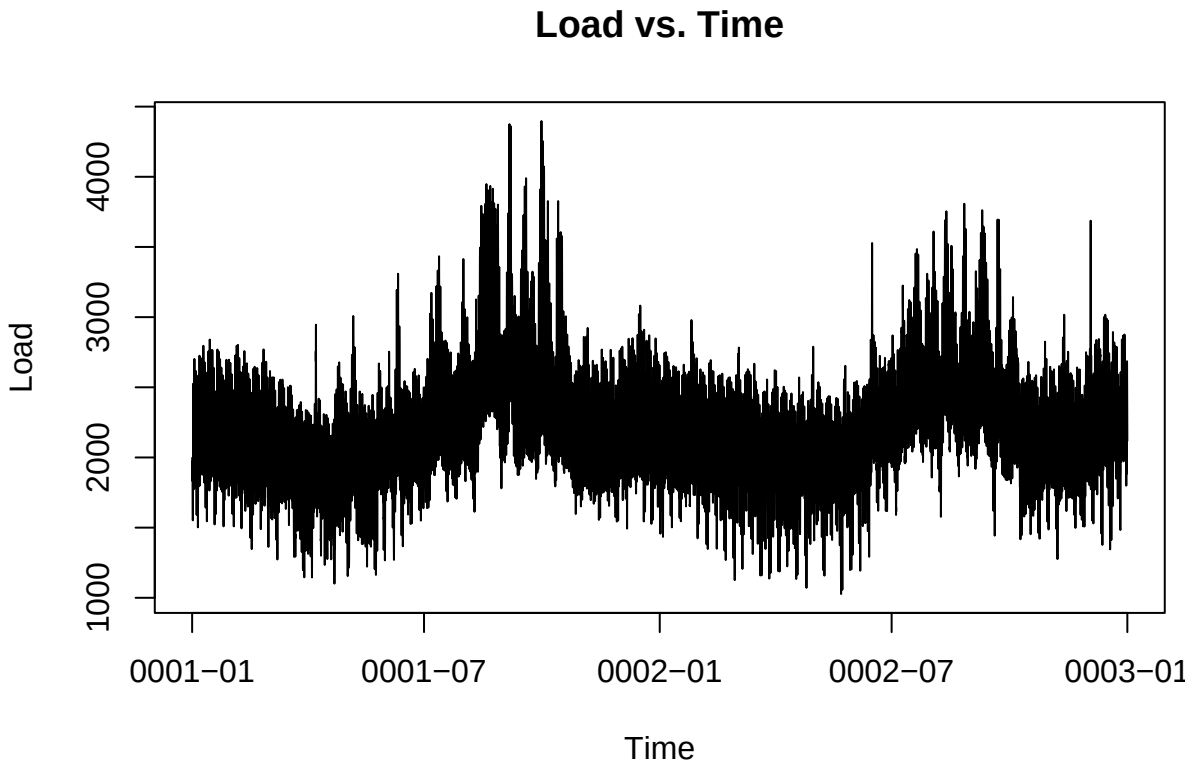
Load: Electricity load at the node of interest
Site-X-Temp: Temperature at a random location within the node of interest
Site-X-GHI: GHI at a random location within the node of interest. GHI is the total solar radiation incident on a horizontal surface

```
# create a column of datetime
train$datetime <- make_datetime(train$Year, train$Month, train$Day, train$Hour)
# convert load column to dbl
train$Load = as.numeric(gsub(",", "", train$Load))

head(train)
```

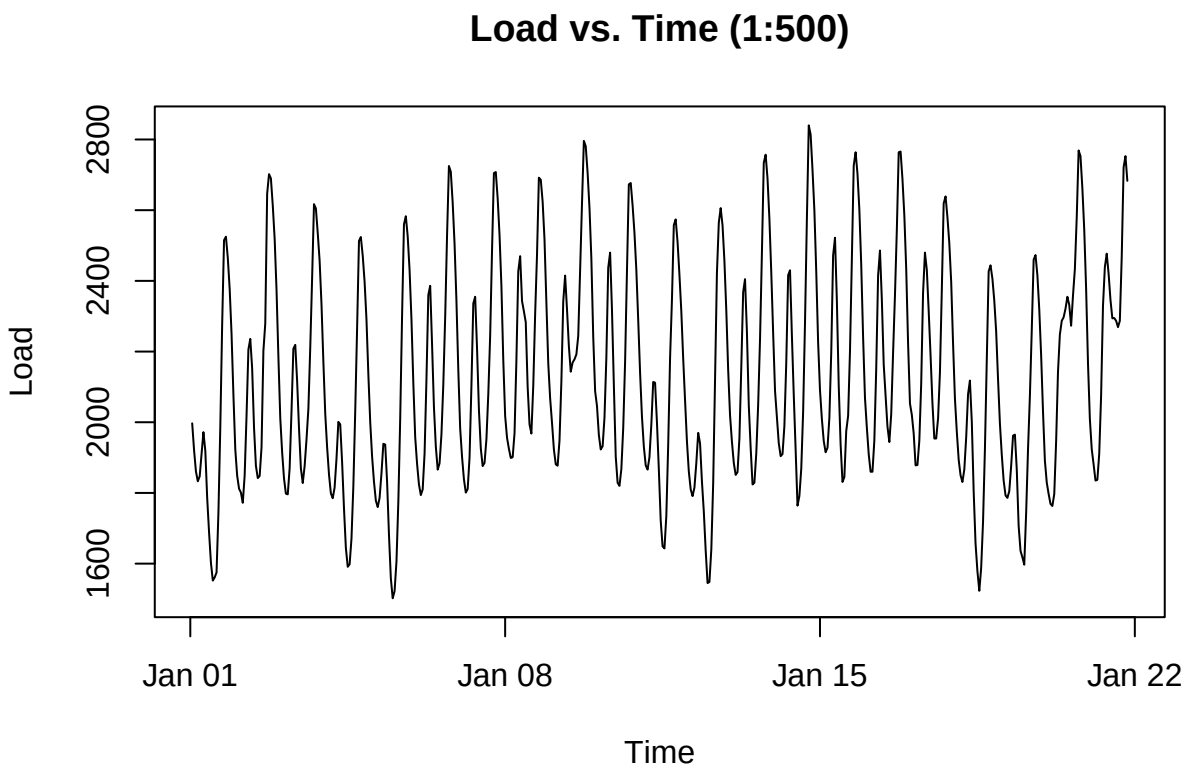
```
##   Year Month Day Hour Load Site.1.Temp Site.2.Temp Site.3.Temp Site.4.Temp
## 1    1     1   1    1 1997         8.0         8.2         5.3         9.4
## 2    1     1   1    2 1921         8.3         8.6         5.2         8.6
## 3    1     1   1    3 1861         8.1         8.8         5.1         8.7
## 4    1     1   1    4 1833         7.6         8.1         4.3         8.5
## 5    1     1   1    5 1847         7.3         7.5         4.0         8.6
## 6    1     1   1    6 1910         6.6         7.3         4.0         7.8
##   Site.5.Temp Site.1.GHI Site.2.GHI Site.3.GHI Site.4.GHI Site.5.GHI  X X.1 X.2
## 1         8.1          0          0          0          0          0 NA  NA  NA
## 2         7.1          0          0          0          0          0 NA  NA  NA
## 3         6.2          0          0          0          0          0 NA  NA  NA
## 4         6.0          0          0          0          0          0 NA  NA  NA
## 5         6.9          0          0          0          0          0 NA  NA  NA
## 6         7.3          0          0          0          0          0 NA  NA  NA
```

```
##   X.3 X.4 X.5 X.6 X.7            datetime
## 1  NA  NA  NA  NA  NA 0001-01-01 01:00:00
## 2  NA  NA  NA  NA  NA 0001-01-01 02:00:00
## 3  NA  NA  NA  NA  NA 0001-01-01 03:00:00
## 4  NA  NA  NA  NA  NA 0001-01-01 04:00:00
## 5  NA  NA  NA  NA  NA 0001-01-01 05:00:00
## 6  NA  NA  NA  NA  NA 0001-01-01 06:00:00
```

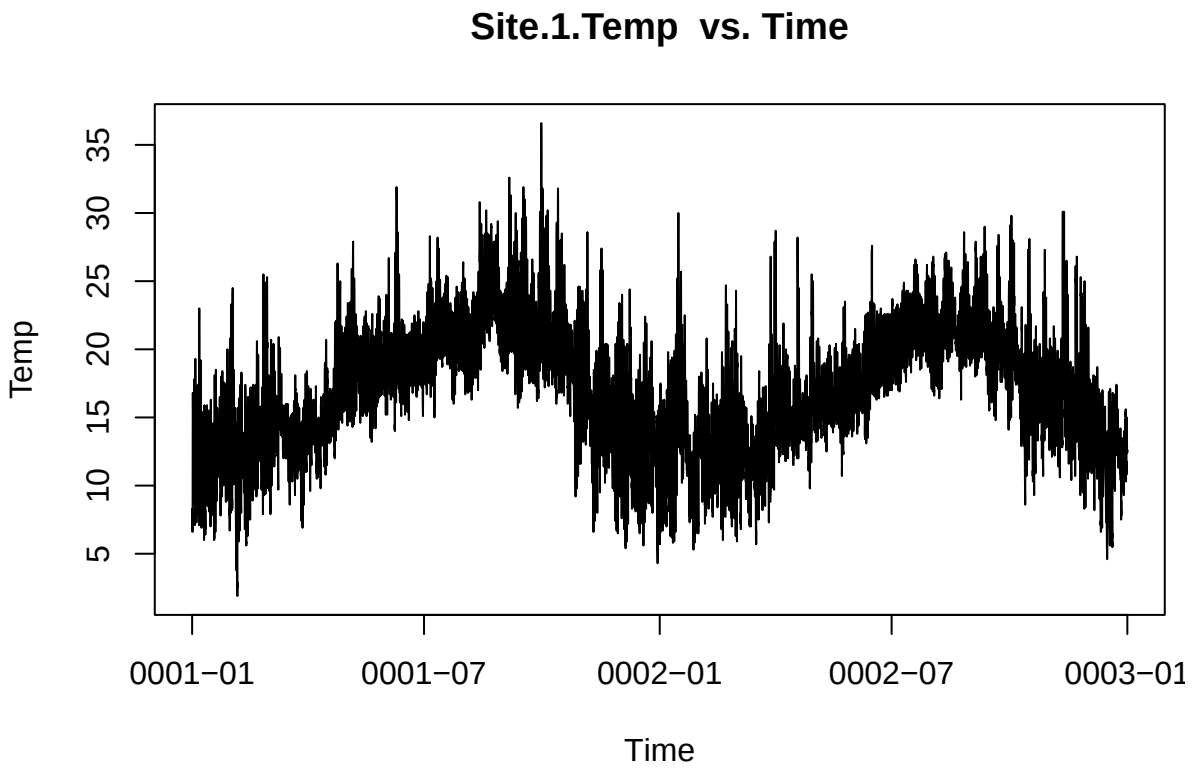
```r
# Load vs. Time
plot(train$datetime, train$Load, type = "l", main = "Load vs. Time",
     ylab = "Load", xlab = "Time")
```

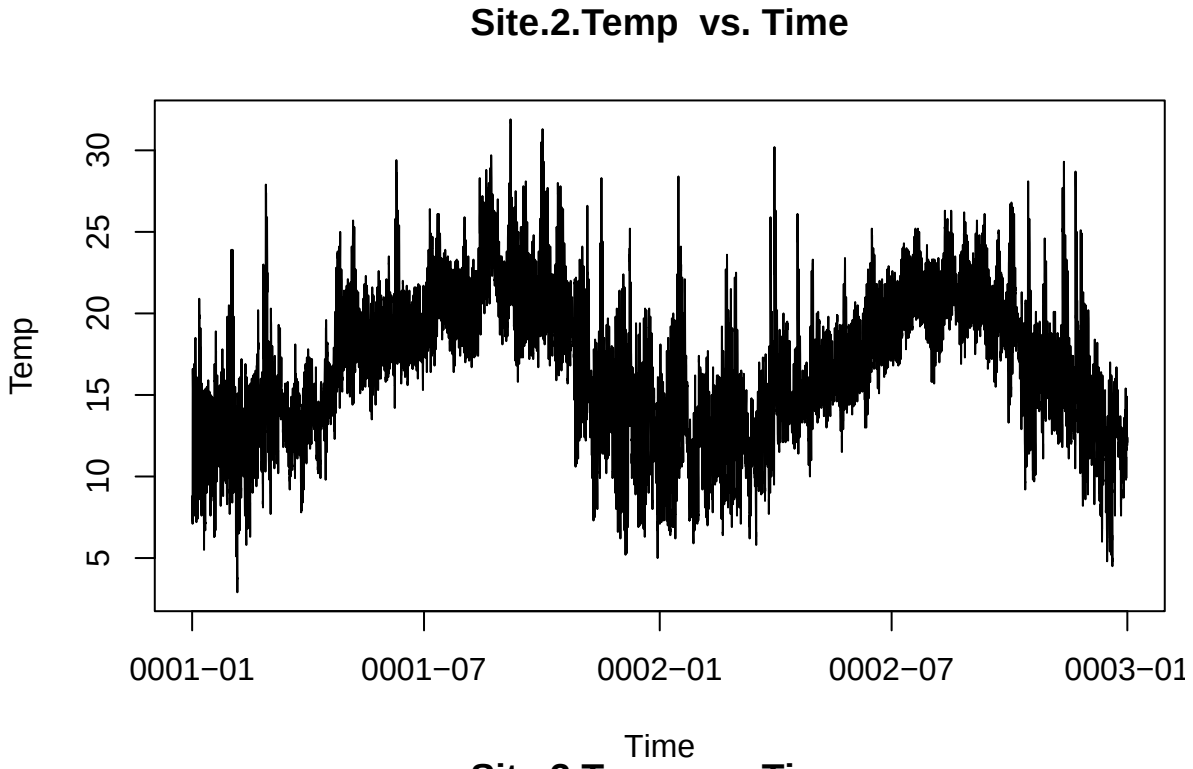


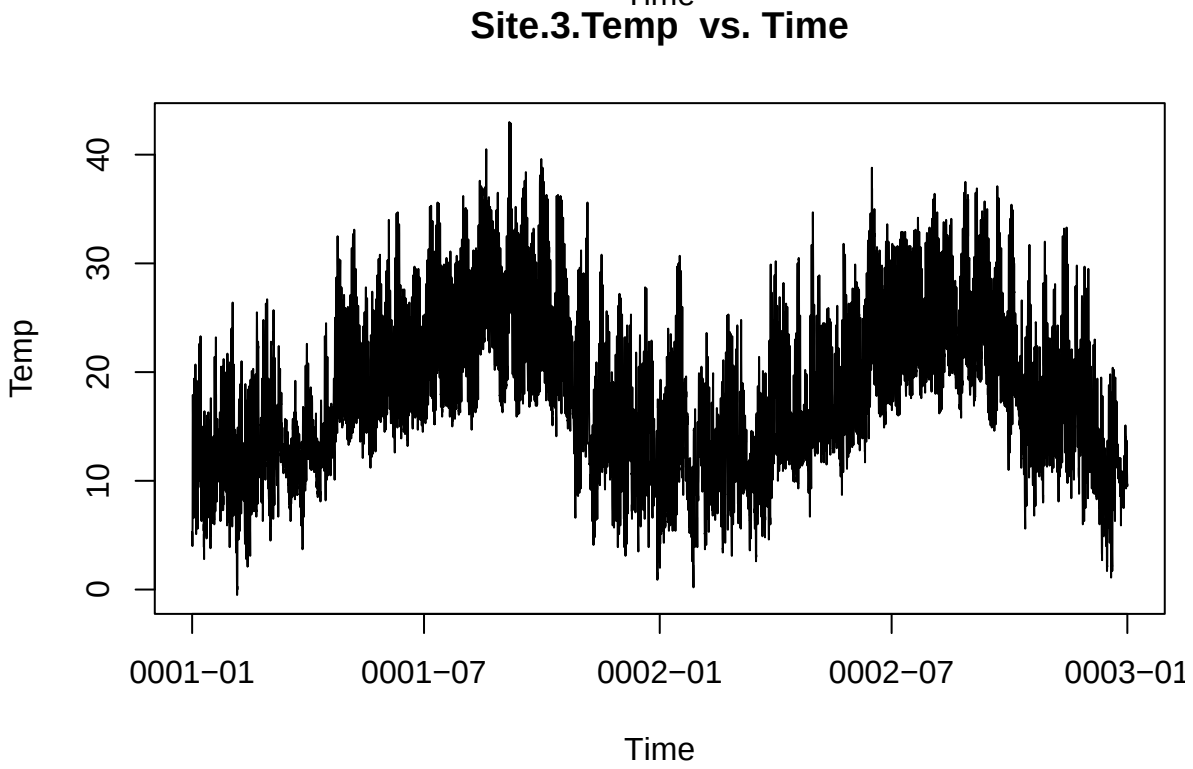

```r
plot(train$datetime[1:500], train$Load[1:500], type = "l",
     main = "Load vs. Time (1:500)", ylab = "Load", xlab = "Time")
```

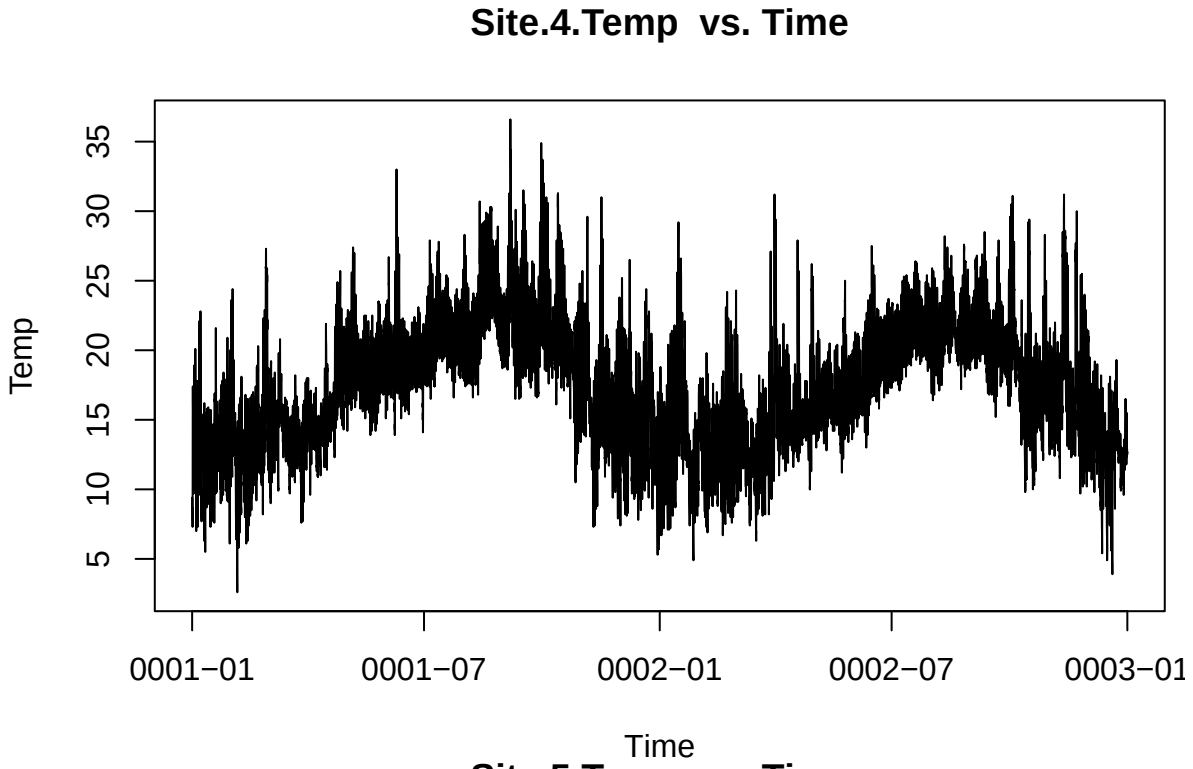

Load vs. Time (1:500)
Load
1600
2000
2400
2800
Jan 01
Jan 08
Jan 15
Jan 22
Time

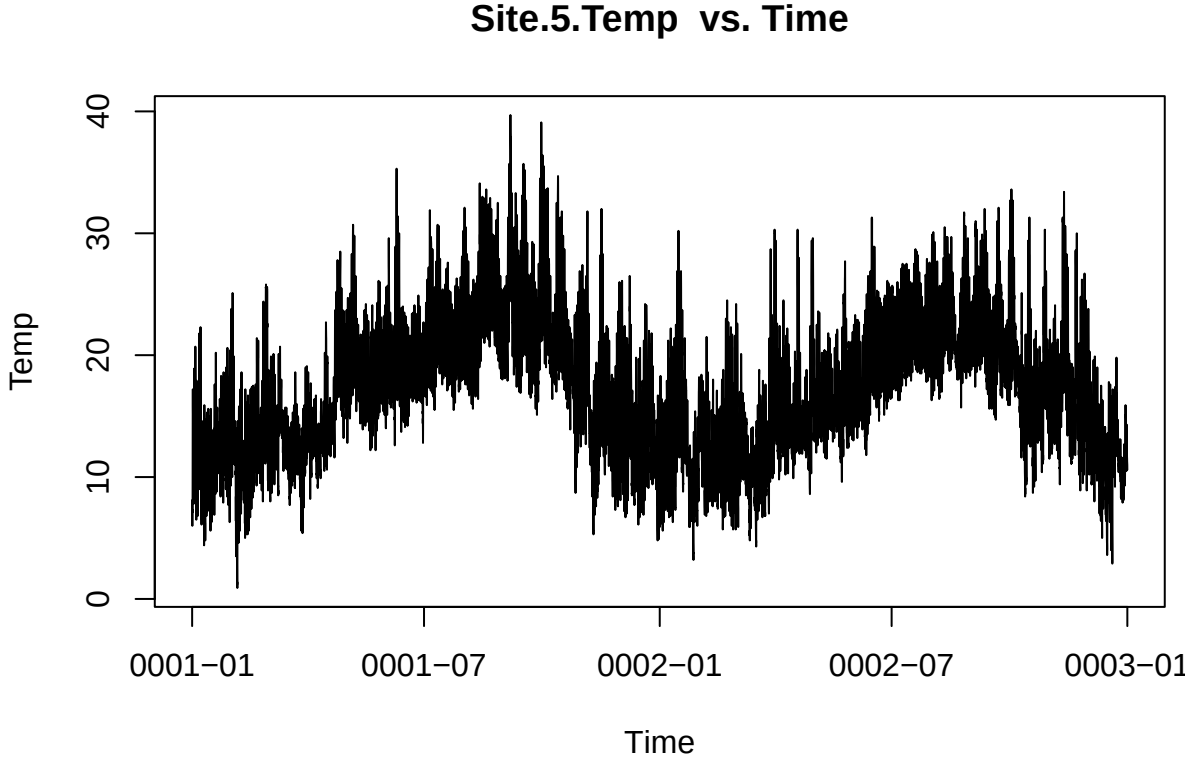
```r
# 5 Temp columns vs. Time
temp_colnames = c("Site.1.Temp", "Site.2.Temp", "Site.3.Temp",
                  "Site.4.Temp", "Site.5.Temp")
for (colname in temp_colnames) {
  main_suffix = " vs. Time"
  plot(train$datetime,train[,colname], main = paste(colname, main_suffix),
       type = "l", ylab = "Temp", xlab = "Time")
}
```



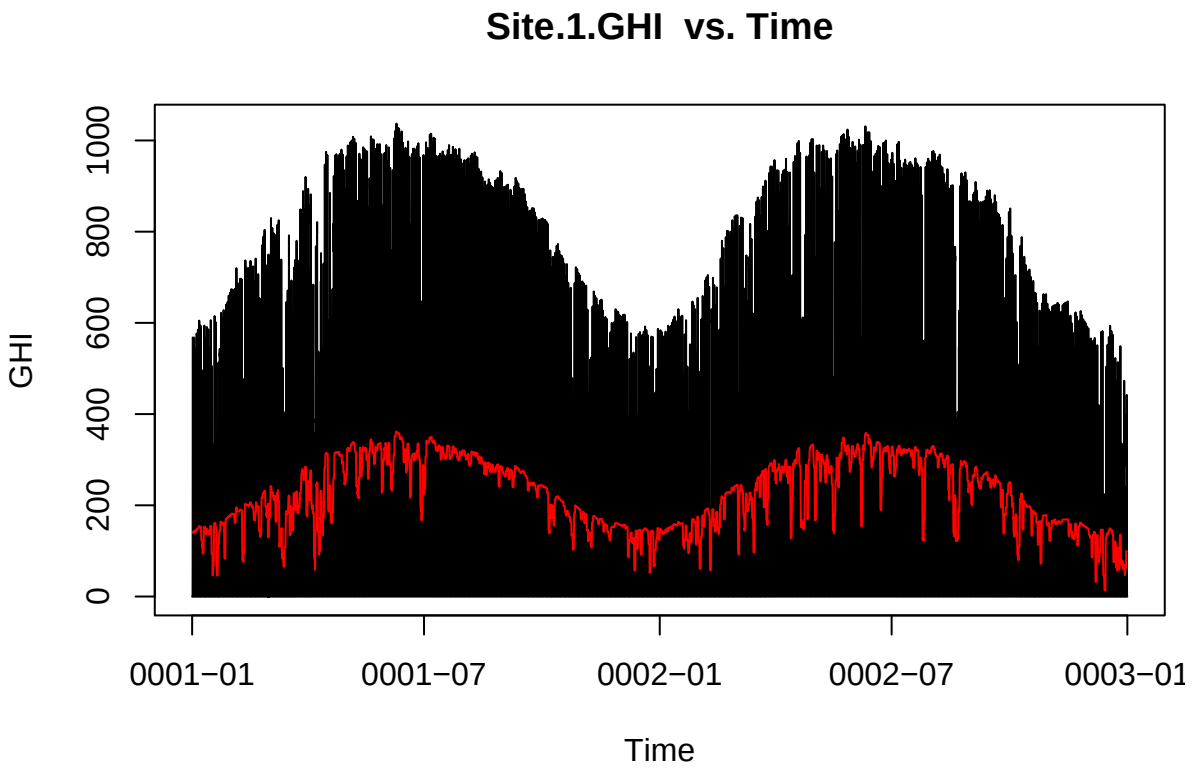

Site.2.Temp vs. Time
Temp
5
10
15
20
25
30
0001−01
0001−07
0002−01
0002−07
0003−01
Time

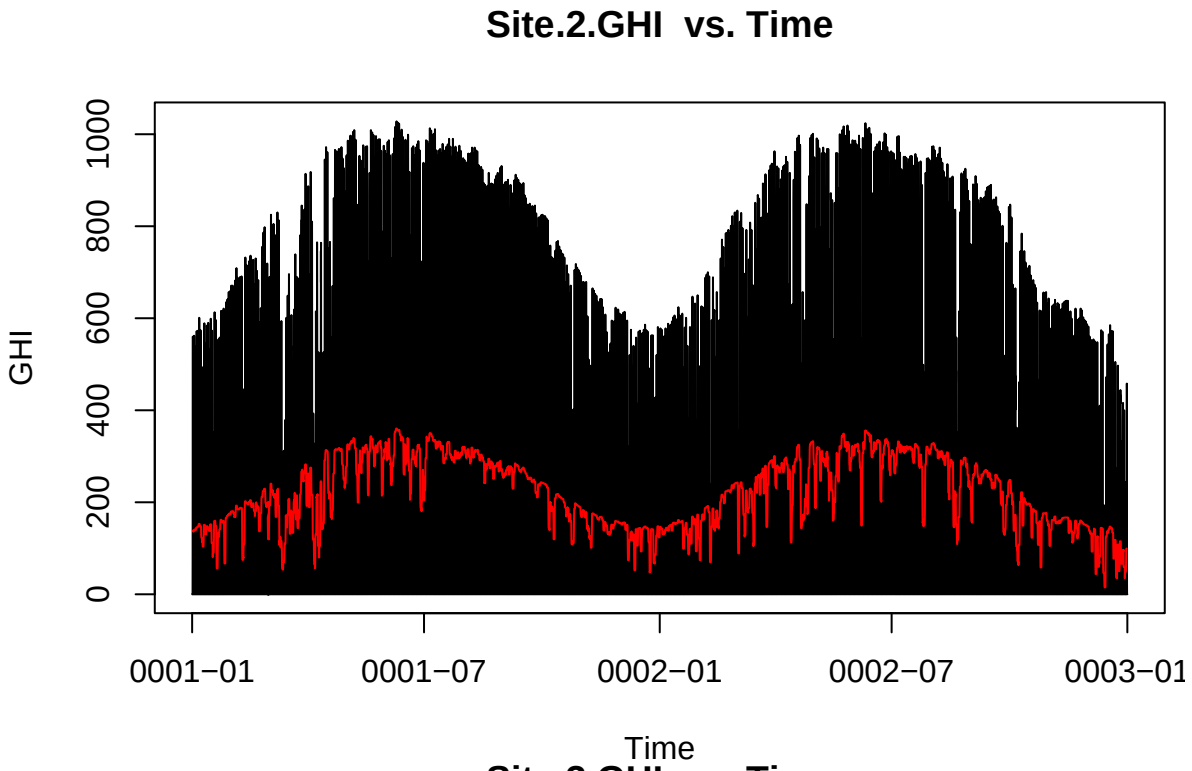

Site.3.Temp vs. Time
Temp
0
10
20
30
40
0001−01
0001−07
0002−01
0002−07
0003−01
Time

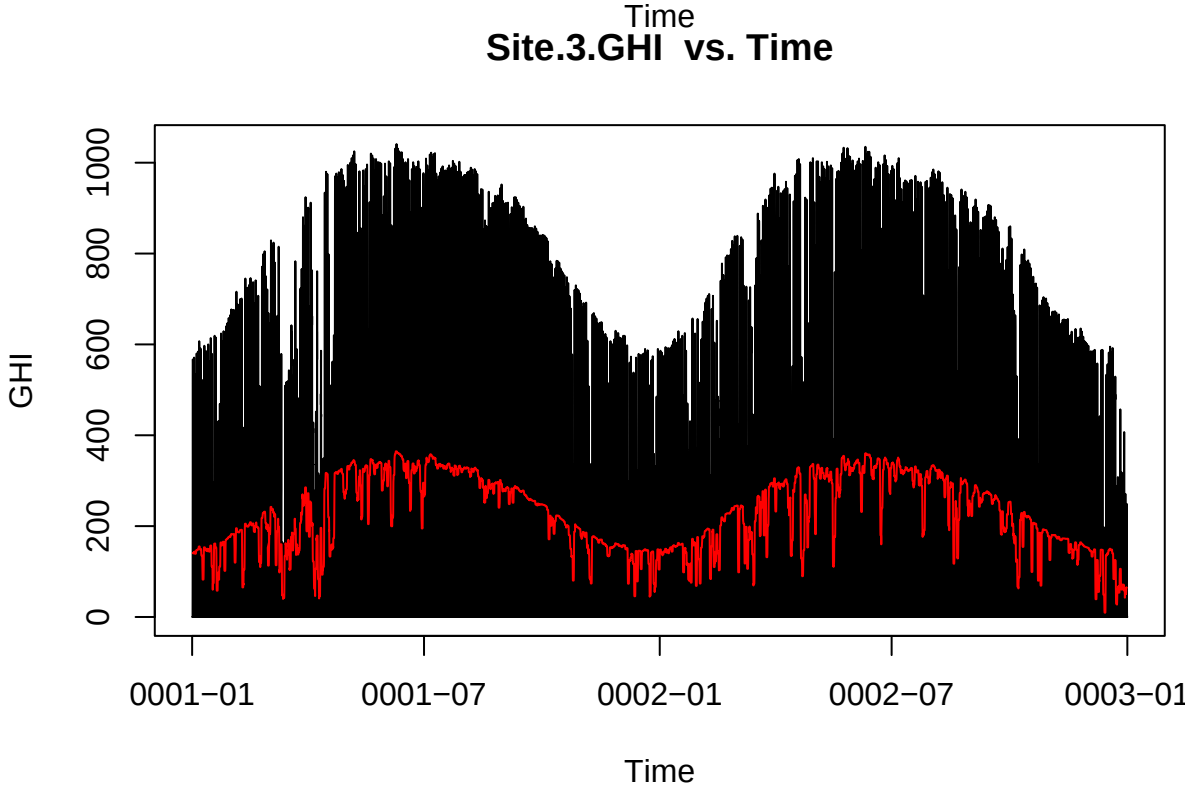

Site.4.Temp vs. Time
Temp
5
10
15
20
25
30
35
0001−01
0001−07
0002−01
0002−07
0003−01
Time

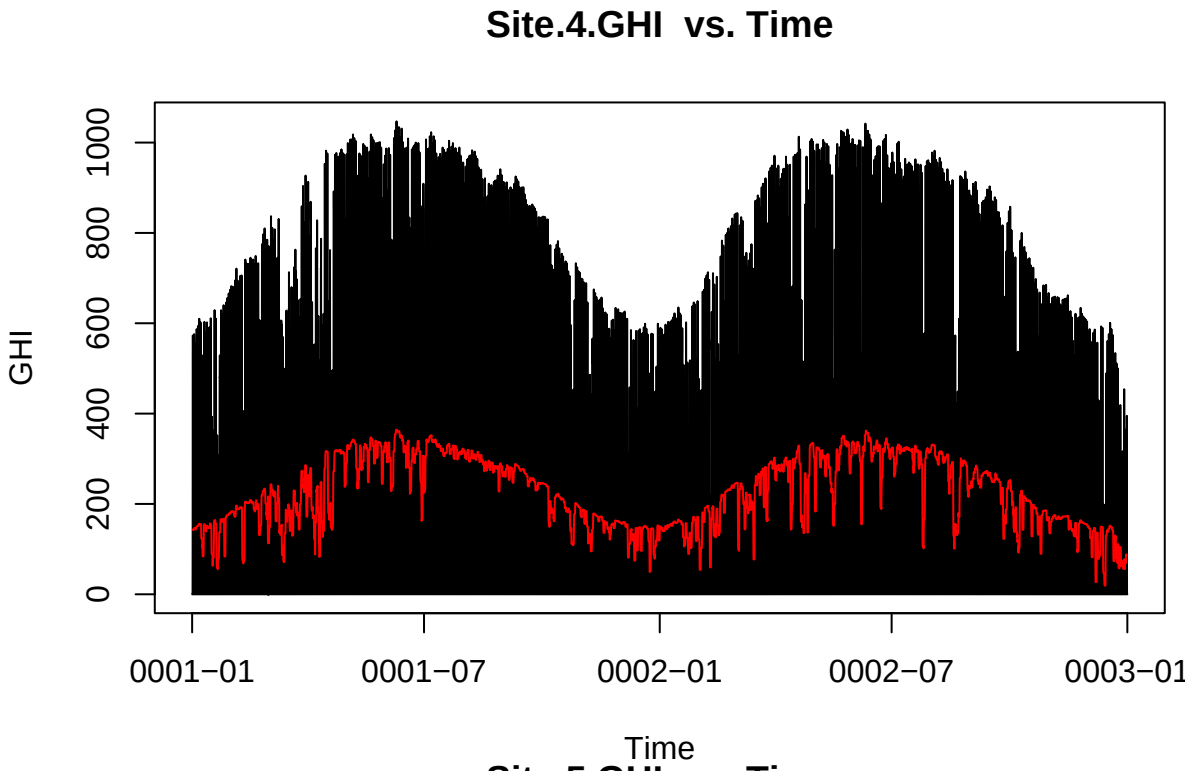

Site.5.Temp vs. Time
Temp
0
10
20
30
40
0001−01
0001−07
0002−01
0002−07
0003−01
Time

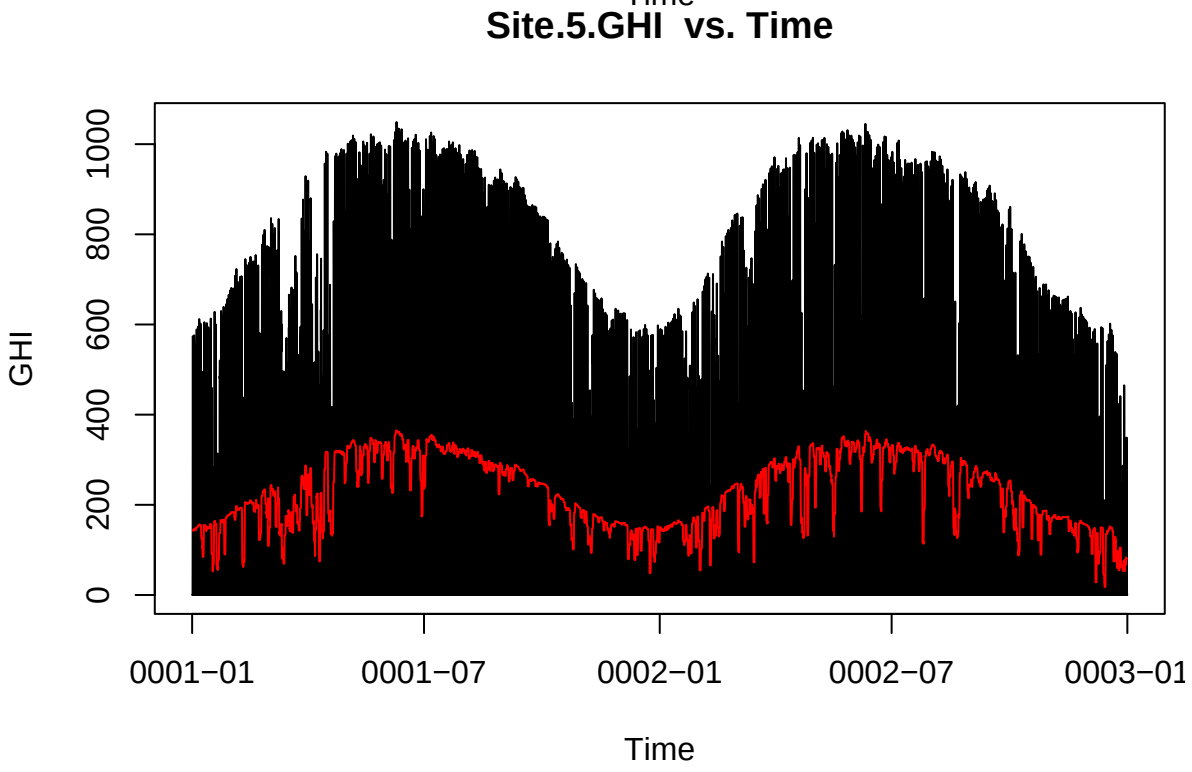
```r
# 5 GHI columns vs. Time
GHI_colnames = c("Site.1.GHI", "Site.2.GHI", "Site.3.GHI",
                 "Site.4.GHI", "Site.5.GHI")
for (colname in GHI_colnames) {
  # compute rolling average over 24h window
  ma = rollmean(train[,colname], fill=NA, k = 24)

  main_suffix = " vs. Time"
  plot(train$datetime, train[,colname], main = paste(colname, main_suffix),
       type = "l", ylab = "GHI", xlab = "Time")
  lines(train$datetime, ma, col = "red")
}
```



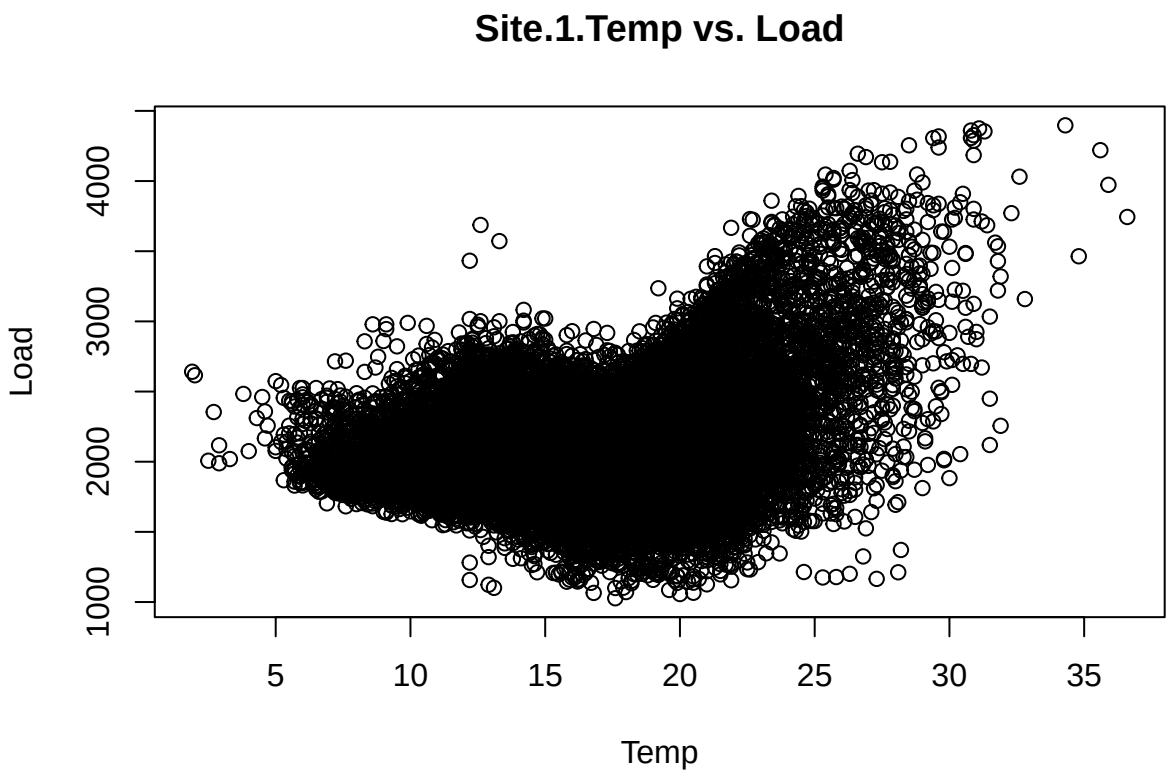

Site.2.GHI vs. Time
GHI
0
200
400
600
800
1000
0001−01
0001−07
0002−01
0002−07
0003−01
Time


Site.3.GHI vs. Time
GHI
0
200
400
600
800
1000
0001−01
0001−07
0002−01
0002−07
0003−01
Time

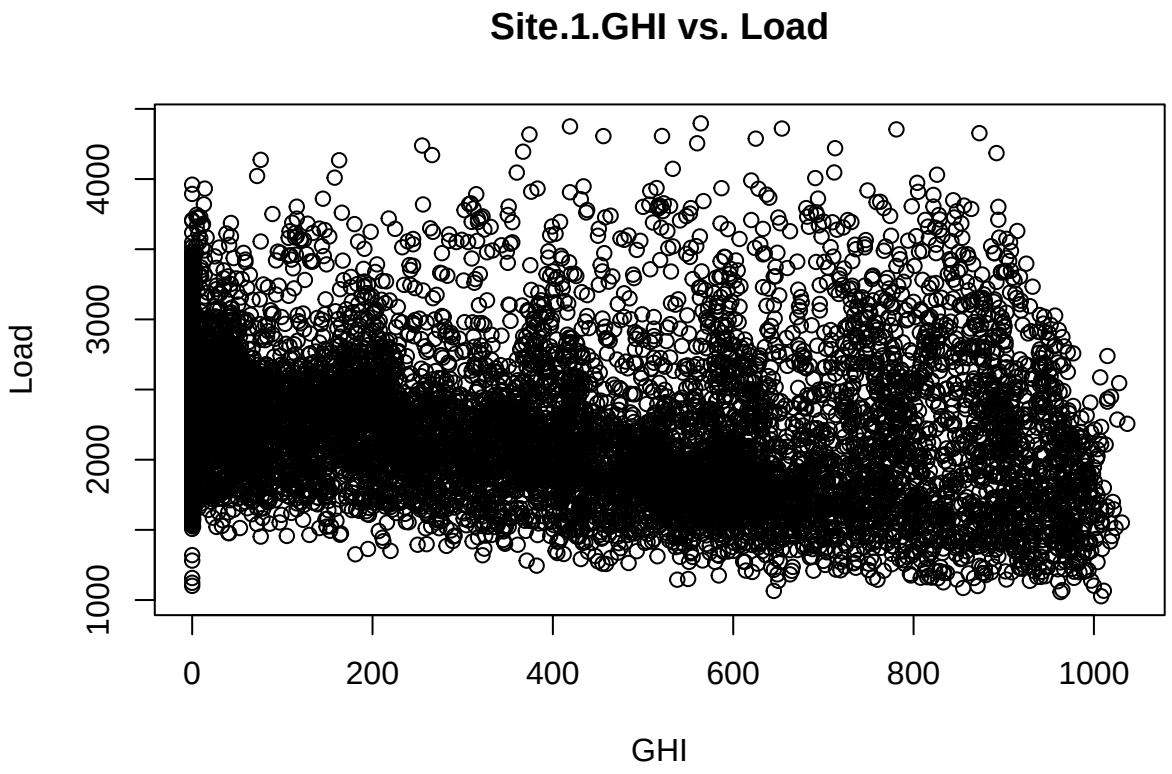

Site.4.GHI vs. Time
GHI
0
200
400
600
800
1000
0001−01
0001−07
0002−01
0002−07
0003−01
Time


Site.5.GHI vs. Time
GHI
0
200
400
600
800
1000
0001−01
0001−07
0002−01
0002−07
0003−01
Time

```r
# correlation
plot(train$Site.1.Temp, train$Load, main = "Site.1.Temp vs. Load",
     ylab = "Load", xlab = "Temp")
```




```r
plot(train$Site.1.GHI, train$Load, main = "Site.1.GHI vs. Load",
     ylab = "Load", xlab = "GHI")
```




```
plot(rollmean(train$Site.1.GHI, fill=NA, k = 24), train$Load,
     main = "Site.1.GHI Moving Average vs. Load", ylab = "Load", xlab = "GHI MA")
```