EDA

Jiashen Wang

2025-02-05

Variables:

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

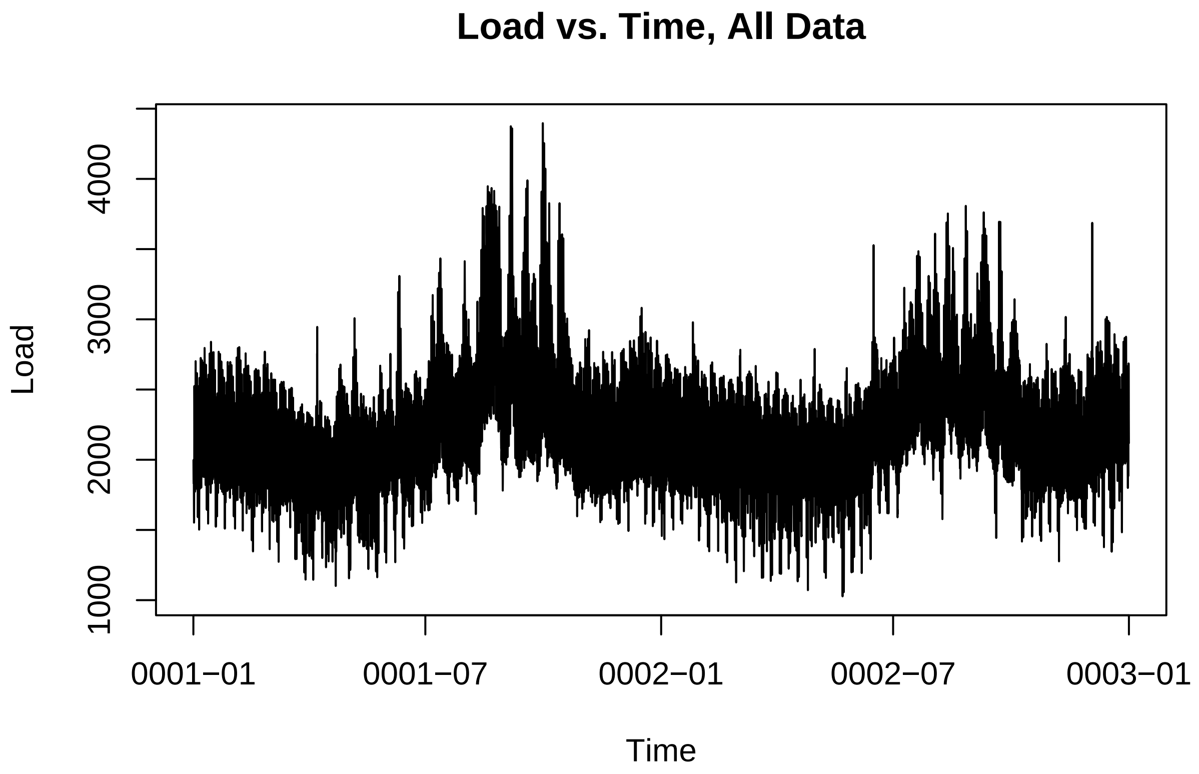
```
# data pre-processing
# create a column of datetime
train$datetime <- make_datetime(train$Year, train$Month, train$Day, train$Hour)
# convert load column to dbl
train$Load = as.numeric(gsub(",", "", train$Load))
```

```
head(train)
```

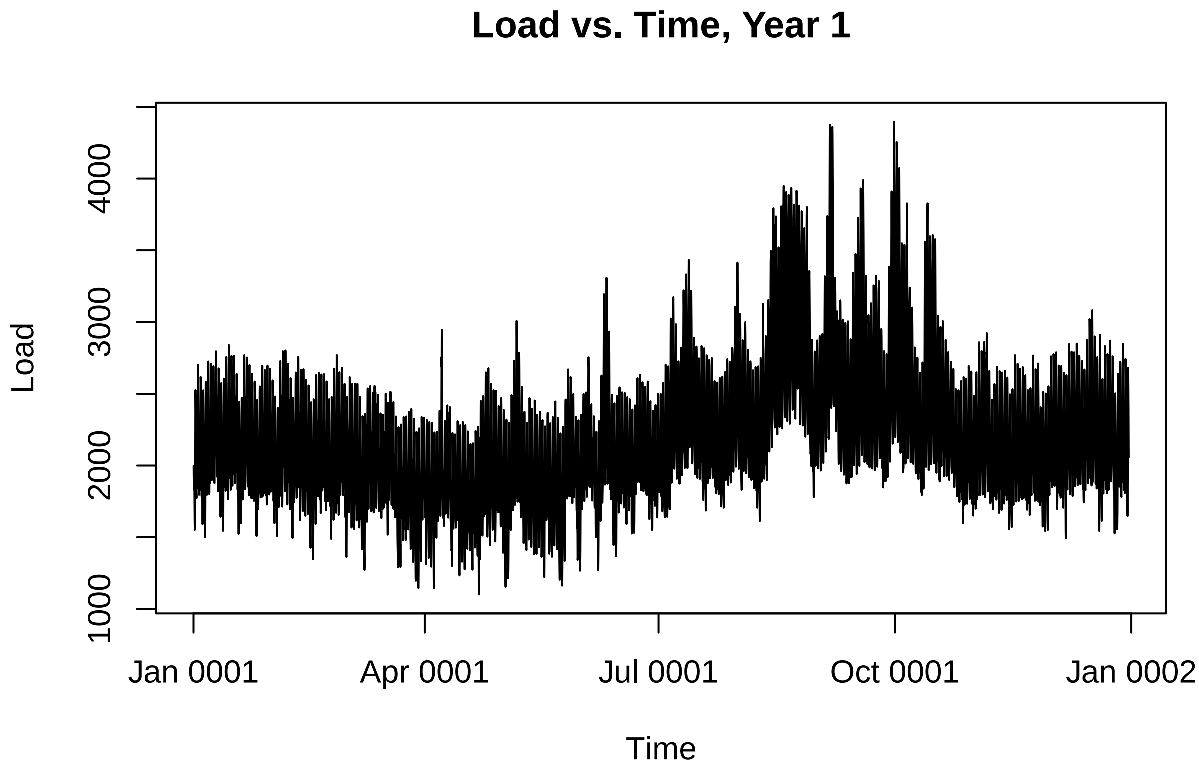
```
##   Year Month Day Hour Load Site.1.Temp Site.2.Temp Site.3.Temp Site.4.Temp
## 1    1     1   1    1 1997         8.0         8.2         5.3         9.4
## 2    1     1   1    2 1921         8.3         8.6         5.2         8.6
## 3    1     1   1    3 1861         8.1         8.8         5.1         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
## 1         8.1          0          0          0          0          0
## 2         7.1          0          0          0          0          0
## 3         6.2          0          0          0          0          0
## 4         6.0          0          0          0          0          0
## 5         6.9          0          0          0          0          0
## 6         7.3          0          0          0          0          0
##              datetime
## 1 0001-01-01 01:00:00
## 2 0001-01-01 02:00:00
## 3 0001-01-01 03:00:00
## 4 0001-01-01 04:00:00
## 5 0001-01-01 05:00:00
## 6 0001-01-01 06:00:00
```

EDA plots

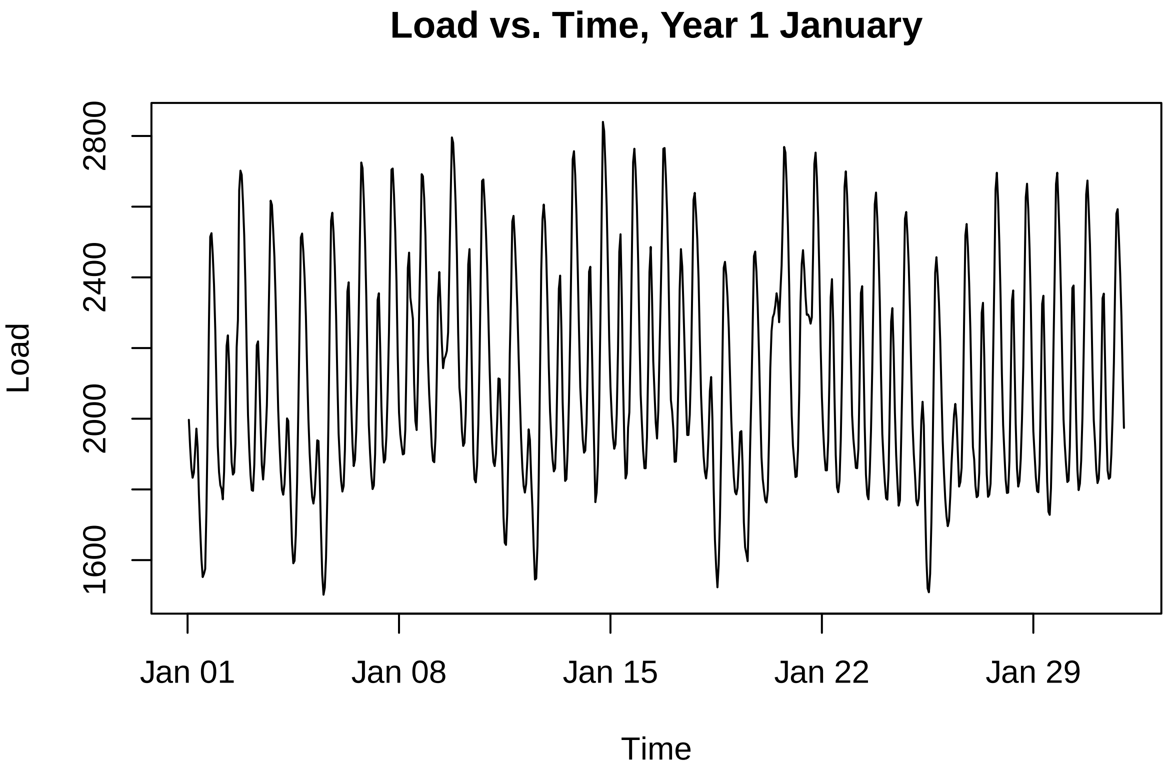
```
# Load vs. Time
plot(train$datetime, train$Load, type = "l", main = "Load vs. Time, All Data",
     ylab = "Load", xlab = "Time")
```



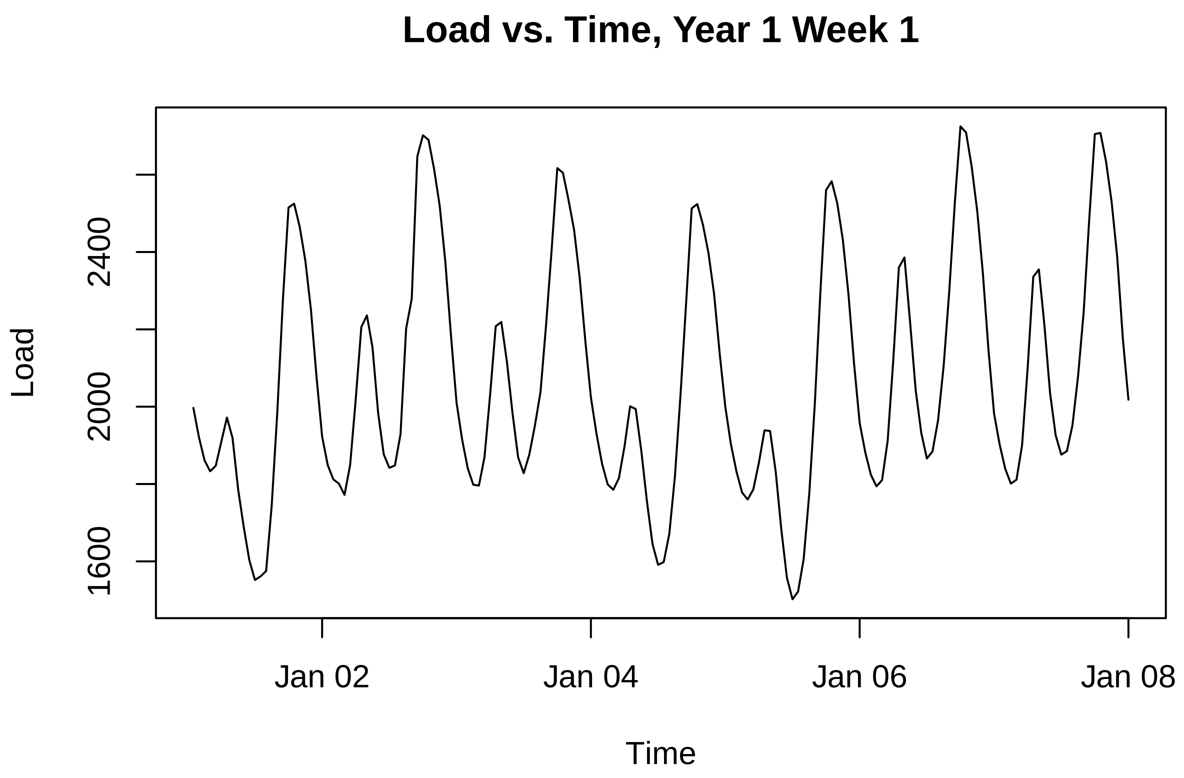
```
plot(train$datetime[1:8760], train$Load[1:8760], type = "l", main = "Load vs. Time, Year 1",
     ylab = "Load", xlab = "Time")
```



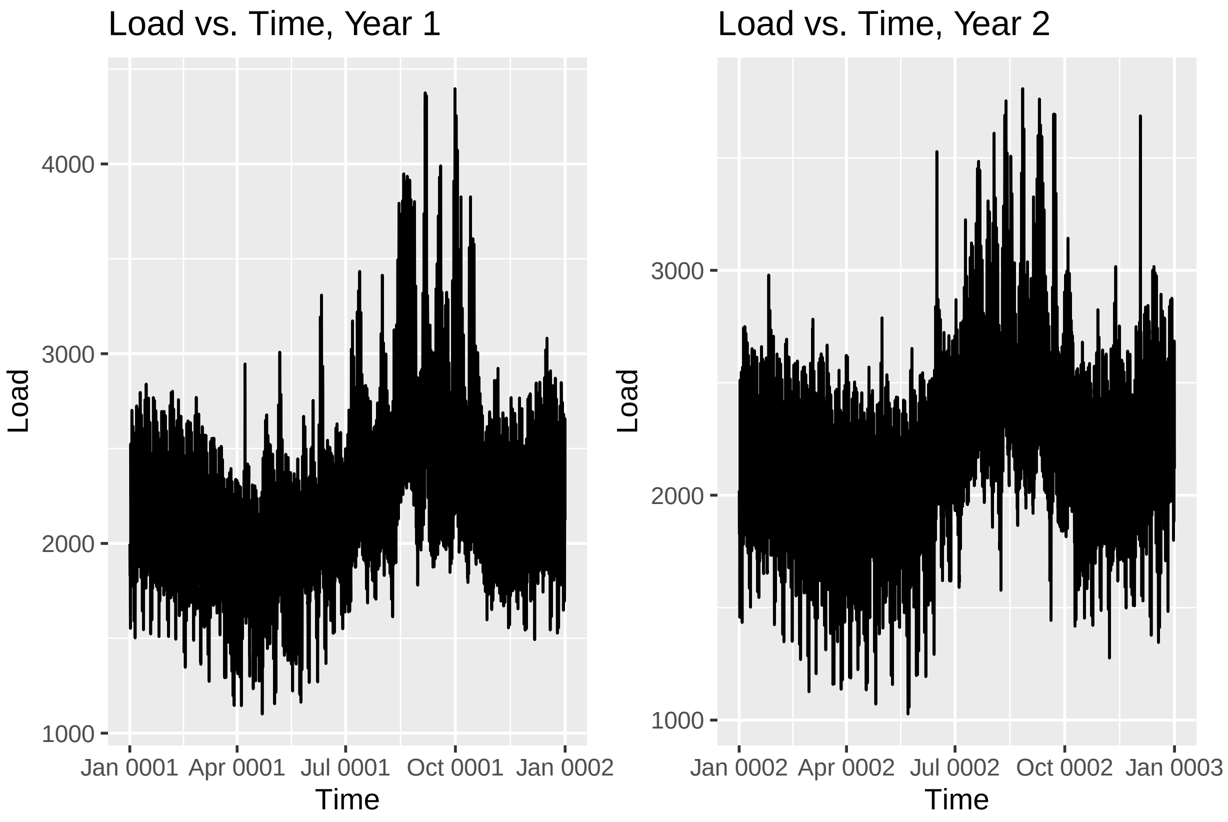
```
plot(train$datetime[1:744], train$Load[1:744], type = "l",
     main = "Load vs. Time, Year 1 January", ylab = "Load", xlab = "Time")
```



```
plot(train$datetime[1:168], train$Load[1:168], type = "l",
     main = "Load vs. Time, Year 1 Week 1", ylab = "Load", xlab = "Time")
```

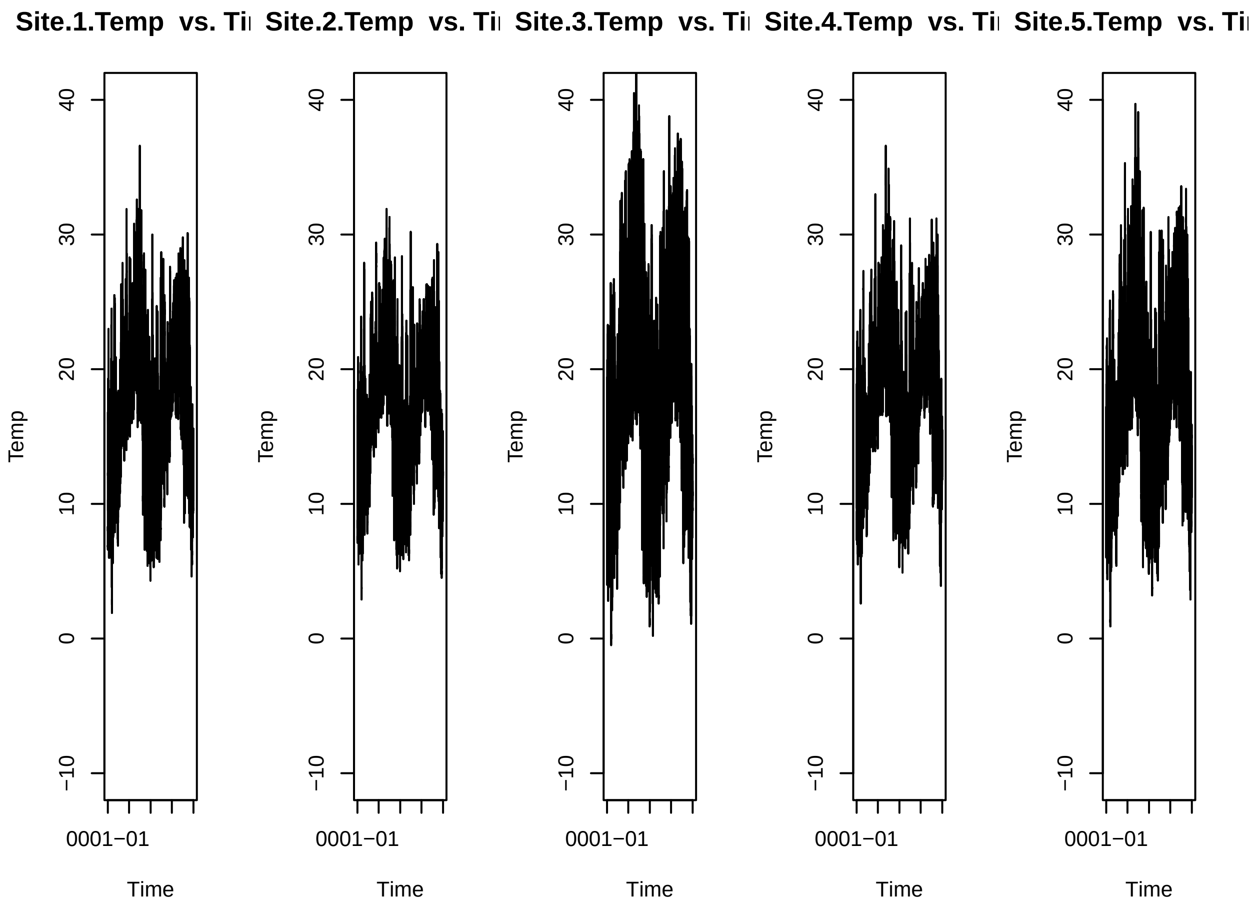


```
p_year1 <- train |>
  filter(Year == 1) |>
  ggplot(aes(x = datetime, y = Load)) +
  geom_line() + # Adding a line plot
  labs(
    title = "Load vs. Time, Year 1",
    x = "Time",
    y = "Load"
  )

p_year2 <- train |>
  filter(Year == 2) |>
  ggplot(aes(x = datetime, y = Load)) +
  geom_line() + # Adding a line plot
  labs(
    title = "Load vs. Time, Year 2",
    x = "Time",
    y = "Load"
  )


p_year1 | p_year2
```

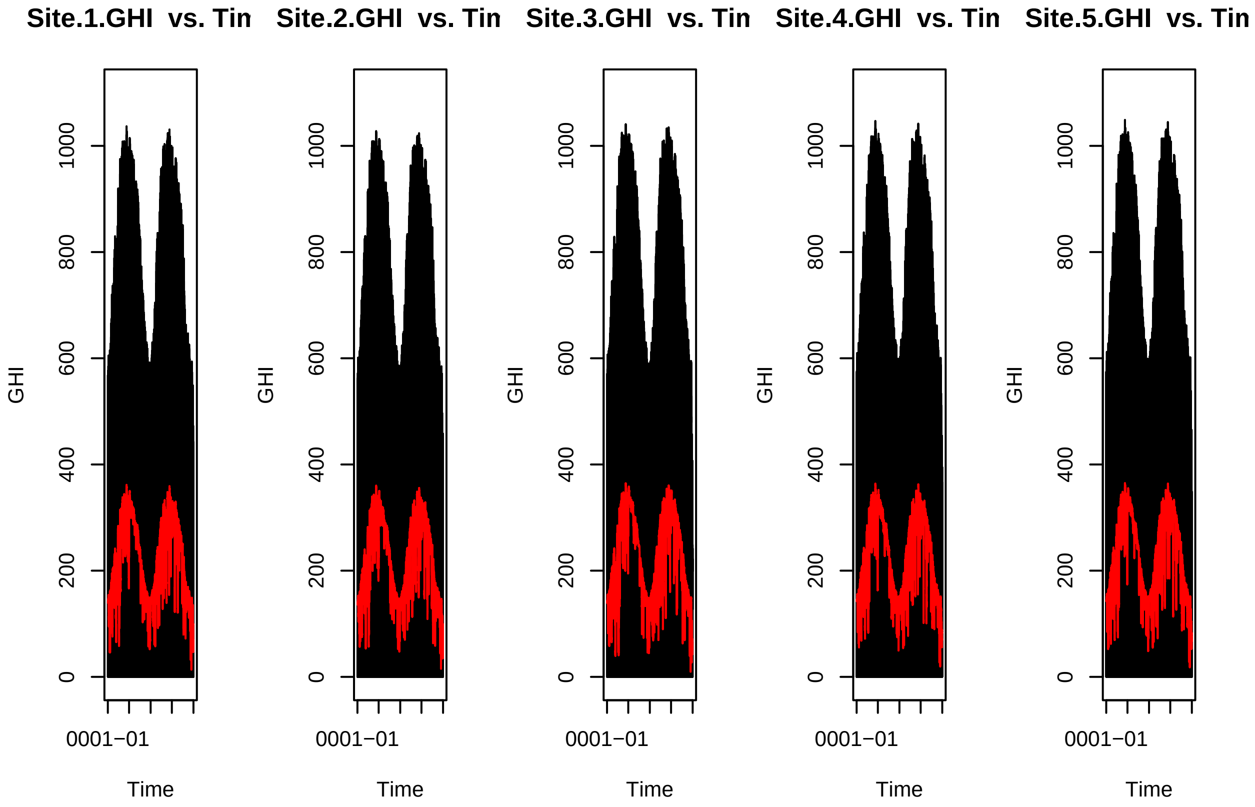


```r
plots_list <- list()  # initialize list

# loop through each month
for (month in 1:12) {
  # filter specific month
  plot_data <- train %>%
    filter(Year == 1, Month == month)

  # plot
  p <- ggplot(plot_data, aes(x = datetime, y = Load)) +
    geom_line() +
    labs(
      title = month,
      x = "Time",
      y = "Load"
    ) +
    scale_y_continuous(limits = c(1000, 4500))

  # append to list
  plots_list[[month]] <- p
}

plot_grid <- wrap_plots(plots_list, ncol = 6)

plot_grid
```

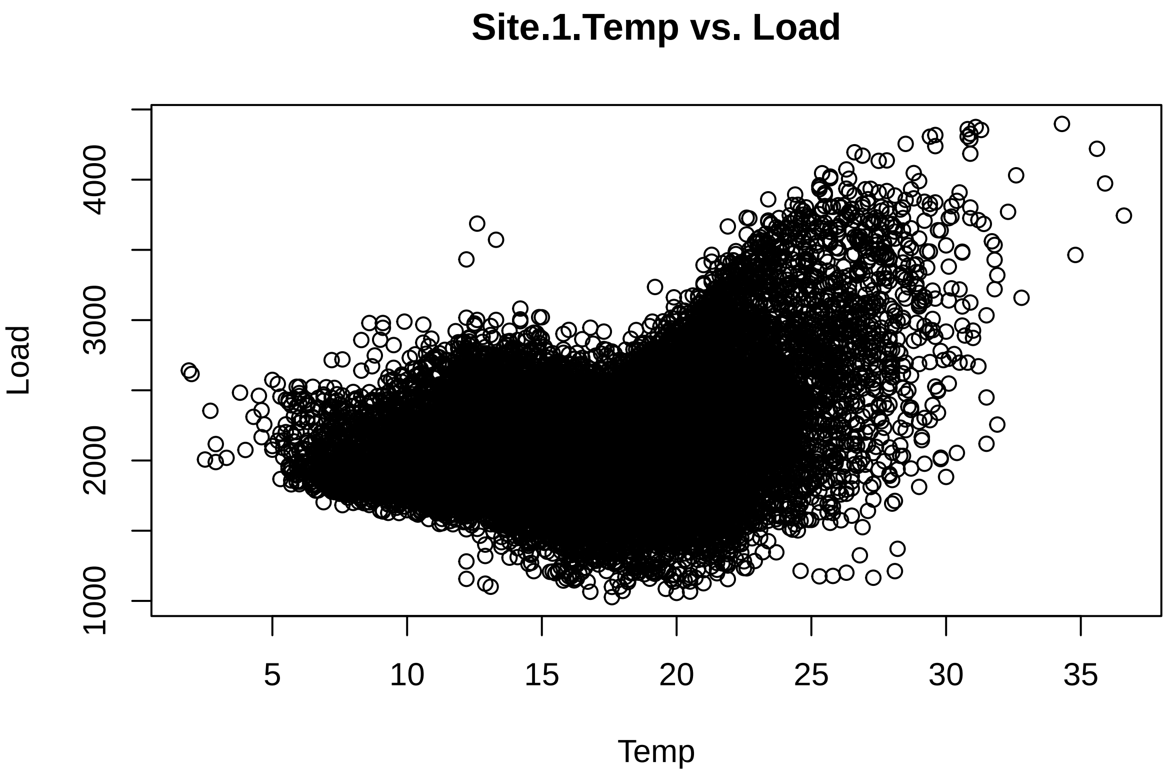
```
# 5 Temp columns vs. Time
temp_colnames = c("Site.1.Temp", "Site.2.Temp", "Site.3.Temp",
                  "Site.4.Temp", "Site.5.Temp")

par(mfrow = c(1, 5))
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", ylim = c(-10, 40))
}
```



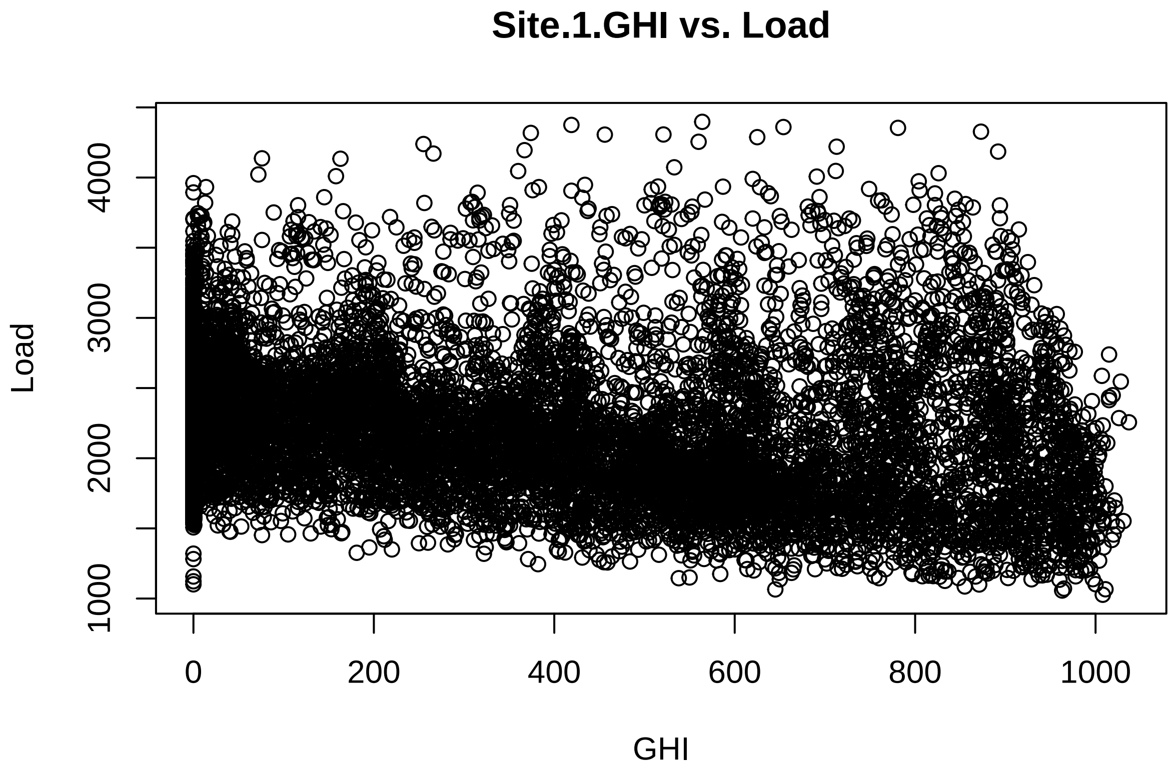
```
# 5 GHI columns vs. Time
GHI_colnames = c("Site.1.GHI", "Site.2.GHI", "Site.3.GHI",
                 "Site.4.GHI", "Site.5.GHI")

par(mfrow = c(1,5))
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$datetime, train[,colname], main = paste(colname, main_suffix),
       type = "l", ylab = "GHI", xlab = "Time", ylim = c(0, 1100))
  lines(train$datetime, ma, col = "red")
}
```



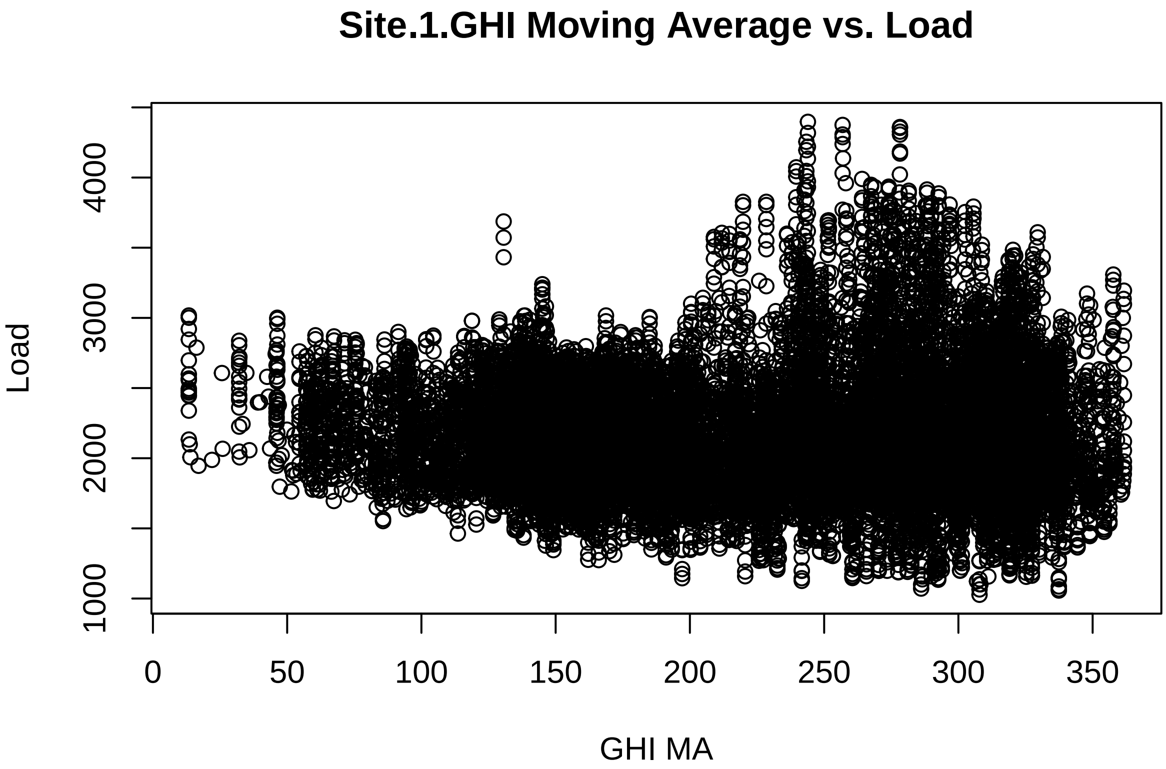
```
# correlation
plot(train$Site.1.Temp, train$Load, main = "Site.1.Temp vs. Load",
     ylab = "Load", xlab = "Temp")
```



```
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")
```


Site.1.GHI Moving Average vs. Load
Load
1000
2000
3000
4000
0
50
100
150
200
250
300
350
GHI MA