DPA_Project

Shriya, Girish, Ranjan, Raghukarn

2023-03-23

```
#remove.packages('vctrs')
#install.packages('rlang')
#install.packages('vctrs')
```

#Precipatation df1

```
Precipitation_124 <- read.table('precipitation_from_weather_station_124.txt',
header = TRUE, sep = ",")
Precipitation_124 <- subset(Precipitation_124, select = -c(Date_time))

head(Precipitation_124)
```

```
##     WY Year Month Day Hour Minute ppt_a perc_snow
## 1 2004 2003    10   1    0      0     0      1.00
## 2 2004 2003    10   1    1      0     0      1.00
## 3 2004 2003    10   1    2      0     0      1.00
## 4 2004 2003    10   1    3      0     0      1.00
## 5 2004 2003    10   1    4      0     0      1.00
## 6 2004 2003    10   1    5      0     0      0.95
```

```
summary(Precipitation_124)
```

```
##        WY            Year          Month             Day              Hour
##  Min.   :2004   Min.   :2003   Min.   : 1.000   Min.   : 1.00   Min.   :
0.00
##  1st Qu.:2006   1st Qu.:2006   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.75
##  Median :2009   Median :2009   Median : 7.000   Median :16.00   Median
:11.50
##  Mean   :2009   Mean   :2009   Mean   : 6.523   Mean   :15.73   Mean
:11.50
##  3rd Qu.:2012   3rd Qu.:2011   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.25
##  Max.   :2014   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.00
##      Minute      ppt_a            perc_snow
##  Min.   :0   Min.   : 0.00000   Min.   :0.0000
##  1st Qu.:0   1st Qu.: 0.00000   1st Qu.:0.0000
##  Median :0   Median : 0.00000   Median :1.0000
##  Mean   :0   Mean   : 0.06421   Mean   :0.6993
##  3rd Qu.:0   3rd Qu.: 0.00000   3rd Qu.:1.0000
##  Max.   :0   Max.   :17.10000   Max.   :1.0000
```

```
#check if NA's Exist
list_na <- colnames(Precipitation_124)[ apply(Precipitation_124, 2, anyNA) ]
list_na

## character(0)
```

#Precipatation df2

```
Precipitation_124b <-
read.table('precipitation_from_weather_station_124b.txt', header = TRUE, sep
= ",")
Precipitation_124b <- subset(Precipitation_124b, select = -c(Date_time,X))

head(Precipitation_124b)

##     WY Year Month Day Hour Minute ppt_a perc_snow
## 1 2004 2003    10   1    0      0     0      1.00
## 2 2004 2003    10   1    1      0     0      1.00
## 3 2004 2003    10   1    2      0     0      1.00
## 4 2004 2003    10   1    3      0     0      1.00
## 5 2004 2003    10   1    4      0     0      0.68
## 6 2004 2003    10   1    5      0     0      0.68

summary(Precipitation_124b)

##        WY            Year          Month             Day              Hour
##  Min.   :2004   Min.   :2003   Min.   : 1.000   Min.   : 1.00   Min.   :
0.00
##  1st Qu.:2006   1st Qu.:2006   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.75
##  Median :2009   Median :2009   Median : 7.000   Median :16.00   Median
:11.50
##  Mean   :2009   Mean   :2009   Mean   : 6.523   Mean   :15.73   Mean
:11.50
##  3rd Qu.:2012   3rd Qu.:2011   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.25
##  Max.   :2014   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.00
##      Minute      ppt_a            perc_snow
##  Min.   :0   Min.   : 0.00000   Min.   :0.0000
##  1st Qu.:0   1st Qu.: 0.00000   1st Qu.:0.0000
##  Median :0   Median : 0.00000   Median :1.0000
##  Mean   :0   Mean   : 0.07984   Mean   :0.6649
##  3rd Qu.:0   3rd Qu.: 0.00000   3rd Qu.:1.0000
##  Max.   :0   Max.   :17.60000   Max.   :1.0000

#check if NA's Exist
list_na <- colnames(Precipitation_124b)[ apply(Precipitation_124b, 2, anyNA)
]
list_na
```

```
## character(0)
```

#Precipatation df3

```
Precipitation_125 <- read.table('precipitation_from_weather_station_125.txt',
header = TRUE, sep = ",")
Precipitation_125 <- subset(Precipitation_125, select = -c(Date_time))

head(Precipitation_125)

##     WY Year Month Day Hour Minute ppt_a perc_snow
## 1 2004 2003    10   1    0      0     0      1.00
## 2 2004 2003    10   1    1      0     0      1.00
## 3 2004 2003    10   1    2      0     0      1.00
## 4 2004 2003    10   1    3      0     0      1.00
## 5 2004 2003    10   1    4      0     0      0.85
## 6 2004 2003    10   1    5      0     0      0.68

summary(Precipitation_125)

##        WY            Year          Month             Day              Hour
##  Min.   :2004   Min.   :2003   Min.   : 1.000   Min.   : 1.00   Min.   :
0.00
##  1st Qu.:2006   1st Qu.:2006   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.75
##  Median :2009   Median :2009   Median : 7.000   Median :16.00   Median
:11.50
##  Mean   :2009   Mean   :2009   Mean   : 6.523   Mean   :15.73   Mean
:11.50
##  3rd Qu.:2012   3rd Qu.:2011   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.25
##  Max.   :2014   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.00
##      Minute      ppt_a            perc_snow
##  Min.   :0   Min.   : 0.00000   Min.   :0.0000
##  1st Qu.:0   1st Qu.: 0.00000   1st Qu.:0.0000
##  Median :0   Median : 0.00000   Median :1.0000
##  Mean   :0   Mean   : 0.06428   Mean   :0.6343
##  3rd Qu.:0   3rd Qu.: 0.00000   3rd Qu.:1.0000
##  Max.   :0   Max.   :31.00000   Max.   :1.0000

#check if NA's Exist
list_na <- colnames(Precipitation_125)[ apply(Precipitation_125, 2, anyNA) ]
list_na

## character(0)
```

Precipatation mearged for all 3 df

```
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

#Merge all 3 precipitation datasets
Precipitation_merged<-bind_rows(Precipitation_124, Precipitation_124b,
Precipitation_125) %>%
          group_by(WY,Year,Month,Day,Hour,Minute) %>%
          summarise_each(funs(mean))

## Warning: `summarise_each_()` was deprecated in dplyr 0.7.0.
## ℹ Please use `across()` instead.
## ℹ The deprecated feature was likely used in the dplyr package.
##   Please report the issue at
<]8;;https://github.com/tidyverse/dplyr/issueshttps://github.com/tidyverse/dp
lyr/issues]8;;>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

## Warning: `funs()` was deprecated in dplyr 0.8.0.
## ℹ Please use a list of either functions or lambdas:
##
## # Simple named list: list(mean = mean, median = median)
##
## # Auto named with `tibble::lst()`: tibble::lst(mean, median)
##
## # Using lambdas list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

head(Precipitation_merged)

## # A tibble: 6 × 8
## # Groups:   WY, Year, Month, Day, Hour [6]
##      WY  Year Month   Day  Hour Minute ppt_a perc_snow
##   <int> <int> <int> <int> <int>  <int> <dbl>     <dbl>
## 1  2004  2003    10     1     0      0     0     1
## 2  2004  2003    10     1     1      0     0     1
## 3  2004  2003    10     1     2      0     0     1
## 4  2004  2003    10     1     3      0     0     1
## 5  2004  2003    10     1     4      0     0     0.843
## 6  2004  2003    10     1     5      0     0     0.77
```

```
summary(Precipitation_merged)
```

```
##        WY            Year          Month             Day              Hour
##  Min.   :2004   Min.   :2003   Min.   : 1.000   Min.   : 1.00   Min.   :
0.00
##  1st Qu.:2006   1st Qu.:2006   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.75
##  Median :2009   Median :2009   Median : 7.000   Median :16.00   Median
:11.50
##  Mean   :2009   Mean   :2009   Mean   : 6.523   Mean   :15.73   Mean
:11.50
##  3rd Qu.:2012   3rd Qu.:2011   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.25
##  Max.   :2014   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.00
##      Minute      ppt_a            perc_snow
##  Min.   :0   Min.   : 0.00000   Min.   :0.0000
##  1st Qu.:0   1st Qu.: 0.00000   1st Qu.:0.0500
##  Median :0   Median : 0.00000   Median :1.0000
##  Mean   :0   Mean   : 0.06945   Mean   :0.6662
##  3rd Qu.:0   3rd Qu.: 0.00000   3rd Qu.:1.0000
##  Max.   :0   Max.   :16.33333   Max.   :1.0000
```

```
#install.packages("writexl")
library("writexl")
```

```
## Warning: package 'writexl' was built under R version 4.2.3
```

```
write_xlsx(Precipitation_merged,"Precipitation_merged.xlsx")
```

#Weather df1

```
weather_data_124 <- read.table('weather_data_124.txt', header = TRUE, sep =
",")
weather_data_124 <- subset(weather_data_124, select = -c(Date_time))

head(weather_data_124)
```

```
##     WY Year Month Day Hour Minute  T_a   RH e_a  T_d S_i w_s   w_d
## 1 2004 2003    10   1    0      0 18.3 0.25 526 -1.8   0 1.7 130.2
## 2 2004 2003    10   1    1      0 18.9 0.25 546 -1.4   0 1.7 114.9
## 3 2004 2003    10   1    2      0 16.8 0.28 536 -1.6   0 1.6 262.1
## 4 2004 2003    10   1    3      0 16.8 0.30 574 -0.8   0 1.2 109.9
## 5 2004 2003    10   1    4      0 17.0 0.30 581 -0.6   0 2.0 102.9
## 6 2004 2003    10   1    5      0 16.7 0.31 589 -0.4   0 1.9 121.3
```

```
summary(weather_data_124)
```

```
##        WY            Year          Month             Day              Hour
##  Min.   :2004   Min.   :2003   Min.   : 1.000   Min.   : 1.00   Min.   :
0.00
##  1st Qu.:2006   1st Qu.:2006   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
```

```
5.75
##  Median :2009   Median :2009   Median : 7.000   Median :16.00   Median
:11.50
##  Mean   :2009   Mean   :2009   Mean   : 6.523   Mean   :15.73   Mean
:11.50
##  3rd Qu.:2012   3rd Qu.:2011   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.25
##  Max.   :2014   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.00
##      Minute       T_a                RH              e_a
##  Min.   :0   Min.   :-18.600   Min.   :0.0400   Min.   :  23.0
##  1st Qu.:0   1st Qu.: -0.500   1st Qu.:0.3400   1st Qu.: 360.0
##  Median :0   Median :  5.700   Median :0.5600   Median : 506.0
##  Mean   :0   Mean   :  7.018   Mean   :0.5642   Mean   : 533.6
##  3rd Qu.:0   3rd Qu.: 14.600   3rd Qu.:0.7900   3rd Qu.: 664.0
##  Max.   :0   Max.   : 33.200   Max.   :1.0000   Max.   :1836.0
##       T_d               S_i              w_s              w_d
##  Min.   :-34.600   Min.   :   0.0   Min.   : 0.400   Min.   :  0
##  1st Qu.: -6.300   1st Qu.:   0.0   1st Qu.: 2.300   1st Qu.:177
##  Median : -2.300   Median :   6.0   Median : 3.600   Median :240
##  Mean   : -2.569   Mean   : 193.6   Mean   : 4.460   Mean   :218
##  3rd Qu.:  1.100   3rd Qu.: 341.0   3rd Qu.: 5.825   3rd Qu.:271
##  Max.   : 16.100   Max.   :1102.0   Max.   :24.200   Max.   :360
```

```
#check if NA's Exist
list_na <- colnames(weather_data_124)[ apply(weather_data_124, 2, anyNA) ]
list_na
```

```
## character(0)
```

```
#check If missing values -9999 exist
any(weather_data_124==-9999)
```

```
## [1] FALSE
```

#Weather df2

```
weather_data_124b <- read.table('weather_data_124b.txt', header = TRUE, sep =
",")
weather_data_124b <- subset(weather_data_124b, select = -c(Date_time))

head(weather_data_124b)
```

```
##     WY Year Month Day Hour Minute  T_a   RH e_a  T_d S_i w_s   w_d
## 1 2004 2003    10   1    0      0 17.2 0.28 549 -1.3   0 0.7 255.5
## 2 2004 2003    10   1    1      0 16.3 0.30 556 -1.1   0 0.9 240.7
## 3 2004 2003    10   1    2      0 15.6 0.33 584 -0.5   0 0.7 142.0
## 4 2004 2003    10   1    3      0 14.2 0.36 582 -0.6   0 0.6   6.5
## 5 2004 2003    10   1    4      0 14.2 0.38 615  0.1   0 0.6 332.6
## 6 2004 2003    10   1    5      0 15.0 0.36 613  0.1   0 0.6 129.7
```

```
summary(weather_data_124b)
```

```
##        WY            Year          Month             Day              Hour
##  Min.   :2004   Min.   :2003   Min.   : 1.000   Min.   : 1.00   Min.   :
0.00
##  1st Qu.:2006   1st Qu.:2006   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.75
##  Median :2009   Median :2009   Median : 7.000   Median :16.00   Median
:11.50
##  Mean   :2009   Mean   :2009   Mean   : 6.523   Mean   :15.73   Mean
:11.50
##  3rd Qu.:2012   3rd Qu.:2011   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.25
##  Max.   :2014   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.00
##      Minute       T_a                RH              e_a
##  Min.   :0   Min.   :-19.600   Min.   :0.0500   Min.   :  48
##  1st Qu.:0   1st Qu.: -0.300   1st Qu.:0.3700   1st Qu.: 375
##  Median :0   Median :  5.700   Median :0.5700   Median : 525
##  Mean   :0   Mean   :  6.949   Mean   :0.5819   Mean   : 552
##  3rd Qu.:0   3rd Qu.: 14.000   3rd Qu.:0.8000   3rd Qu.: 688
##  Max.   :0   Max.   : 33.800   Max.   :1.0000   Max.   :1779
##       T_d               S_i              w_s              w_d
##  Min.   :-27.600   Min.   :   0.0   Min.   : 0.40   Min.   :  0
##  1st Qu.: -5.800   1st Qu.:   0.0   1st Qu.: 1.00   1st Qu.:169
##  Median : -1.800   Median :   5.0   Median : 1.60   Median :225
##  Mean   : -2.117   Mean   : 195.6   Mean   : 1.85   Mean   :217
##  3rd Qu.:  1.700   3rd Qu.: 338.0   3rd Qu.: 2.40   3rd Qu.:292
##  Max.   : 15.700   Max.   :1141.0   Max.   :16.30   Max.   :360
```

```
#check if NA's Exist
list_na <- colnames(weather_data_124b)[ apply(weather_data_124b, 2, anyNA) ]
list_na
```

```
## character(0)
```

```
#check If missing values -9999 exist
any(weather_data_124b==-9999)
```

```
## [1] FALSE
```

#Weather df3

```
weather_data_125 <- read.table('weather_data_125.txt', header = TRUE, sep =
",")
weather_data_125 <- subset(weather_data_125, select = -c(Date_time))

head(weather_data_125)
```

```
##     WY Year Month Day Hour Minute  T_a   RH e_a  T_d S_i w_s   w_d
## 1 2004 2003    10   1    0      0 13.7 0.34 533 -1.6   0 0.7 255.5
```

```
## 2 2004 2003    10   1    1      0 13.1 0.36 543 -1.4   0 0.9 240.7
## 3 2004 2003    10   1    2      0 12.4 0.39 562 -1.0   0 0.7 142.0
## 4 2004 2003    10   1    3      0 12.1 0.41 579 -0.7   0 0.6   6.5
## 5 2004 2003    10   1    4      0 12.7 0.41 602 -0.2   0 0.6 332.6
## 6 2004 2003    10   1    5      0 12.7 0.42 617  0.1   0 0.6 129.7

summary(weather_data_125)

##        WY            Year          Month             Day              Hour
##  Min.   :2004   Min.   :2003   Min.   : 1.000   Min.   : 1.00   Min.   :
0.00
##  1st Qu.:2006   1st Qu.:2006   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.75
##  Median :2009   Median :2009   Median : 7.000   Median :16.00   Median
:11.50
##  Mean   :2009   Mean   :2009   Mean   : 6.523   Mean   :15.73   Mean
:11.50
##  3rd Qu.:2012   3rd Qu.:2011   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.25
##  Max.   :2014   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.00
##      Minute       T_a                RH              e_a
##  Min.   :0   Min.   :-20.900   Min.   :0.0400   Min.   :  57
##  1st Qu.:0   1st Qu.:  0.600   1st Qu.:0.3600   1st Qu.: 398
##  Median :0   Median :  6.800   Median :0.5800   Median : 545
##  Mean   :0   Mean   :  8.084   Mean   :0.5686   Mean   : 574
##  3rd Qu.:0   3rd Qu.: 15.000   3rd Qu.:0.7700   3rd Qu.: 710
##  Max.   :0   Max.   : 36.900   Max.   :1.0000   Max.   :1932
##       T_d               S_i              w_s             w_d
##  Min.   :-26.100   Min.   :   0.0   Min.   :0.400   Min.   :  0.0
##  1st Qu.: -5.100   1st Qu.:   0.0   1st Qu.:0.900   1st Qu.:127.0
##  Median : -1.400   Median :   4.0   Median :1.600   Median :214.1
##  Mean   : -1.567   Mean   : 172.5   Mean   :1.778   Mean   :195.2
##  3rd Qu.:  2.100   3rd Qu.: 296.0   3rd Qu.:2.300   3rd Qu.:260.0
##  Max.   : 16.900   Max.   :1044.0   Max.   :9.700   Max.   :360.0

#check if NA's Exist
list_na <- colnames(weather_data_125)[apply(weather_data_125, 2, anyNA) ]
list_na

## character(0)

#check If missing values -9999 exist
any(weather_data_125==-9999)

## [1] FALSE
```

#Weather df4

```
library(ggplot2)
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.2.3
```

```
weather_data_jdt1 <- read.table('weather_data_jdt1.txt', header = TRUE, sep =
",")
weather_data_jdt1 <- subset(weather_data_jdt1, select = -c(Date_time))
```

```
head(weather_data_jdt1)
```

```
##     WY Year Month Day Hour Minute  T_a   RH e_a  T_d
## 1 2006 2005    11   5    0      0  0.2 0.87 539 -1.5
## 2 2006 2005    11   5    1      0  0.8 0.77 499 -2.4
## 3 2006 2005    11   5    2      0  0.3 0.75 468 -3.2
## 4 2006 2005    11   5    3      0 -0.4 0.73 431 -4.2
## 5 2006 2005    11   5    4      0 -1.1 0.79 441 -3.9
## 6 2006 2005    11   5    5      0 -1.2 0.76 420 -4.5
```

```
#check if NA's Exist
list_na <- colnames(weather_data_jdt1)[ apply(weather_data_jdt1, 2, anyNA) ]
list_na
```

```
## character(0)
```

```
#check If missing values -9999 exist
any(weather_data_jdt1==-9999)
```

```
## [1] TRUE
```

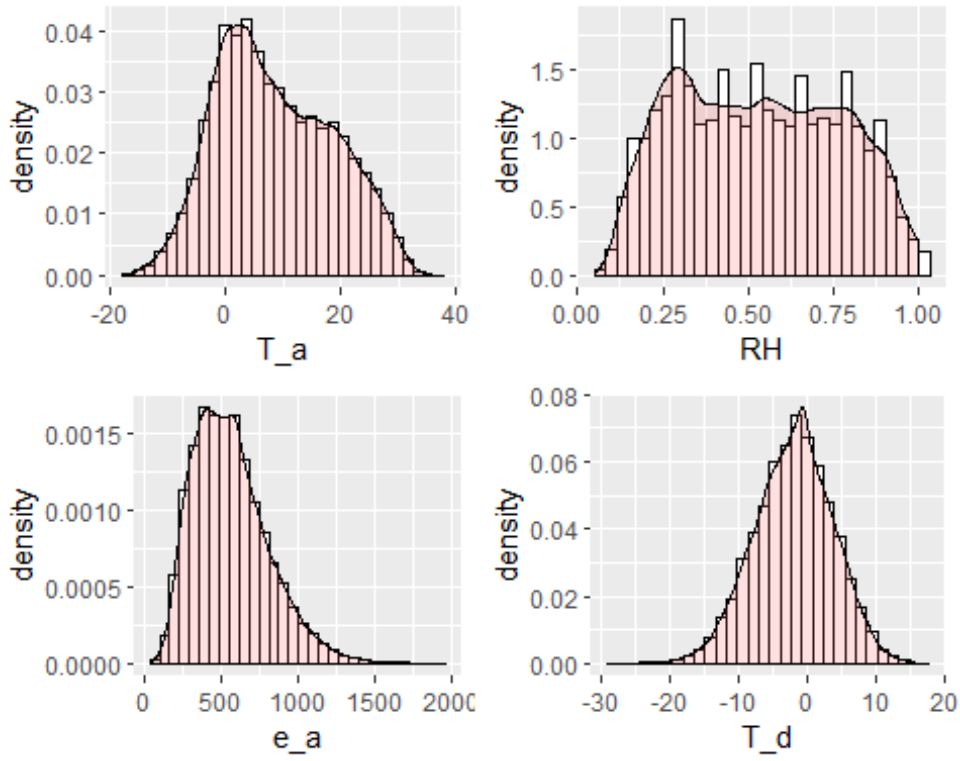
```
# replace -9999 with Na's
weather_data_jdt1 <- na_if(weather_data_jdt1, -9999)
#check if NA's Exist
list_na <- colnames(weather_data_jdt1)[ apply(weather_data_jdt1, 2, anyNA) ]
list_na
```

```
## [1] "T_a" "RH"  "e_a" "T_d"
```

```
#Density plot to see distribution of data within the feature visulization
T_a <- ggplot(weather_data_jdt1, aes(x=T_a)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

RH <- ggplot(weather_data_jdt1, aes(x=RH)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

e_a <- ggplot(weather_data_jdt1, aes(x=e_a)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

T_d <- ggplot(weather_data_jdt1, aes(x=T_d)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")
```

```
ggarrange(T_a,RH,e_a,T_d, ncol=2, nrow=2)
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2
3.4.0.
## ℹ Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 2391 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 2391 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 2391 rows containing non-finite values (`stat_bin()`).
## Removed 2391 rows containing non-finite values (`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 2391 rows containing non-finite values (`stat_bin()`).
## Removed 2391 rows containing non-finite values (`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 2391 rows containing non-finite values (`stat_bin()`).
## Removed 2391 rows containing non-finite values (`stat_density()`).
```




```
summary(weather_data_jdt1)

##        WY            Year          Month             Day              Hour
##  Min.   :2006   Min.   :2005   Min.   : 1.000   Min.   : 1.00   Min.   :
0.0
##  1st Qu.:2008   1st Qu.:2008   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.0
##  Median :2010   Median :2010   Median : 6.000   Median :16.00   Median
:11.0
##  Mean   :2010   Mean   :2010   Mean   : 6.485   Mean   :15.74   Mean
:11.5
##  3rd Qu.:2012   3rd Qu.:2012   3rd Qu.: 9.000   3rd Qu.:23.00   3rd
Qu.:17.0
##  Max.   :2015   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.0
##
##      Minute       T_a                RH              e_a
##  Min.   :0   Min.   :-17.600   Min.   :0.0500   Min.   :  43.0
##  1st Qu.:0   1st Qu.:  0.800   1st Qu.:0.3200   1st Qu.: 375.0
##  Median :0   Median :  7.300   Median :0.5200   Median : 528.0
##  Mean   :0   Mean   :  8.577   Mean   :0.5294   Mean   : 558.8
##  3rd Qu.:0   3rd Qu.: 16.300   3rd Qu.:0.7300   3rd Qu.: 701.0
##  Max.   :0   Max.   : 36.400   Max.   :1.0000   Max.   :1914.0
##              NA's   :2391      NA's   :2391     NA's   :2391
##       T_d
##  Min.   :-28.800
##  1st Qu.: -5.800
```

```
##  Median : -1.800
##  Mean   : -1.997
##  3rd Qu.:  1.900
##  Max.   : 16.800
##  NA's   :2391
```

#Find the median of values of that column that has NA

```
weather_data_jdt1$T_a[is.na(weather_data_jdt1$T_a)]<-
median(weather_data_jdt1$T_a, na.rm=TRUE)

weather_data_jdt1$RH[is.na(weather_data_jdt1$RH)]<-
median(weather_data_jdt1$RH, na.rm=TRUE)

weather_data_jdt1$e_a[is.na(weather_data_jdt1$e_a)]<-
median(weather_data_jdt1$e_a, na.rm=TRUE)

weather_data_jdt1$T_d[is.na(weather_data_jdt1$T_d)]<-
mean(weather_data_jdt1$T_d, na.rm=TRUE)

#check if NA's Exist
list_na <- colnames(weather_data_jdt1)[ apply(weather_data_jdt1, 2, anyNA) ]
list_na

## character(0)

#install.packages("ggpubr")
#install.packages("rlang")
#remove.packages("rlang")
#remove.packages("vtcrs")
```

#Weather df5

```
weather_data_jdt2 <- read.table('weather_data_jdt2.txt', header = TRUE, sep =
",")
weather_data_jdt2 <- subset(weather_data_jdt2, select = -c(Date_time))

head(weather_data_jdt2)

##     WY Year Month Day Hour Minute  T_a   RH e_a  T_d
## 1 2006 2005    11   5    0      0  0.2 0.83 515 -2.0
## 2 2006 2005    11   5    1      0  0.5 0.77 488 -2.7
## 3 2006 2005    11   5    2      0  0.0 0.75 458 -3.5
## 4 2006 2005    11   5    3      0 -0.6 0.71 413 -4.7
## 5 2006 2005    11   5    4      0 -1.4 0.79 430 -4.2
## 6 2006 2005    11   5    5      0 -1.6 0.77 412 -4.7

#check if NA's Exist
list_na <- colnames(weather_data_jdt2)[ apply(weather_data_jdt2, 2, anyNA) ]
list_na
```

```
## character(0)

#check If missing values -9999 exist
any(weather_data_jdt2==-9999)

## [1] TRUE

# replace -9999 with Na's
weather_data_jdt2 <- na_if(weather_data_jdt2, -9999)
#check if NA's Exist
list_na <- colnames(weather_data_jdt2)[ apply(weather_data_jdt2, 2, anyNA) ]
list_na

## [1] "T_a" "RH"  "e_a" "T_d"

summary(weather_data_jdt2)

##        WY            Year          Month             Day              Hour
##  Min.   :2006   Min.   :2005   Min.   : 1.000   Min.   : 1.00   Min.   :
0.0
##  1st Qu.:2008   1st Qu.:2008   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.0
##  Median :2010   Median :2010   Median : 6.000   Median :16.00   Median
:11.0
##  Mean   :2010   Mean   :2010   Mean   : 6.485   Mean   :15.74   Mean
:11.5
##  3rd Qu.:2012   3rd Qu.:2012   3rd Qu.: 9.000   3rd Qu.:23.00   3rd
Qu.:17.0
##  Max.   :2015   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.0
##
##      Minute       T_a                RH              e_a
##  Min.   :0   Min.   :-17.30   Min.   :0.0400   Min.   :  31.0
##  1st Qu.:0   1st Qu.:  0.60   1st Qu.:0.3100   1st Qu.: 359.0
##  Median :0   Median :  7.10   Median :0.5100   Median : 504.0
##  Mean   :0   Mean   :  8.37   Mean   :0.5141   Mean   : 532.7
##  3rd Qu.:0   3rd Qu.: 16.40   3rd Qu.:0.7100   3rd Qu.: 668.0
##  Max.   :0   Max.   : 34.90   Max.   :1.0000   Max.   :1722.0
##              NA's   :1002     NA's   :1002     NA's   :1002
##       T_d
##  Min.   :-31.900
##  1st Qu.: -6.300
##  Median : -2.300
##  Mean   : -2.594
##  3rd Qu.:  1.200
##  Max.   : 15.200
##  NA's   :1002
```

#Find the median of values of that column that has NA

```
weather_data_jdt2$T_a[is.na(weather_data_jdt2$T_a)]<-
median(weather_data_jdt2$T_a, na.rm=TRUE)
```

```
weather_data_jdt2$RH[is.na(weather_data_jdt2$RH)]<-
median(weather_data_jdt2$RH, na.rm=TRUE)

weather_data_jdt2$e_a[is.na(weather_data_jdt2$e_a)]<-
median(weather_data_jdt2$e_a, na.rm=TRUE)

weather_data_jdt2$T_d[is.na(weather_data_jdt2$T_d)]<-
mean(weather_data_jdt2$T_d, na.rm=TRUE)

#check if NA's Exist
list_na <- colnames(weather_data_jdt2)[ apply(weather_data_jdt2, 2, anyNA) ]
list_na

## character(0)
```

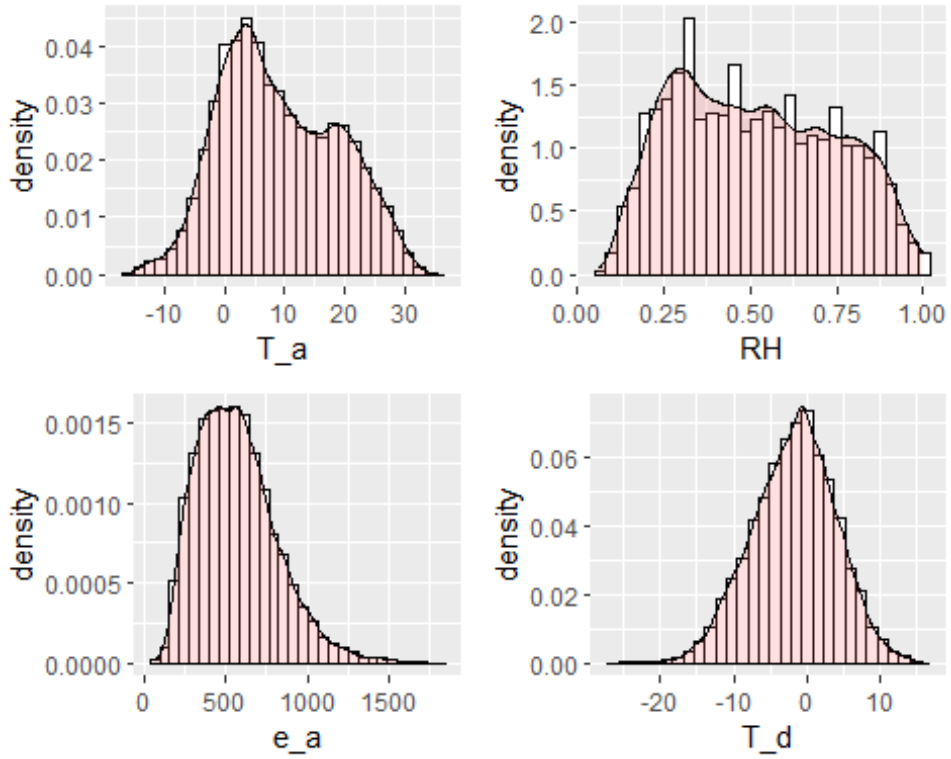
#Weather df6

```
weather_data_jdt2b <- read.table('weather_data_jdt2b.txt', header = TRUE, sep
= ",")
weather_data_jdt2b <- subset(weather_data_jdt2b, select = -c(Date_time))

head(weather_data_jdt2b)

##     WY Year Month Day Hour Minute   T_a    RH   e_a   T_d   w_s   w_d
## 1 2006 2005    11   5    0      0 -9999 -9999 -9999 -9999 -9999 -9999
## 2 2006 2005    11   5    1      0 -9999 -9999 -9999 -9999 -9999 -9999
## 3 2006 2005    11   5    2      0 -9999 -9999 -9999 -9999 -9999 -9999
## 4 2006 2005    11   5    3      0 -9999 -9999 -9999 -9999 -9999 -9999
## 5 2006 2005    11   5    4      0 -9999 -9999 -9999 -9999 -9999 -9999
## 6 2006 2005    11   5    5      0 -9999 -9999 -9999 -9999 -9999 -9999

#check if NA's Exist
list_na <- colnames(weather_data_jdt2b)[ apply(weather_data_jdt2b, 2, anyNA)
]
list_na

## character(0)

#check If missing values -9999 exist
any(weather_data_jdt2b==-9999)

## [1] TRUE

# replace -9999 with Na's
weather_data_jdt2b <- na_if(weather_data_jdt2b, -9999)
#check if NA's Exist
list_na <- colnames(weather_data_jdt2b)[ apply(weather_data_jdt2b, 2, anyNA)
]
list_na

## [1] "T_a" "RH"  "e_a" "T_d" "w_s" "w_d"
```

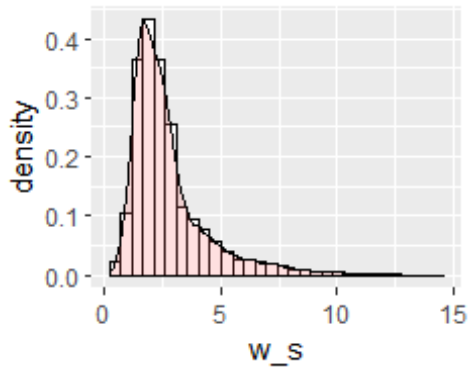
```r
#Density plot to see distribution of data within the feature Visulization
T_a <- ggplot(weather_data_jdt2b, aes(x=T_a)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

RH <- ggplot(weather_data_jdt2b, aes(x=RH)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

e_a <- ggplot(weather_data_jdt2b, aes(x=e_a)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

T_d <- ggplot(weather_data_jdt2b, aes(x=T_d)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

w_s <- ggplot(weather_data_jdt2b, aes(x=w_s)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

w_d <- ggplot(weather_data_jdt2b, aes(x=w_d)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

ggarrange(T_a,RH,e_a,T_d,w_s,w_d, ncol=2, nrow=2)
```

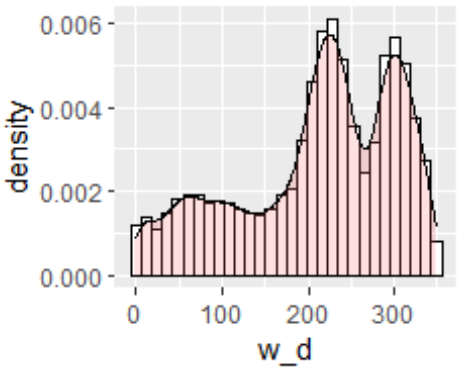
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 49303 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 49303 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 49303 rows containing non-finite values (`stat_bin()`).
## Removed 49303 rows containing non-finite values (`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 49303 rows containing non-finite values (`stat_bin()`).
## Removed 49303 rows containing non-finite values (`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 49303 rows containing non-finite values (`stat_bin()`).
## Removed 49303 rows containing non-finite values (`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 49233 rows containing non-finite values (`stat_bin()`).
```

```
## Warning: Removed 49233 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 50422 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 50422 rows containing non-finite values
(`stat_density()`).

## $`1`
```




```
##
## $`2`
```






```
##
## attr(,"class")
## [1] "list"      "ggarrange"

summary(weather_data_jdt2b)

##        WY            Year          Month             Day              Hour
##  Min.   :2006   Min.   :2005   Min.   : 1.000   Min.   : 1.00   Min.   :
0.0
##  1st Qu.:2008   1st Qu.:2008   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.0
##  Median :2010   Median :2010   Median : 6.000   Median :16.00   Median
:11.0
##  Mean   :2010   Mean   :2010   Mean   : 6.485   Mean   :15.74   Mean
:11.5
##  3rd Qu.:2012   3rd Qu.:2012   3rd Qu.: 9.000   3rd Qu.:23.00   3rd
Qu.:17.0
##  Max.   :2015   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.0
##
##      Minute       T_a                RH              e_a              T_d
##  Min.   :0   Min.   :-16.10   Min.   :0.06    Min.   :  55.0   Min.   :-
26.40
##  1st Qu.:0   1st Qu.:  1.60   1st Qu.:0.32    1st Qu.: 381.0   1st Qu.: -
5.60
##  Median :0   Median :  7.80   Median :0.50    Median : 539.0   Median : -
1.50
```

```
##  Mean   :0   Mean   :  9.08   Mean   :0.52    Mean   : 566.5   Mean   : -
1.81
##  3rd Qu.:0   3rd Qu.: 16.90   3rd Qu.:0.70    3rd Qu.: 711.0   3rd Qu.:
2.10
##  Max.   :0   Max.   : 36.20   Max.   :1.00    Max.   :1825.0   Max.   :
16.10
##              NA's   :49303    NA's   :49303   NA's   :49303    NA's
:49303
##       w_s              w_d
##  Min.   : 0.4    Min.   :  0.0
##  1st Qu.: 1.7    1st Qu.:153.5
##  Median : 2.3    Median :226.2
##  Mean   : 2.8    Mean   :210.7
##  3rd Qu.: 3.2    3rd Qu.:288.6
##  Max.   :14.3    Max.   :348.8
##  NA's   :49233   NA's   :50422
```

#Find the median of values of that column that has NA

```
weather_data_jdt2b$T_a[is.na(weather_data_jdt2b$T_a)]<-
median(weather_data_jdt2b$T_a, na.rm=TRUE)

weather_data_jdt2b$RH[is.na(weather_data_jdt2b$RH)]<-
median(weather_data_jdt2b$RH, na.rm=TRUE)

weather_data_jdt2b$e_a[is.na(weather_data_jdt2b$e_a)]<-
median(weather_data_jdt2b$e_a, na.rm=TRUE)

weather_data_jdt2b$T_d[is.na(weather_data_jdt2b$T_d)]<-
mean(weather_data_jdt2b$T_d, na.rm=TRUE)

weather_data_jdt2b$w_s[is.na(weather_data_jdt2b$w_s)]<-
median(weather_data_jdt2b$w_s, na.rm=TRUE)

weather_data_jdt2b$w_d[is.na(weather_data_jdt2b$w_d)]<-
median(weather_data_jdt2b$w_d, na.rm=TRUE)

#check if NA's Exist
list_na <- colnames(weather_data_jdt2b)[ apply(weather_data_jdt2b, 2, anyNA)
]
list_na

## character(0)
```

#Weather df7

```
weather_data_jdt3 <- read.table('weather_data_jdt3.txt', header = TRUE, sep =
",")
weather_data_jdt3 <- subset(weather_data_jdt3, select = -c(Date_time))
```

```
head(weather_data_jdt3)
```

```
##     WY Year Month Day Hour Minute  T_a   RH e_a  T_d w_s   w_d
## 1 2006 2005    11   5    0      0  0.0 0.84 513 -2.1 5.8 262.9
## 2 2006 2005    11   5    1      0 -0.1 0.80 485 -2.8 5.2 255.5
## 3 2006 2005    11   5    2      0 -0.6 0.79 459 -3.4 5.2 267.8
## 4 2006 2005    11   5    3      0 -1.1 0.72 401 -5.0 4.7 254.3
## 5 2006 2005    11   5    4      0 -1.9 0.84 438 -4.0 3.1 262.6
## 6 2006 2005    11   5    5      0 -2.1 0.79 405 -4.9 3.4 257.1
```

```
#check if NA's Exist
list_na <- colnames(weather_data_jdt3)[ apply(weather_data_jdt3, 2, anyNA) ]
list_na
```

```
## character(0)
```

```
#check If missing values -9999 exist
any(weather_data_jdt3==-9999)
```

```
## [1] TRUE
```

```
# replace -9999 with Na's
weather_data_jdt3 <- na_if(weather_data_jdt3, -9999)

#check if NA's Exist
list_na <- colnames(weather_data_jdt3)[ apply(weather_data_jdt3, 2, anyNA) ]
list_na
```

```
## [1] "T_a" "RH"  "e_a" "T_d" "w_s" "w_d"
```

```
summary(weather_data_jdt3)
```

```
##        WY            Year          Month             Day              Hour
##  Min.   :2006   Min.   :2005   Min.   : 1.000   Min.   : 1.00   Min.   :
0.0
##  1st Qu.:2008   1st Qu.:2008   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.0
##  Median :2010   Median :2010   Median : 6.000   Median :16.00   Median
:11.0
##  Mean   :2010   Mean   :2010   Mean   : 6.485   Mean   :15.74   Mean
:11.5
##  3rd Qu.:2012   3rd Qu.:2012   3rd Qu.: 9.000   3rd Qu.:23.00   3rd
Qu.:17.0
##  Max.   :2015   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.0
##
##      Minute       T_a                RH              e_a
##  Min.   :0   Min.   :-17.100   Min.   :0.0400   Min.   :  25
##  1st Qu.:0   1st Qu.:  0.500   1st Qu.:0.3100   1st Qu.: 356
##  Median :0   Median :  6.900   Median :0.5100   Median : 500
##  Mean   :0   Mean   :  8.168   Mean   :0.5196   Mean   : 530
```

```
##  3rd Qu.:0   3rd Qu.: 16.100   3rd Qu.:0.7200   3rd Qu.: 665
##  Max.   :0   Max.   : 35.000   Max.   :1.0000   Max.   :1736
##              NA's   :1019      NA's   :1019     NA's   :1019
##       T_d               w_s              w_d
##  Min.   :-34.000   Min.   : 0.400   Min.   :  0.1
##  1st Qu.: -6.400   1st Qu.: 1.800   1st Qu.:174.6
##  Median : -2.400   Median : 2.600   Median :234.4
##  Mean   : -2.661   Mean   : 2.691   Mean   :206.0
##  3rd Qu.:  1.200   3rd Qu.: 3.400   3rd Qu.:269.8
##  Max.   : 15.300   Max.   :10.400   Max.   :360.0
##  NA's   :1019      NA's   :35       NA's   :58
```

#Find the median of values of that column that has NA

```
weather_data_jdt3$T_a[is.na(weather_data_jdt3$T_a)]<-
median(weather_data_jdt3$T_a, na.rm=TRUE)

weather_data_jdt3$RH[is.na(weather_data_jdt3$RH)]<-
median(weather_data_jdt3$RH, na.rm=TRUE)

weather_data_jdt3$e_a[is.na(weather_data_jdt3$e_a)]<-
median(weather_data_jdt3$e_a, na.rm=TRUE)

weather_data_jdt3$T_d[is.na(weather_data_jdt3$T_d)]<-
mean(weather_data_jdt3$T_d, na.rm=TRUE)

weather_data_jdt3$w_s[is.na(weather_data_jdt3$w_s)]<-
median(weather_data_jdt3$w_s, na.rm=TRUE)

weather_data_jdt3$w_d[is.na(weather_data_jdt3$w_d)]<-
median(weather_data_jdt3$w_d, na.rm=TRUE)

#check if NA's Exist
list_na <- colnames(weather_data_jdt3)[ apply(weather_data_jdt3, 2, anyNA) ]
list_na

## character(0)
```

#Weather df8

```
weather_data_jdt3b <- read.table('weather_data_jdt3b.txt', header = TRUE, sep
= ",")
weather_data_jdt3b <- subset(weather_data_jdt3b, select = -c(Date_time))

#check if NA's Exist
list_na <- colnames(weather_data_jdt3b)[ apply(weather_data_jdt3b, 2, anyNA)
]
list_na

## character(0)
```

```
#check If missing values -9999 exist
any(weather_data_jdt3b==-9999)

## [1] TRUE

# replace -9999 with Na's
weather_data_jdt3b <- na_if(weather_data_jdt3b, -9999)
any(weather_data_jdt3b==-9999)

## [1] NA

#check if NA's Exist
list_na <- colnames(weather_data_jdt3b)[ apply(weather_data_jdt3b, 2, anyNA)
]
list_na

## [1] "T_a" "RH"  "e_a" "T_d" "w_s" "w_d"

summary(weather_data_jdt3b)

##        WY            Year          Month              Day              Hour
##  Min.   :2006   Min.   :2005   Min.   : 1.000   Min.   : 1.00   Min.   :
0.0
##  1st Qu.:2008   1st Qu.:2008   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.0
##  Median :2010   Median :2010   Median : 6.000   Median :16.00   Median
:11.0
##  Mean   :2010   Mean   :2010   Mean   : 6.485   Mean   :15.74   Mean
:11.5
##  3rd Qu.:2012   3rd Qu.:2012   3rd Qu.: 9.000   3rd Qu.:23.00   3rd
Qu.:17.0
##  Max.   :2015   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.0
##
##      Minute       T_a                RH             e_a              T_d
##  Min.   :0   Min.   :-16.60   Min.   :0.05    Min.   :  62.0   Min.   :-
25.20
##  1st Qu.:0   1st Qu.:  0.80   1st Qu.:0.31    1st Qu.: 374.0   1st Qu.: -
5.80
##  Median :0   Median :  6.80   Median :0.49    Median : 528.0   Median : -
1.80
##  Mean   :0   Mean   :  8.41   Mean   :0.51    Mean   : 555.2   Mean   : -
2.07
##  3rd Qu.:0   3rd Qu.: 16.50   3rd Qu.:0.71    3rd Qu.: 697.0   3rd Qu.:
1.80
##  Max.   :0   Max.   : 35.30   Max.   :1.00    Max.   :1702.0   Max.   :
15.00
##              NA's   :47373    NA's   :49306   NA's   :49306    NA's
:49306
##       w_s              w_d
##  Min.   : 0.40   Min.   :  0.1
```

```
##  1st Qu.: 2.00   1st Qu.:140.9
##  Median : 2.70   Median :231.4
##  Mean   : 3.07   Mean   :208.4
##  3rd Qu.: 3.60   3rd Qu.:286.2
##  Max.   :15.60   Max.   :348.5
##  NA's   :47355   NA's   :47648
```

#Find the median of values of that column that has NA

```
weather_data_jdt3b$T_a[is.na(weather_data_jdt3b$T_a)]<-
median(weather_data_jdt3b$T_a, na.rm=TRUE)

weather_data_jdt3b$RH[is.na(weather_data_jdt3b$RH)]<-
median(weather_data_jdt3b$RH, na.rm=TRUE)

weather_data_jdt3b$e_a[is.na(weather_data_jdt3b$e_a)]<-
median(weather_data_jdt3b$e_a, na.rm=TRUE)

weather_data_jdt3b$T_d[is.na(weather_data_jdt3b$T_d)]<-
mean(weather_data_jdt3b$T_d, na.rm=TRUE)

weather_data_jdt3b$w_s[is.na(weather_data_jdt3b$w_s)]<-
median(weather_data_jdt3b$w_s, na.rm=TRUE)

weather_data_jdt3b$w_d[is.na(weather_data_jdt3b$w_d)]<-
median(weather_data_jdt3b$w_d, na.rm=TRUE)

#check if NA's Exist
list_na <- colnames(weather_data_jdt3b)[ apply(weather_data_jdt3b, 2, anyNA)
]
list_na

## character(0)
```

#Weather df9

```
weather_data_jdt4 <- read.table('weather_data_jdt4.txt', header = TRUE, sep =
",")
weather_data_jdt4 <- subset(weather_data_jdt4, select = -c(Date_time))

head(weather_data_jdt4)

##     WY Year Month Day Hour Minute  T_a   RH e_a  T_d
## 1 2006 2005    11   5    0      0 -0.6 0.88 511 -2.1
## 2 2006 2005    11   5    1      0 -0.4 0.82 485 -2.8
## 3 2006 2005    11   5    2      0 -0.9 0.81 459 -3.4
## 4 2006 2005    11   5    3      0 -1.5 0.77 415 -4.6
## 5 2006 2005    11   5    4      0 -2.3 0.85 429 -4.2
## 6 2006 2005    11   5    5      0 -2.4 0.81 405 -4.9
```

```r
#check if NA's Exist
list_na <- colnames(weather_data_jdt4)[ apply(weather_data_jdt4, 2, anyNA) ]
list_na
```

```
## character(0)
```

```r
#check If missing values -9999 exist
any(weather_data_jdt4==-9999)
```

```
## [1] TRUE
```

```r
# replace -9999 with Na's
weather_data_jdt4 <- na_if(weather_data_jdt4, -9999)
any(weather_data_jdt4==-9999)
```

```
## [1] NA
```

```r
#check if NA's Exist
list_na <- colnames(weather_data_jdt4)[ apply(weather_data_jdt4, 2, anyNA) ]
list_na
```

```
## [1] "T_a" "RH"  "e_a" "T_d"
```

```r
summary(weather_data_jdt4)
```

```
##        WY            Year          Month             Day              Hour
##  Min.   :2006   Min.   :2005   Min.   : 1.000   Min.   : 1.00   Min.   :
0.0
##  1st Qu.:2008   1st Qu.:2008   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.0
##  Median :2010   Median :2010   Median : 6.000   Median :16.00   Median
:11.0
##  Mean   :2010   Mean   :2010   Mean   : 6.485   Mean   :15.74   Mean
:11.5
##  3rd Qu.:2012   3rd Qu.:2012   3rd Qu.: 9.000   3rd Qu.:23.00   3rd
Qu.:17.0
##  Max.   :2015   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.0
##
##      Minute       T_a                RH              e_a
##  Min.   :0   Min.   :-17.300   Min.   :0.0300   Min.   :  22.0
##  1st Qu.:0   1st Qu.:  0.200   1st Qu.:0.3100   1st Qu.: 355.0
##  Median :0   Median :  6.600   Median :0.5100   Median : 501.0
##  Mean   :0   Mean   :  8.042   Mean   :0.5285   Mean   : 531.6
##  3rd Qu.:0   3rd Qu.: 16.400   3rd Qu.:0.7400   3rd Qu.: 668.0
##  Max.   :0   Max.   : 34.300   Max.   :1.0000   Max.   :1736.0
##              NA's   :1823      NA's   :1823     NA's   :1823
##       T_d
##  Min.   :-35.100
##  1st Qu.: -6.400
##  Median : -2.400
##  Mean   : -2.647
```

```
##  3rd Qu.:  1.200
##  Max.   : 15.300
##  NA's   :1823
```

#Find the median of values of that column that has NA

```
weather_data_jdt4$T_a[is.na(weather_data_jdt4$T_a)]<-
median(weather_data_jdt4$T_a, na.rm=TRUE)

weather_data_jdt4$RH[is.na(weather_data_jdt4$RH)]<-
median(weather_data_jdt4$RH, na.rm=TRUE)

weather_data_jdt4$e_a[is.na(weather_data_jdt4$e_a)]<-
median(weather_data_jdt4$e_a, na.rm=TRUE)

weather_data_jdt4$T_d[is.na(weather_data_jdt4$T_d)]<-
mean(weather_data_jdt4$T_d, na.rm=TRUE)

#check if NA's Exist
list_na <- colnames(weather_data_jdt4)[ apply(weather_data_jdt4, 2, anyNA) ]
list_na
```

```
## character(0)
```

#Weather df10

```
weather_data_jdt4b <- read.table('weather_data_jdt4b.txt', header = TRUE, sep
= ",")
weather_data_jdt4b <- subset(weather_data_jdt4b, select = -c(Date_time))

head(weather_data_jdt4b)
```

```
##     WY Year Month Day Hour Minute   T_a    RH   e_a   T_d   w_s   w_d
## 1 2006 2005    11   5    0      0 -9999 -9999 -9999 -9999 -9999 -9999
## 2 2006 2005    11   5    1      0 -9999 -9999 -9999 -9999 -9999 -9999
## 3 2006 2005    11   5    2      0 -9999 -9999 -9999 -9999 -9999 -9999
## 4 2006 2005    11   5    3      0 -9999 -9999 -9999 -9999 -9999 -9999
## 5 2006 2005    11   5    4      0 -9999 -9999 -9999 -9999 -9999 -9999
## 6 2006 2005    11   5    5      0 -9999 -9999 -9999 -9999 -9999 -9999
```

```
#check if NA's Exist
list_na <- colnames(weather_data_jdt4b)[ apply(weather_data_jdt4b, 2, anyNA)
]
list_na
```

```
## character(0)
```

```
#check If missing values -9999 exist
any(weather_data_jdt4b==-9999)
```

```
## [1] TRUE
```

```
# replace -9999 with Na's
weather_data_jdt4b <- na_if(weather_data_jdt4b, -9999)
#check if NA's Exist
list_na <- colnames(weather_data_jdt4b)[ apply(weather_data_jdt4b, 2, anyNA)
]
list_na

## [1] "T_a" "RH"  "e_a" "T_d" "w_s" "w_d"

summary(weather_data_jdt4b)

##        WY            Year          Month             Day              Hour
##  Min.   :2006   Min.   :2005   Min.   : 1.000   Min.   : 1.00   Min.   :
0.0
##  1st Qu.:2008   1st Qu.:2008   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.0
##  Median :2010   Median :2010   Median : 6.000   Median :16.00   Median
:11.0
##  Mean   :2010   Mean   :2010   Mean   : 6.485   Mean   :15.74   Mean
:11.5
##  3rd Qu.:2012   3rd Qu.:2012   3rd Qu.: 9.000   3rd Qu.:23.00   3rd
Qu.:17.0
##  Max.   :2015   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.0
##
##      Minute       T_a               RH             e_a              T_d
##  Min.   :0   Min.   :-17.10   Min.   :0.05    Min.   :  57.0   Min.   :-
26.00
##  1st Qu.:0   1st Qu.:  1.20   1st Qu.:0.31    1st Qu.: 372.0   1st Qu.: -
5.90
##  Median :0   Median :  7.40   Median :0.49    Median : 525.0   Median : -
1.80
##  Mean   :0   Mean   :  8.82   Mean   :0.52    Mean   : 550.6   Mean   : -
2.17
##  3rd Qu.:0   3rd Qu.: 16.80   3rd Qu.:0.72    3rd Qu.: 690.0   3rd Qu.:
1.70
##  Max.   :0   Max.   : 36.10   Max.   :1.00    Max.   :1697.0   Max.   :
14.90
##              NA's   :49305    NA's   :49305   NA's   :49305    NA's
:49305
##       w_s              w_d
##  Min.   : 0.4    Min.   :  0.0
##  1st Qu.: 1.7    1st Qu.:188.0
##  Median : 2.3    Median :240.8
##  Mean   : 2.9    Mean   :225.1
##  3rd Qu.: 3.6    3rd Qu.:292.4
##  Max.   :15.2    Max.   :349.4
##  NA's   :49763   NA's   :49844
```

#Find the median of values of that column that has NA

```
weather_data_jdt4b$T_a[is.na(weather_data_jdt4b$T_a)]<-
median(weather_data_jdt4b$T_a, na.rm=TRUE)

weather_data_jdt4b$RH[is.na(weather_data_jdt4b$RH)]<-
median(weather_data_jdt4b$RH, na.rm=TRUE)

weather_data_jdt4b$e_a[is.na(weather_data_jdt4b$e_a)]<-
median(weather_data_jdt4b$e_a, na.rm=TRUE)

weather_data_jdt4b$T_d[is.na(weather_data_jdt4b$T_d)]<-
mean(weather_data_jdt4b$T_d, na.rm=TRUE)

weather_data_jdt4b$w_s[is.na(weather_data_jdt4b$w_s)]<-
median(weather_data_jdt4b$w_s, na.rm=TRUE)

weather_data_jdt4b$w_d[is.na(weather_data_jdt4b$w_d)]<-
median(weather_data_jdt4b$w_d, na.rm=TRUE)

#check if NA's Exist
list_na <- colnames(weather_data_jdt4b)[ apply(weather_data_jdt4b, 2, anyNA)
]
list_na

## character(0)
```

#Weather df11

```
weather_data_jdt5 <- read.table('weather_data_jdt5.txt', header = TRUE, sep =
",")
weather_data_jdt5 <- subset(weather_data_jdt5, select = -c(Date_time))

head(weather_data_jdt5)

##     WY Year Month Day Hour Minute  T_a   RH e_a  T_d
## 1 2006 2005    11   5    0      0 -0.9 0.87 493 -2.6
## 2 2006 2005    11   5    1      0 -1.0 0.85 478 -2.9
## 3 2006 2005    11   5    2      0 -1.5 0.84 453 -3.6
## 4 2006 2005    11   5    3      0 -2.1 0.81 416 -4.6
## 5 2006 2005    11   5    4      0 -2.6 0.86 423 -4.4
## 6 2006 2005    11   5    5      0 -2.8 0.82 397 -5.1

#check if NA's Exist
list_na <- colnames(weather_data_jdt5)[ apply(weather_data_jdt5, 2, anyNA) ]
list_na

## character(0)

#check If missing values -9999 exist
any(weather_data_jdt5==-9999)

## [1] TRUE
```

```
# replace -9999 with Na's
weather_data_jdt5 <- na_if(weather_data_jdt5, -9999)
#check if NA's Exist
list_na <- colnames(weather_data_jdt5)[ apply(weather_data_jdt5, 2, anyNA) ]
list_na

## [1] "T_a" "RH"  "e_a" "T_d"

summary(weather_data_jdt5)

##        WY            Year          Month              Day              Hour
##  Min.   :2006   Min.   :2005   Min.   : 1.000   Min.   : 1.00   Min.   :
0.0
##  1st Qu.:2008   1st Qu.:2008   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.0
##  Median :2010   Median :2010   Median : 6.000   Median :16.00   Median
:11.0
##  Mean   :2010   Mean   :2010   Mean   : 6.485   Mean   :15.74   Mean
:11.5
##  3rd Qu.:2012   3rd Qu.:2012   3rd Qu.: 9.000   3rd Qu.:23.00   3rd
Qu.:17.0
##  Max.   :2015   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.0
##
##      Minute       T_a               RH              e_a
##  Min.   :0   Min.   :-19.40   Min.   :0.0500   Min.   :  40.0
##  1st Qu.:0   1st Qu.: -0.20   1st Qu.:0.3300   1st Qu.: 356.0
##  Median :0   Median :  6.00   Median :0.5400   Median : 499.0
##  Mean   :0   Mean   :  7.35   Mean   :0.5467   Mean   : 526.4
##  3rd Qu.:0   3rd Qu.: 14.90   3rd Qu.:0.7600   3rd Qu.: 658.0
##  Max.   :0   Max.   : 34.40   Max.   :1.0000   Max.   :1814.0
##              NA's   :998      NA's   :999      NA's   :999
##       T_d
##  Min.   :-29.500
##  1st Qu.: -6.400
##  Median : -2.400
##  Mean   : -2.727
##  3rd Qu.:  1.000
##  Max.   : 16.000
##  NA's   :999
```

#Find the median of values of that column that has NA

```
weather_data_jdt5$T_a[is.na(weather_data_jdt5$T_a)]<-
median(weather_data_jdt5$T_a, na.rm=TRUE)

weather_data_jdt5$RH[is.na(weather_data_jdt5$RH)]<-
median(weather_data_jdt5$RH, na.rm=TRUE)

weather_data_jdt5$e_a[is.na(weather_data_jdt5$e_a)]<-
median(weather_data_jdt5$e_a, na.rm=TRUE)
```

```
weather_data_jdt5$T_d[is.na(weather_data_jdt5$T_d)]<-
mean(weather_data_jdt5$T_d, na.rm=TRUE)

#check if NA's Exist
list_na <- colnames(weather_data_jdt5)[ apply(weather_data_jdt5, 2, anyNA) ]
list_na

## character(0)
```

#Weather mearged for all 11 df which contains all 11 stations

```
library(dplyr)
weather_data_merged<-bind_rows(weather_data_124, weather_data_124b,
weather_data_125, weather_data_jdt1,weather_data_jdt2, weather_data_jdt2b,
weather_data_jdt3, weather_data_jdt3b, weather_data_jdt3b, weather_data_jdt4,
weather_data_jdt4b, weather_data_jdt5) %>%
          group_by(WY,Year,Month,Day,Hour,Minute) %>%
          summarise_each(funs(mean))
head(weather_data_merged)

## # A tibble: 6 × 13
## # Groups:   WY, Year, Month, Day, Hour [6]
##      WY  Year Month   Day  Hour Minute   T_a    RH   e_a     T_d   S_i
w_s
##   <int> <int> <int> <int> <int>  <int> <dbl> <dbl> <dbl>   <dbl> <dbl>
<dbl>
## 1  2004  2003    10     1     0      0  16.4 0.29   536  -1.57       0
1.03
## 2  2004  2003    10     1     1      0  16.1 0.303  548. -1.3        0
1.17
## 3  2004  2003    10     1     2      0  14.9 0.333  561. -1.03       0  1
## 4  2004  2003    10     1     3      0  14.4 0.357  578. -0.7        0
0.8
## 5  2004  2003    10     1     4      0  14.6 0.363  599. -0.233      0
1.07
## 6  2004  2003    10     1     5      0  14.8 0.363  606. -0.0667     0
1.03
## # … with 1 more variable: w_d <dbl>

summary(weather_data_merged)

##        WY            Year          Month             Day             Hour
##  Min.   :2004   Min.   :2003   Min.   : 1.000   Min.   : 1.00   Min.   :
0.0
##  1st Qu.:2006   1st Qu.:2006   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.0
##  Median :2009   Median :2009   Median : 7.000   Median :16.00   Median
:11.0
##  Mean   :2009   Mean   :2009   Mean   : 6.523   Mean   :15.73   Mean
:11.5
```

```
##  3rd Qu.:2012   3rd Qu.:2011   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.0
##  Max.   :2015   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.0
##
##      Minute       T_a                RH               e_a
##  Min.   :0   Min.   :-16.792   Min.   :0.06333   Min.   :  61.17
##  1st Qu.:0   1st Qu.:  1.725   1st Qu.:0.37500   1st Qu.: 410.33
##  Median :0   Median :  6.642   Median :0.53333   Median : 522.42
##  Mean   :0   Mean   :  7.758   Mean   :0.53987   Mean   : 548.29
##  3rd Qu.:0   3rd Qu.: 13.633   3rd Qu.:0.69917   3rd Qu.: 652.75
##  Max.   :0   Max.   : 34.717   Max.   :1.00000   Max.   :1716.75
##
##       T_d                S_i               w_s             w_d
##  Min.   :-25.3583   Min.   :   0.00   Min.   :0.43    Min.   :  0.77
##  1st Qu.: -4.9687   1st Qu.:   0.00   1st Qu.:1.30    1st Qu.:132.09
##  Median : -1.9667   Median :   5.33   Median :2.03    Median :213.07
##  Mean   : -2.0857   Mean   : 180.49   Mean   :2.32    Mean   :197.03
##  3rd Qu.:  0.8167   3rd Qu.: 309.42   3rd Qu.:2.97    3rd Qu.:255.84
##  Max.   : 15.1167   Max.   :1040.33   Max.   :9.87    Max.   :359.33
##                     NA's   :78049     NA's   :78049   NA's   :78049
```

#download the weather datasets which is merged into excel for further analysis.

```
library("writexl")
write_xlsx(weather_data_merged,"weather_data_merged.xlsx")
```

#snow depth df for all 11 stations

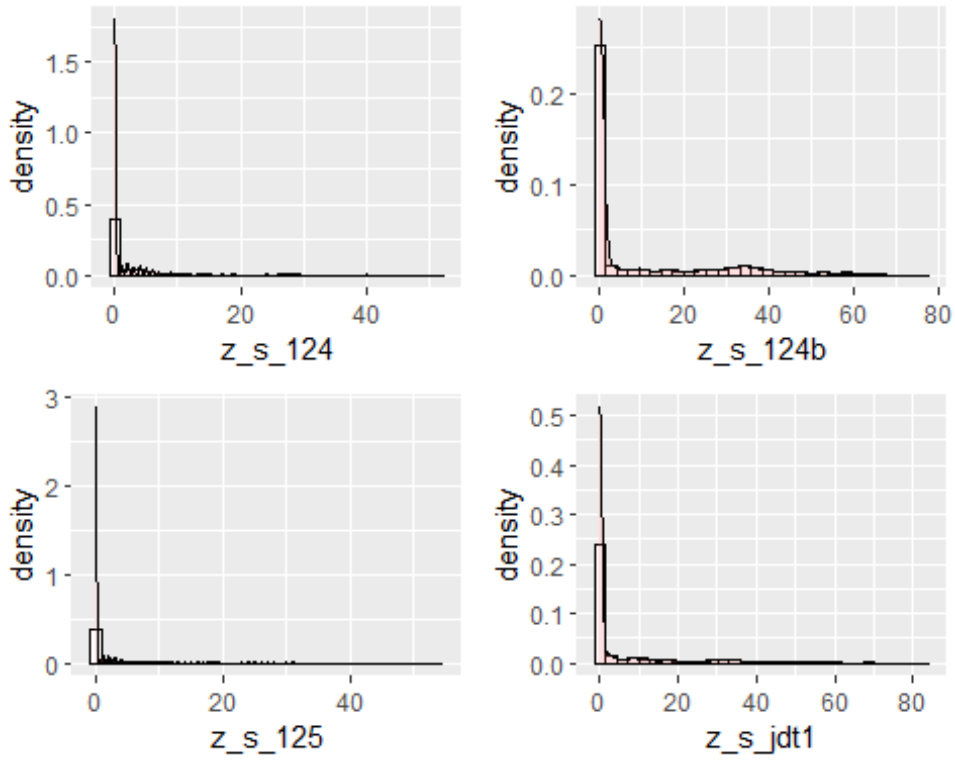
```
Snow_depth <- read.table('rc.tg_.dc_.jd_sc.txt', header = TRUE, sep = ",")
Snow_depth <- subset(Snow_depth, select = -c(Date_time))
head(Snow_depth)

##     WY Year Month Day Hour Minute z_s_124 z_s_124b z_s_125 z_s_jdt1
z_s_jdt2
## 1 2004 2003    10   1    0      0       0    -9999       0    -9999    -
9999
## 2 2004 2003    10   1    1      0       0    -9999       0    -9999    -
9999
## 3 2004 2003    10   1    2      0       0    -9999       0    -9999    -
9999
## 4 2004 2003    10   1    3      0       0    -9999       0    -9999    -
9999
## 5 2004 2003    10   1    4      0       0    -9999       0    -9999    -
9999
## 6 2004 2003    10   1    5      0       0    -9999       0    -9999    -
9999
##   z_s_jdt3 z_s_jdt4 z_s_jdt5 z_s_jdt2b z_s_jdt3b z_s_jdt4b
## 1    -9999    -9999    -9999     -9999     -9999     -9999
## 2    -9999    -9999    -9999     -9999     -9999     -9999
## 3    -9999    -9999    -9999     -9999     -9999     -9999
```

```
## 4    -9999    -9999    -9999     -9999     -9999     -9999
## 5    -9999    -9999    -9999     -9999     -9999     -9999
## 6    -9999    -9999    -9999     -9999     -9999     -9999

summary(Snow_depth)

##        WY            Year          Month             Day              Hour
##  Min.   :2004   Min.   :2003   Min.   : 1.000   Min.   : 1.00   Min.   :
0.00
##  1st Qu.:2006   1st Qu.:2006   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.75
##  Median :2009   Median :2009   Median : 7.000   Median :16.00   Median
:11.50
##  Mean   :2009   Mean   :2009   Mean   : 6.523   Mean   :15.73   Mean
:11.50
##  3rd Qu.:2012   3rd Qu.:2011   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.25
##  Max.   :2014   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.00
##      Minute     z_s_124           z_s_124b          z_s_125
##  Min.   :0   Min.   :-9999.00   Min.   :-9999.0   Min.   :-9999.000
##  1st Qu.:0   1st Qu.:    0.00   1st Qu.:-9999.0   1st Qu.:    0.000
##  Median :0   Median :    0.00   Median :    0.0   Median :    0.000
##  Mean   :0   Mean   : -468.49   Mean   :-2827.2   Mean   :    0.686
##  3rd Qu.:0   3rd Qu.:    1.95   3rd Qu.:    0.0   3rd Qu.:    1.230
##  Max.   :0   Max.   :   51.62   Max.   :   76.7   Max.   :   54.000
##     z_s_jdt1          z_s_jdt2          z_s_jdt3           z_s_jdt4
##  Min.   :-9999.0   Min.   :-9999.0   Min.   :-9999.00   Min.   :-9999.00
##  1st Qu.:    0.0   1st Qu.:    0.0   1st Qu.:-9999.00   1st Qu.:    0.00
##  Median :    0.0   Median :    0.0   Median :    0.00   Median :    0.00
##  Mean   :-2081.7   Mean   :-1945.1   Mean   :-3635.06   Mean   :-1910.58
##  3rd Qu.:    1.7   3rd Qu.:    1.0   3rd Qu.:    0.00   3rd Qu.:   11.63
##  Max.   :   83.0   Max.   :   59.2   Max.   :   84.67   Max.   :   92.00
##     z_s_jdt5       z_s_jdt2b          z_s_jdt3b         z_s_jdt4b
##  Min.   :-9999   Min.   :-9999.00   Min.   :-9999.0   Min.   :-9999.00
##  1st Qu.:    0   1st Qu.:-9999.00   1st Qu.:-9999.0   1st Qu.:-9999.00
##  Median :    0   Median :-9999.00   Median :-9999.0   Median :-9999.00
##  Mean   :-1906   Mean   :-6573.95   Mean   :-6545.8   Mean   :-6557.07
##  3rd Qu.:    2   3rd Qu.:    0.00   3rd Qu.:    0.0   3rd Qu.:    0.00
##  Max.   :   56   Max.   :   25.43   Max.   :   25.9   Max.   :   31.07

#check if NA's Exist
list_na <- colnames(Snow_depth)[ apply(Snow_depth, 2, anyNA) ]
list_na

## character(0)

#check If missing values -9999 exist
any(Snow_depth==-9999)

## [1] TRUE
```

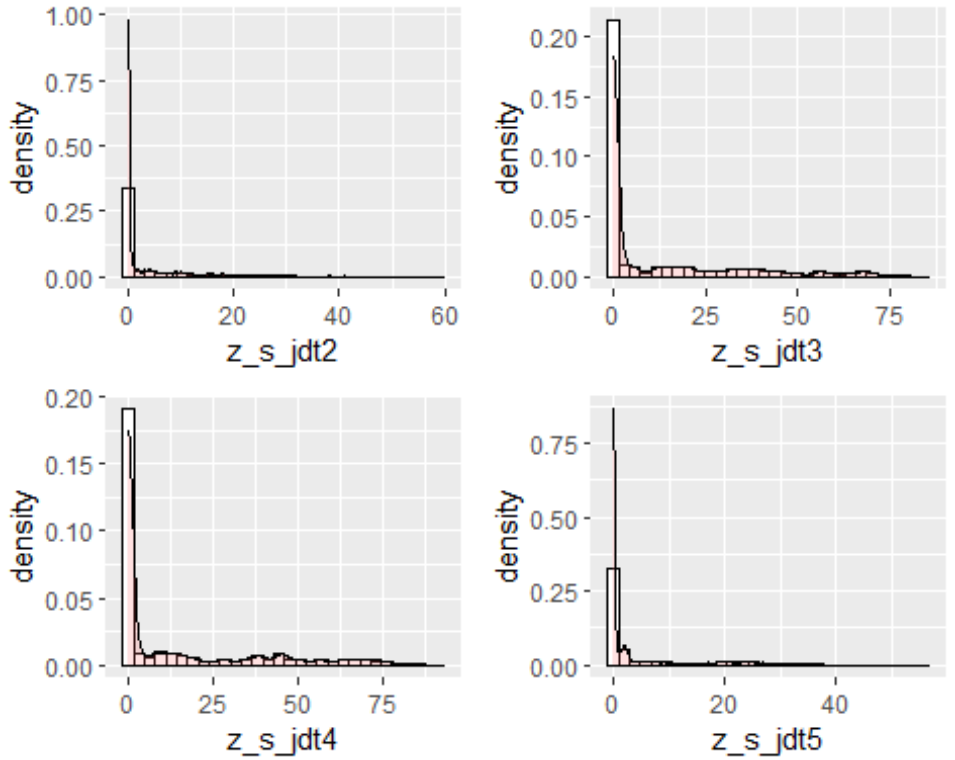
```r
# replace -9999 with Na's
Snow_depth <- na_if(Snow_depth, -9999)
#check if NA's Exist
list_na <- colnames(Snow_depth)[ apply(Snow_depth, 2, anyNA) ]
list_na

##  [1] "z_s_124"   "z_s_124b"  "z_s_125"   "z_s_jdt1"  "z_s_jdt2"
"z_s_jdt3"
##  [7] "z_s_jdt4"  "z_s_jdt5"  "z_s_jdt2b" "z_s_jdt3b" "z_s_jdt4b"

#Density plot to see distribution of data within the feature by Visulization
z_s_124 <- ggplot(Snow_depth, aes(x=z_s_124)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

z_s_124b <- ggplot(Snow_depth, aes(x=z_s_124b)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

z_s_125 <- ggplot(Snow_depth, aes(x=z_s_125)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

z_s_jdt1 <- ggplot(Snow_depth, aes(x=z_s_jdt1)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

z_s_jdt2 <- ggplot(Snow_depth, aes(x=z_s_jdt2)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

z_s_jdt3 <- ggplot(Snow_depth, aes(x=z_s_jdt3)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

z_s_jdt4 <- ggplot(Snow_depth, aes(x=z_s_jdt4)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

z_s_jdt5 <- ggplot(Snow_depth, aes(x=z_s_jdt5)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

z_s_jdt2b <- ggplot(Snow_depth, aes(x=z_s_jdt2b)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

z_s_jdt3b <- ggplot(Snow_depth, aes(x=z_s_jdt3b)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
```

```
 geom_density(alpha=.2, fill="#FF6666")

z_s_jdt4b <- ggplot(Snow_depth, aes(x=z_s_jdt4b)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

ggarrange(z_s_124,z_s_124b,z_s_125,z_s_jdt1,z_s_jdt2,z_s_jdt3,z_s_jdt4,z_s_jd
t5,z_s_jdt2b,z_s_jdt3b,z_s_jdt4b, ncol=2, nrow=2)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 4542 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 4542 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 27329 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 27329 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 20 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 20 rows containing non-finite values (`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 20133 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 20133 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 18790 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 18790 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 35130 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 35130 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 18533 rows containing non-finite values (`stat_bin()`).
```

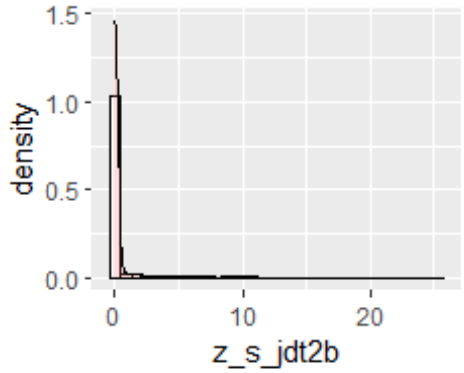
```
## Warning: Removed 18533 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 18417 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 18417 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 63402 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 63402 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 63133 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 63133 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 63242 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 63242 rows containing non-finite values
(`stat_density()`).

## $`1`
```

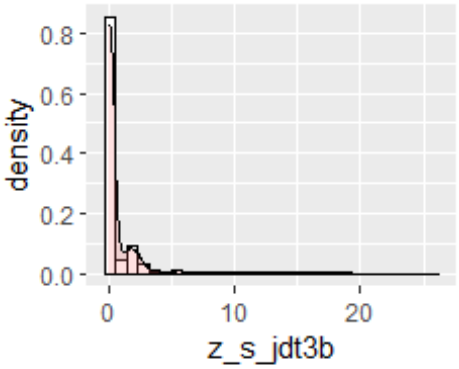

density
z_s_124
z_s_124b
z_s_125
z_s_jdt1

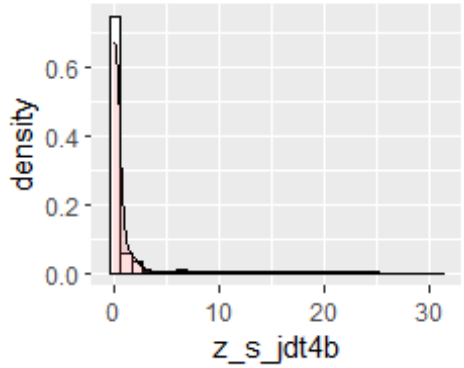
```
##
## $`2`
```


density
z_s_jdt2
z_s_jdt3
z_s_jdt4
z_s_jdt5

```
##
## $`3`
```








```
##
## attr(,"class")
## [1] "list"      "ggarrange"

summary(Snow_depth)

##        WY            Year          Month             Day              Hour
##  Min.   :2004   Min.   :2003   Min.   : 1.000   Min.   : 1.00   Min.   :
0.00
##  1st Qu.:2006   1st Qu.:2006   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.75
##  Median :2009   Median :2009   Median : 7.000   Median :16.00   Median
:11.50
##  Mean   :2009   Mean   :2009   Mean   : 6.523   Mean   :15.73   Mean
:11.50
##  3rd Qu.:2012   3rd Qu.:2011   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.25
##  Max.   :2014   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.00
##
##      Minute     z_s_124         z_s_124b         z_s_125          z_s_jdt1
##  Min.   :0   Min.   : 0.00   Min.   : 0.000   Min.   : 0.000   Min.   :
0.00
##  1st Qu.:0   1st Qu.: 0.00   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.:
```

```
0.00
##  Median :0   Median : 0.00   Median : 0.000   Median : 0.000   Median :
0.00
##  Mean   :0   Mean   : 2.59   Mean   : 9.072   Mean   : 2.761   Mean   :
7.45
##  3rd Qu.:0   3rd Qu.: 2.00   3rd Qu.:13.000   3rd Qu.: 1.235   3rd Qu.:
7.10
##  Max.   :0   Max.   :51.62   Max.   :76.700   Max.   :54.000   Max.
:83.00
##              NA's   :4542    NA's   :27329    NA's   :20       NA's
:20133
##     z_s_jdt2        z_s_jdt3        z_s_jdt4        z_s_jdt5
##  Min.   : 0.00   Min.   : 0.00   Min.   : 0.00   Min.   : 0.000
##  1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.: 0.000
##  Median : 0.00   Median : 0.00   Median : 0.00   Median : 0.000
##  Mean   : 4.04   Mean   :11.88   Mean   :13.74   Mean   : 4.851
##  3rd Qu.: 3.62   3rd Qu.:18.27   3rd Qu.:19.60   3rd Qu.: 4.000
##  Max.   :59.20   Max.   :84.67   Max.   :92.00   Max.   :56.000
##  NA's   :18790   NA's   :35130   NA's   :18533   NA's   :18417
##    z_s_jdt2b       z_s_jdt3b       z_s_jdt4b
##  Min.   : 0.00   Min.   : 0.00   Min.   : 0.00
##  1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.: 0.00
##  Median : 0.00   Median : 0.00   Median : 0.00
##  Mean   : 0.54   Mean   : 1.17   Mean   : 1.37
##  3rd Qu.: 0.00   3rd Qu.: 0.00   3rd Qu.: 0.00
##  Max.   :25.43   Max.   :25.90   Max.   :31.07
##  NA's   :63402   NA's   :63133   NA's   :63242
```

```
Snow_depth <- replace(Snow_depth, is.na(Snow_depth), 0)

#check if NA's Exist
list_na <- colnames(Snow_depth)[ apply(Snow_depth, 2, anyNA) ]
list_na

## character(0)

#Merge snow depth feature for all stations into one feature
Snow_depth$z_s <- rowMeans(Snow_depth[ ,
c('z_s_124','z_s_124b','z_s_125','z_s_jdt1','z_s_jdt2','z_s_jdt3','z_s_jdt4',
'z_s_jdt5','z_s_jdt2b','z_s_jdt3b','z_s_jdt4b')])

Snow_depth <- subset(Snow_depth, select = -
c(z_s_124,z_s_124b,z_s_125,z_s_jdt1,z_s_jdt2,z_s_jdt3,z_s_jdt4,z_s_jdt5,z_s_j
dt2b,z_s_jdt3b,z_s_jdt4b))

summary(Snow_depth)

##        WY            Year          Month             Day              Hour
##  Min.   :2004   Min.   :2003   Min.   : 1.000   Min.   : 1.00   Min.   :
0.00
```

```
##  1st Qu.:2006   1st Qu.:2006   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.75
##  Median :2009   Median :2009   Median : 7.000   Median :16.00   Median
:11.50
##  Mean   :2009   Mean   :2009   Mean   : 6.523   Mean   :15.73   Mean
:11.50
##  3rd Qu.:2012   3rd Qu.:2011   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.25
##  Max.   :2014   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.00
##      Minute       z_s
##  Min.   :0   Min.   : 0.000
##  1st Qu.:0   1st Qu.: 0.000
##  Median :0   Median : 0.000
##  Mean   :0   Mean   : 4.047
##  3rd Qu.:0   3rd Qu.: 4.364
##  Max.   :0   Max.   :42.091

head(Snow_depth)

##     WY Year Month Day Hour Minute z_s
## 1 2004 2003    10   1    0      0   0
## 2 2004 2003    10   1    1      0   0
## 3 2004 2003    10   1    2      0   0
## 4 2004 2003    10   1    3      0   0
## 5 2004 2003    10   1    4      0   0
## 6 2004 2003    10   1    5      0   0
```

#download the SnowDepth datasets which is merged into excel for further analysis.

```
library("writexl")
write_xlsx(Snow_depth,"SnowDepth_merged.xlsx")
```

#Loading 9 soil data files

```
#=====================================Reading-9-
dataframes=================================

T1_Soil <- read.table('rc.tg_.dc_.jd-124ba_stm.txt', header = TRUE, sep =
",")
T1_Soil <- subset(T1_Soil, select = -c(Date_time))
head(T1_Soil)

##     WY Year Month Day Hour Minute T_g_5 T_g_20 T_g_50 T_g_75 T_g_90 s_m_5
## 1 2011 2010    10   1    0      0 -9999  -9999  -9999  -9999  -9999 -9999
## 2 2011 2010    10   1    1      0 -9999  -9999  -9999  -9999  -9999 -9999
## 3 2011 2010    10   1    2      0 -9999  -9999  -9999  -9999  -9999 -9999
## 4 2011 2010    10   1    3      0 -9999  -9999  -9999  -9999  -9999 -9999
## 5 2011 2010    10   1    4      0 -9999  -9999  -9999  -9999  -9999 -9999
## 6 2011 2010    10   1    5      0 -9999  -9999  -9999  -9999  -9999 -9999
##   s_m_20 s_m_50 s_m_75 s_m_90
```

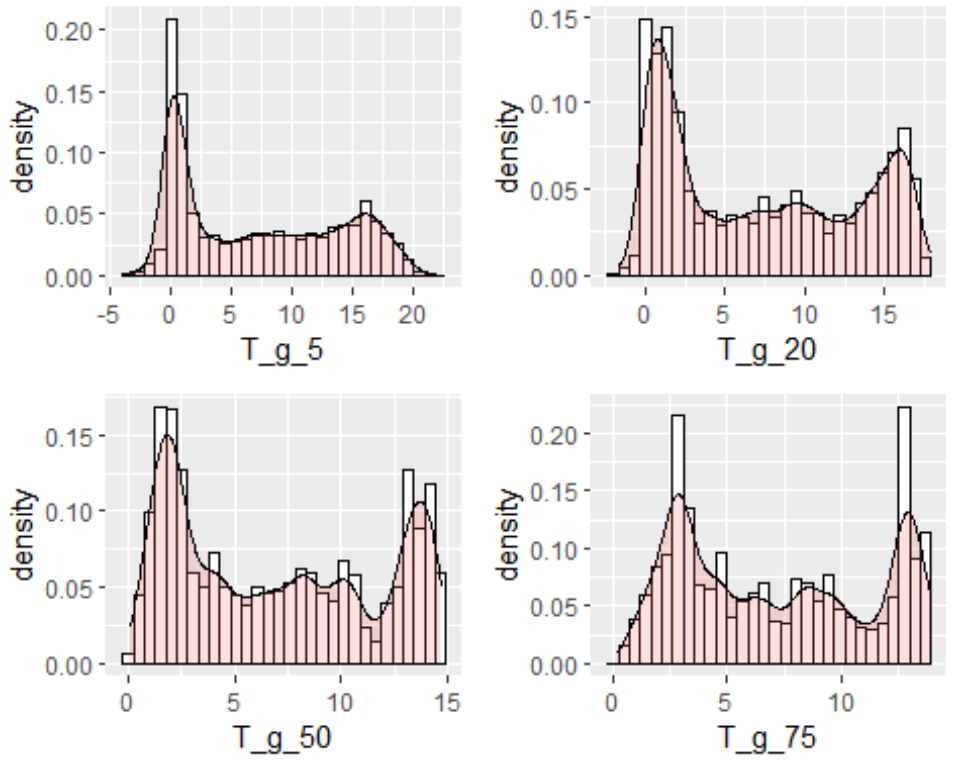
```
## 1  -9999  -9999  -9999  -9999
## 2  -9999  -9999  -9999  -9999
## 3  -9999  -9999  -9999  -9999
## 4  -9999  -9999  -9999  -9999
## 5  -9999  -9999  -9999  -9999
## 6  -9999  -9999  -9999  -9999

summary(T1_Soil)

##        WY            Year          Month             Day              Hour
##  Min.   :2011   Min.   :2010   Min.   : 1.000   Min.   : 1.00   Min.   :
0.00
##  1st Qu.:2012   1st Qu.:2011   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.75
##  Median :2012   Median :2012   Median : 7.000   Median :16.00   Median
:11.50
##  Mean   :2012   Mean   :2012   Mean   : 6.523   Mean   :15.73   Mean
:11.50
##  3rd Qu.:2013   3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.25
##  Max.   :2014   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.00
##      Minute      T_g_5             T_g_20            T_g_50
##  Min.   :0   Min.   :-9999.0   Min.   :-9999.0   Min.   :-9999.000
##  1st Qu.:0   1st Qu.:    0.4   1st Qu.:    1.1   1st Qu.:    1.991
##  Median :0   Median :    5.1   Median :    5.7   Median :    5.790
##  Mean   :0   Mean   : -423.8   Mean   : -423.4   Mean   : -588.864
##  3rd Qu.:0   3rd Qu.:   13.1   3rd Qu.:   12.8   3rd Qu.:   10.870
##  Max.   :0   Max.   :   22.1   Max.   :   17.9   Max.   :   14.740
##      T_g_75             T_g_90             s_m_5
##  Min.   :-9999.000   Min.   :-9999.000   Min.   :-9999.000
##  1st Qu.:-9999.000   1st Qu.:    3.535   1st Qu.:    0.062
##  Median :    2.956   Median :    6.746   Median :    0.145
##  Mean   :-3526.272   Mean   : -353.412   Mean   : -855.921
##  3rd Qu.:    8.680   3rd Qu.:   10.760   3rd Qu.:    0.218
##  Max.   :   13.910   Max.   :   13.320   Max.   :    0.420
##      s_m_20             s_m_50             s_m_75
##  Min.   :-9999.000   Min.   :-9999.000   Min.   :-9999.000
##  1st Qu.:    0.117   1st Qu.:    0.138   1st Qu.:-9999.000
##  Median :    0.269   Median :    0.238   Median :    0.244
##  Mean   : -855.845   Mean   : -995.861   Mean   :-3428.051
##  3rd Qu.:    0.320   3rd Qu.:    0.322   3rd Qu.:    0.298
##  Max.   :    0.393   Max.   :    0.387   Max.   :    0.410
##      s_m_90
##  Min.   :-9999.000
##  1st Qu.:    0.139
##  Median :    0.176
##  Mean   : -855.848
##  3rd Qu.:    0.330
##  Max.   :    0.401
```

```r
#check if NA's Exist
list_na <- colnames(T1_Soil)[ apply(T1_Soil, 2, anyNA) ]
list_na

## character(0)

#check If missing values -9999 exist
any(T1_Soil==-9999)

## [1] TRUE

# replace -9999 with Na's
T1_Soil <- na_if(T1_Soil, -9999)

#check if NA's Exist
list_na <- colnames(T1_Soil)[ apply(T1_Soil, 2, anyNA) ]
list_na

##  [1] "T_g_5"  "T_g_20" "T_g_50" "T_g_75" "T_g_90" "s_m_5"  "s_m_20"
"s_m_50"
##  [9] "s_m_75" "s_m_90"

#Density plot to see distribution of data within the feature by Visulization
T_g_5 <- ggplot(T1_Soil, aes(x=T_g_5)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

T_g_20 <- ggplot(T1_Soil, aes(x=T_g_20)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

T_g_50 <- ggplot(T1_Soil, aes(x=T_g_50)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

T_g_75 <- ggplot(T1_Soil, aes(x=T_g_75)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

T_g_90 <- ggplot(T1_Soil, aes(x=T_g_90)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

s_m_5 <- ggplot(T1_Soil, aes(x=s_m_5)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

s_m_20 <- ggplot(T1_Soil, aes(x=s_m_20)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")
```

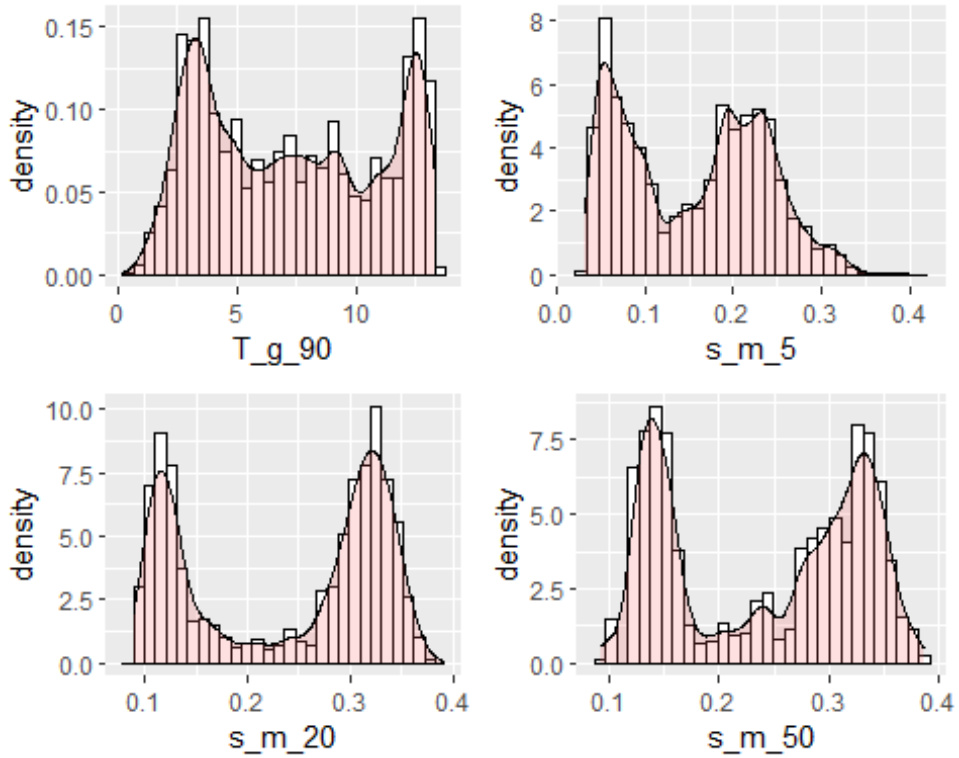
```
s_m_50 <- ggplot(T1_Soil, aes(x=s_m_50)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

s_m_75 <- ggplot(T1_Soil, aes(x=s_m_75)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

s_m_90 <- ggplot(T1_Soil, aes(x=s_m_90)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

ggarrange(T_g_5,T_g_20,T_g_50,T_g_75,T_g_90,s_m_5,s_m_20,s_m_50,s_m_75,s_m_90
, ncol=2, nrow=2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 1510 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 1510 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 1509 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 1509 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 2088 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 2088 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 12382 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 12382 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 1264 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 1264 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

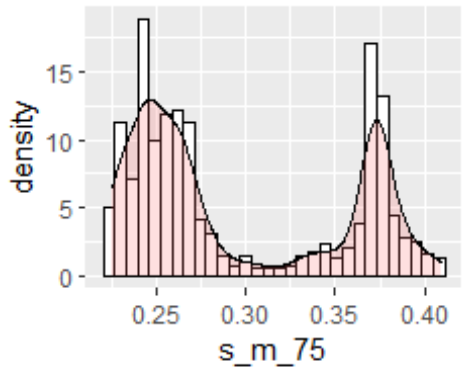
```
## Warning: Removed 3002 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 3002 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 3002 rows containing non-finite values (`stat_bin()`).
## Removed 3002 rows containing non-finite values (`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 3493 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 3493 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 12022 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 12022 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 3002 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 3002 rows containing non-finite values
(`stat_density()`).

## $`1`
```

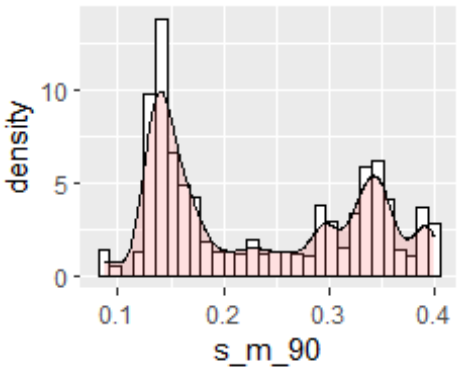



```
##
## $`2`
```



```
##
## $`3`
```






```
##
## attr(,"class")
## [1] "list"      "ggarrange"

summary(T1_Soil)

##        WY            Year          Month             Day              Hour
##  Min.   :2011   Min.   :2010   Min.   : 1.000   Min.   : 1.00   Min.   :
0.00
##  1st Qu.:2012   1st Qu.:2011   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.75
##  Median :2012   Median :2012   Median : 7.000   Median :16.00   Median
:11.50
##  Mean   :2012   Mean   :2012   Mean   : 6.523   Mean   :15.73   Mean
:11.50
##  3rd Qu.:2013   3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.25
##  Max.   :2014   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.00
##
##      Minute      T_g_5            T_g_20           T_g_50
##  Min.   :0   Min.   :-3.800   Min.   :-1.700   Min.   : 0.078
##  1st Qu.:0   1st Qu.: 0.600   1st Qu.: 1.300   1st Qu.: 2.296
##  Median :0   Median : 5.900   Median : 6.300   Median : 6.553
```

```
##  Mean   :0   Mean   : 7.076   Mean   : 7.193   Mean   : 6.974
##  3rd Qu.:0   3rd Qu.:13.400   3rd Qu.:13.300   3rd Qu.:11.470
##  Max.   :0   Max.   :22.100   Max.   :17.900   Max.   :14.740
##              NA's   :1510     NA's   :1509     NA's   :2088
##      T_g_75           T_g_90           s_m_5            s_m_20
##  Min.   : 0.203   Min.   : 0.245   Min.   :0.0310   Min.   :0.0890
##  1st Qu.: 3.133   1st Qu.: 3.715   1st Qu.:0.0740   1st Qu.:0.1280
##  Median : 6.649   Median : 7.090   Median :0.1670   Median :0.2860
##  Mean   : 7.161   Mean   : 7.299   Mean   :0.1554   Mean   :0.2381
##  3rd Qu.:11.470   3rd Qu.:10.930   3rd Qu.:0.2230   3rd Qu.:0.3220
##  Max.   :13.910   Max.   :13.320   Max.   :0.4200   Max.   :0.3930
##  NA's   :12382    NA's   :1264     NA's   :3002     NA's   :3002
##      s_m_50           s_m_75           s_m_90
##  Min.   :0.092   Min.   :0.225   Min.   :0.0870
##  1st Qu.:0.148   1st Qu.:0.245   1st Qu.:0.1430
##  Median :0.271   Median :0.265   Median :0.2070
##  Mean   :0.242   Mean   :0.296   Mean   :0.2351
##  3rd Qu.:0.325   3rd Qu.:0.370   3rd Qu.:0.3340
##  Max.   :0.387   Max.   :0.410   Max.   :0.4010
##  NA's   :3493    NA's   :12022   NA's   :3002
```

```
T2_Soil <- read.table('rc.tg_.dc_.jd-124bs_stm.txt', header = TRUE, sep =
",")
T2_Soil <- subset(T2_Soil, select = -c(Date_time))
head(T2_Soil)
```

```
##     WY Year Month Day Hour Minute T_g_5 T_g_20 T_g_35 T_g_50 s_m_5 s_m_20
## 1 2011 2010    10   1    0      0 -9999  -9999  -9999  -9999 -9999  -9999
## 2 2011 2010    10   1    1      0 -9999  -9999  -9999  -9999 -9999  -9999
## 3 2011 2010    10   1    2      0 -9999  -9999  -9999  -9999 -9999  -9999
## 4 2011 2010    10   1    3      0 -9999  -9999  -9999  -9999 -9999  -9999
## 5 2011 2010    10   1    4      0 -9999  -9999  -9999  -9999 -9999  -9999
## 6 2011 2010    10   1    5      0 -9999  -9999  -9999  -9999 -9999  -9999
##   s_m_35 s_m_50
## 1  -9999  -9999
## 2  -9999  -9999
## 3  -9999  -9999
## 4  -9999  -9999
## 5  -9999  -9999
## 6  -9999  -9999
```

```
summary(T2_Soil)
```

```
##        WY            Year          Month             Day              Hour
##  Min.   :2011   Min.   :2010   Min.   : 1.000   Min.   : 1.00   Min.   :
0.00
##  1st Qu.:2012   1st Qu.:2011   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.75
##  Median :2012   Median :2012   Median : 7.000   Median :16.00   Median
:11.50
##  Mean   :2012   Mean   :2012   Mean   : 6.523   Mean   :15.73   Mean
```

```
:11.50
##  3rd Qu.:2013   3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.25
##  Max.   :2014   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.00
##      Minute      T_g_5            T_g_20           T_g_35
##  Min.   :0   Min.   :-9999.0   Min.   :-9999.0   Min.   :-9999.0
##  1st Qu.:0   1st Qu.:-9999.0   1st Qu.:    1.3   1st Qu.:    1.7
##  Median :0   Median :    0.2   Median :    7.4   Median :    5.9
##  Mean   :0   Mean   :-3231.2   Mean   : -406.3   Mean   : -736.0
##  3rd Qu.:0   3rd Qu.:   12.0   3rd Qu.:   17.0   3rd Qu.:   14.7
##  Max.   :0   Max.   :   41.1   Max.   :   24.9   Max.   :   20.7
##      T_g_50           s_m_5              s_m_20             s_m_35
##  Min.   :-9999.0   Min.   :-9999.000   Min.   :-9999.000   Min.   :-
9999.000
##  1st Qu.:    2.7   1st Qu.:-9999.000   1st Qu.:    0.097   1st Qu.:
0.091
##  Median :    7.3   Median :    0.119   Median :    0.154   Median :
0.121
##  Mean   : -352.0   Mean   :-3758.040   Mean   : -910.938   Mean   :-
1239.743
##  3rd Qu.:   14.4   3rd Qu.:    0.291   3rd Qu.:    0.247   3rd Qu.:
0.240
##  Max.   :   19.1   Max.   :    0.460   Max.   :    0.332   Max.   :
0.310
##      s_m_50
##  Min.   :-9999.000
##  1st Qu.:    0.111
##  Median :    0.172
##  Mean   : -910.922
##  3rd Qu.:    0.265
##  Max.   :    0.482
```

```
#check if NA's Exist
list_na <- colnames(T2_Soil)[ apply(T2_Soil, 2, anyNA) ]
list_na
```

```
## character(0)
```

```
#check If missing values -9999 exist
any(T2_Soil==-9999)
```

```
## [1] TRUE
```

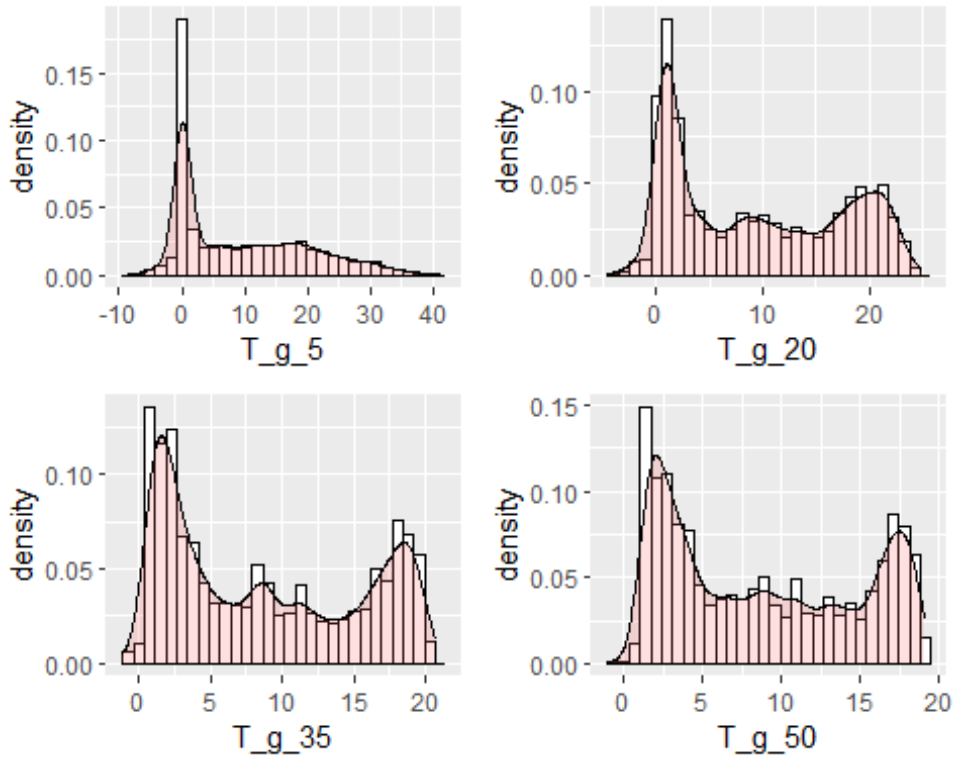
```
# replace -9999 with Na's
T2_Soil <- na_if(T2_Soil, -9999)

#check if NA's Exist
list_na <- colnames(T2_Soil)[ apply(T2_Soil, 2, anyNA) ]
list_na
```

```
## [1] "T_g_5"  "T_g_20" "T_g_35" "T_g_50" "s_m_5"  "s_m_20" "s_m_35"
"s_m_50"
```

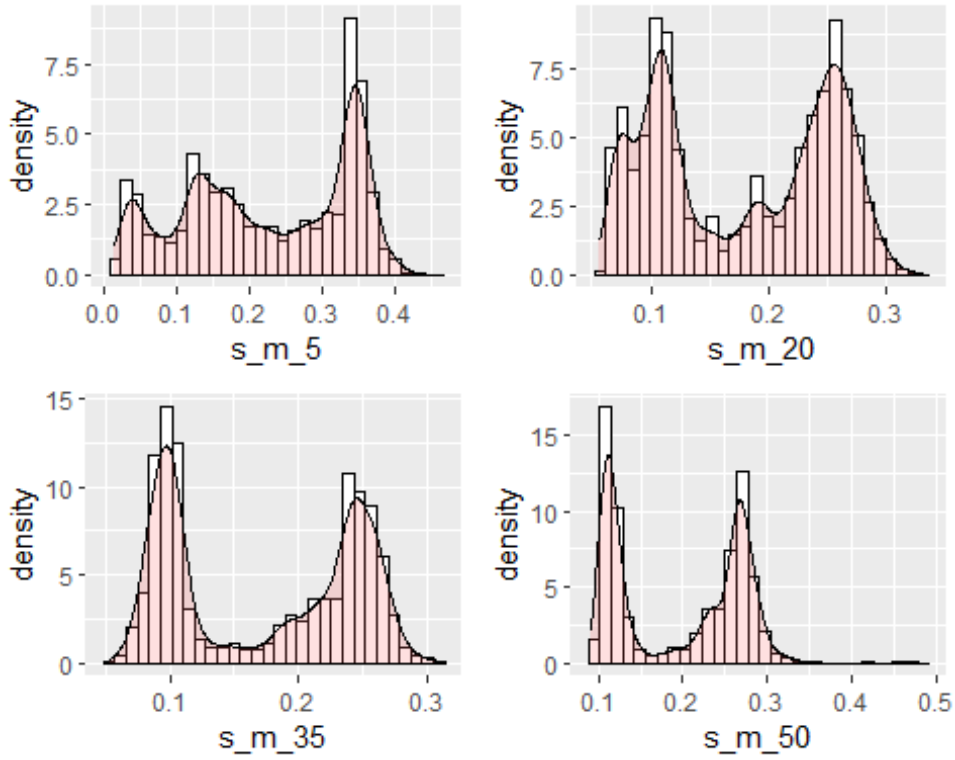
```
T_g_5 <- ggplot(T2_Soil, aes(x=T_g_5)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

T_g_20 <- ggplot(T2_Soil, aes(x=T_g_20)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

T_g_35 <- ggplot(T2_Soil, aes(x=T_g_35)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

T_g_50 <- ggplot(T2_Soil, aes(x=T_g_50)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

s_m_5 <- ggplot(T2_Soil, aes(x=s_m_5)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

s_m_20 <- ggplot(T2_Soil, aes(x=s_m_20)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

s_m_35 <- ggplot(T2_Soil, aes(x=s_m_35)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

s_m_50 <- ggplot(T2_Soil, aes(x=s_m_50)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

ggarrange(T_g_5,T_g_20,T_g_35,T_g_50,s_m_5,s_m_20,s_m_35,s_m_50, ncol=2,
nrow=2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 11353 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 11353 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 1456 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 1456 rows containing non-finite values
(`stat_density()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 2609 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 2609 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 1264 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 1264 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 13179 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 13179 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 3195 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 3195 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 4348 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 4348 rows containing non-finite values
(`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 3195 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 3195 rows containing non-finite values
(`stat_density()`).

## $`1`
```




```
##
## $`2`
```



```
##
## attr(,"class")
## [1] "list"      "ggarrange"
```

```
summary(T2_Soil)
```

```
##        WY            Year          Month             Day             Hour
##  Min.   :2011   Min.   :2010   Min.   : 1.000   Min.   : 1.00   Min.   :
0.00
##  1st Qu.:2012   1st Qu.:2011   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.75
##  Median :2012   Median :2012   Median : 7.000   Median :16.00   Median
:11.50
##  Mean   :2012   Mean   :2012   Mean   : 6.523   Mean   :15.73   Mean
:11.50
##  3rd Qu.:2013   3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.25
##  Max.   :2014   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.00
##
##      Minute      T_g_5            T_g_20           T_g_35
##  Min.   :0   Min.   :-8.600   Min.   :-4.400   Min.   :-1.100
##  1st Qu.:0   1st Qu.: 0.200   1st Qu.: 1.500   1st Qu.: 2.200
##  Median :0   Median : 5.900   Median : 8.100   Median : 7.300
##  Mean   :0   Mean   : 9.284   Mean   : 9.295   Mean   : 8.603
##  3rd Qu.:0   3rd Qu.:17.300   3rd Qu.:17.300   3rd Qu.:15.500
##  Max.   :0   Max.   :41.100   Max.   :24.900   Max.   :20.700
##              NA's   :11353    NA's   :1456     NA's   :2609
##      T_g_50           s_m_5           s_m_20          s_m_35
##  Min.   :-0.700   Min.   :0.014   Min.   :0.057   Min.   :0.054
##  1st Qu.: 3.000   1st Qu.:0.130   1st Qu.:0.107   1st Qu.:0.099
##  Median : 8.000   Median :0.227   Median :0.187   Median :0.187
##  Mean   : 8.793   Mean   :0.225   Mean   :0.178   Mean   :0.172
##  3rd Qu.:14.900   3rd Qu.:0.338   3rd Qu.:0.251   3rd Qu.:0.243
##  Max.   :19.100   Max.   :0.460   Max.   :0.332   Max.   :0.310
##  NA's   :1264     NA's   :13179   NA's   :3195    NA's   :4348
##      s_m_50
##  Min.   :0.090
##  1st Qu.:0.115
##  Median :0.221
##  Mean   :0.196
##  3rd Qu.:0.267
##  Max.   :0.482
##  NA's   :3195
```

```
T3_Soil <- read.table('rc.tg_.dc_.jd-jdt1_stm.txt', header = TRUE, sep = ",")
T3_Soil <- subset(T3_Soil, select = -c(Date_time))
head(T3_Soil)
```

```
##     WY Year Month Day Hour Minute T_g_5 T_g_20 T_g_50 T_g_90 T_g_130
T_g_190
```

```
## 1 2011 2010    10   1    0      0  12.7   14.7   13.4   13.1    12.4
12
## 2 2011 2010    10   1    1      0  12.4   14.6   13.5   13.1    12.4
12
## 3 2011 2010    10   1    2      0  11.9   14.4   13.5   13.1    12.6
12
## 4 2011 2010    10   1    3      0  11.2   14.3   13.5   13.1    12.4
12
## 5 2011 2010    10   1    4      0  11.1   14.1   13.4   13.1    12.4
12
## 6 2011 2010    10   1    5      0  10.7   14.0   13.5   13.1    12.4
12
##   s_m_5 s_m_20 s_m_50 s_m_90 s_m_130 s_m_190
## 1 0.092  0.041  0.066  0.112   0.108   0.148
## 2 0.090  0.039  0.064  0.110   0.108   0.140
## 3 0.088  0.039  0.064  0.108   0.115   0.148
## 4 0.090  0.039  0.064  0.110   0.108   0.148
## 5 0.092  0.045  0.070  0.112   0.114   0.140
## 6 0.092  0.039  0.064  0.112   0.114   0.145

summary(T3_Soil)

##        WY            Year          Month              Day              Hour
##  Min.   :2011   Min.   :2010   Min.   : 1.000   Min.   : 1.00   Min.   :
0.00
##  1st Qu.:2012   1st Qu.:2011   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.75
##  Median :2012   Median :2012   Median : 7.000   Median :16.00   Median
:11.50
##  Mean   :2012   Mean   :2012   Mean   : 6.523   Mean   :15.73   Mean
:11.50
##  3rd Qu.:2013   3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.25
##  Max.   :2014   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.00
##      Minute      T_g_5            T_g_20           T_g_50
##  Min.   :0   Min.   :-9999.0   Min.   :-9999.0   Min.   :-9999.0
##  1st Qu.:0   1st Qu.:    0.3   1st Qu.:    0.7   1st Qu.:    1.9
##  Median :0   Median :    6.5   Median :    6.9   Median :    7.0
##  Mean   :0   Mean   : -195.9   Mean   : -196.5   Mean   : -196.8
##  3rd Qu.:0   3rd Qu.:   17.5   3rd Qu.:   16.5   3rd Qu.:   14.7
##  Max.   :0   Max.   :   39.1   Max.   :   27.5   Max.   :   20.7
##      T_g_90           T_g_130           T_g_190           s_m_5
##  Min.   :-9999.0   Min.   :-9999.0   Min.   :-9999.0   Min.   :-9999.000
##  1st Qu.:    3.2   1st Qu.:    4.2   1st Qu.:    5.5   1st Qu.:    0.084
##  Median :    7.5   Median :    7.6   Median :    8.1   Median :    0.113
##  Mean   : -196.8   Mean   : -196.9   Mean   : -196.8   Mean   : -204.891
##  3rd Qu.:   13.3   3rd Qu.:   12.0   3rd Qu.:   11.1   3rd Qu.:    0.202
##  Max.   :   17.5   Max.   :   15.2   Max.   :   13.5   Max.   :    0.397
##      s_m_20              s_m_50              s_m_90
```

```
##  Min.   :-9999.000   Min.   :-9999.000   Min.   :-9999.000
##  1st Qu.:    0.044   1st Qu.:    0.068   1st Qu.:    0.108
##  Median :    0.088   Median :    0.098   Median :    0.112
##  Mean   : -204.920   Mean   : -204.923   Mean   : -204.899
##  3rd Qu.:    0.191   3rd Qu.:    0.160   3rd Qu.:    0.179
##  Max.   :    0.301   Max.   :    0.226   Max.   :    0.243
##     s_m_130            s_m_190
##  Min.   :-9999.000   Min.   :-9999.000
##  1st Qu.:    0.108   1st Qu.:    0.136
##  Median :    0.112   Median :    0.141
##  Mean   : -204.903   Mean   : -204.884
##  3rd Qu.:    0.124   3rd Qu.:    0.148
##  Max.   :    0.242   Max.   :    0.246
```

```
#check if NA's Exist
list_na <- colnames(T3_Soil)[ apply(T3_Soil, 2, anyNA) ]
list_na
```

```
## character(0)
```

```
#check If missing values -9999 exist
any(T3_Soil==-9999)
```

```
## [1] TRUE
```

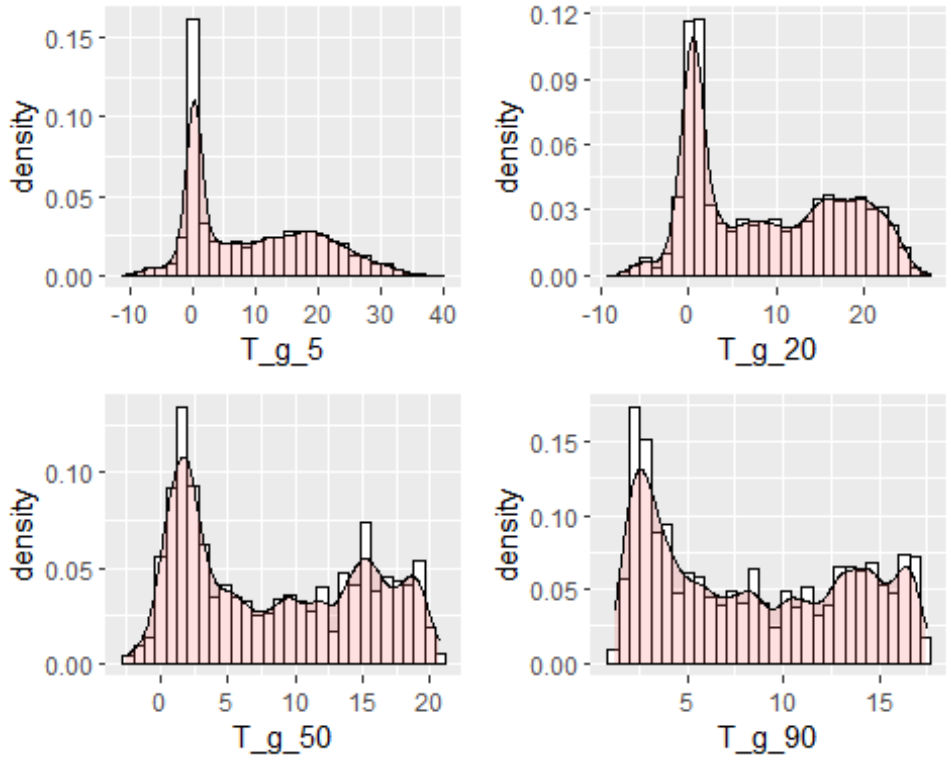
```
# replace -9999 with Na's
T3_Soil <- na_if(T3_Soil, -9999)

#check if NA's Exist
list_na <- colnames(T3_Soil)[ apply(T3_Soil, 2, anyNA) ]
list_na
```

```
##  [1] "T_g_5"   "T_g_20"  "T_g_50"  "T_g_90"  "T_g_130" "T_g_190" "s_m_5"
##  [8] "s_m_20"  "s_m_50"  "s_m_90"  "s_m_130" "s_m_190"
```

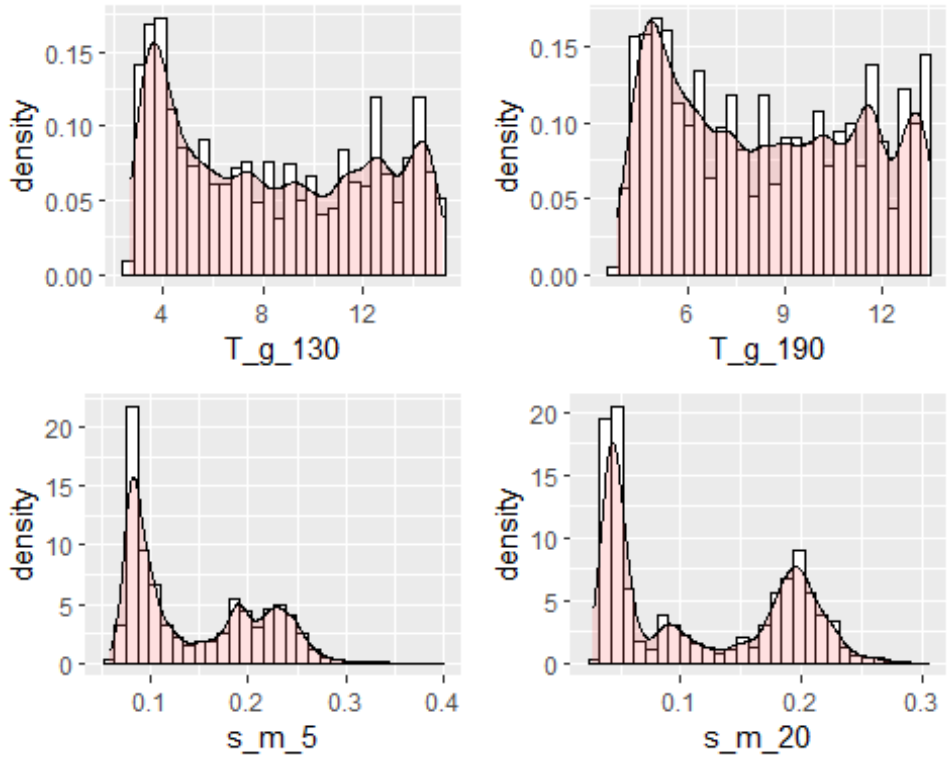
```
T_g_5 <- ggplot(T3_Soil, aes(x=T_g_5)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

T_g_20 <- ggplot(T3_Soil, aes(x=T_g_20)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

T_g_50 <- ggplot(T3_Soil, aes(x=T_g_50)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

T_g_90 <- ggplot(T3_Soil, aes(x=T_g_90)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")
```

```
T_g_130 <- ggplot(T3_Soil, aes(x=T_g_130)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

T_g_190 <- ggplot(T3_Soil, aes(x=T_g_190)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

s_m_5 <- ggplot(T3_Soil, aes(x=s_m_5)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

s_m_20 <- ggplot(T3_Soil, aes(x=s_m_20)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

s_m_50 <- ggplot(T3_Soil, aes(x=s_m_50)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

s_m_90 <- ggplot(T3_Soil, aes(x=s_m_90)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

s_m_130 <- ggplot(T3_Soil, aes(x=s_m_130)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

s_m_190 <- ggplot(T3_Soil, aes(x=s_m_190)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white")+
 geom_density(alpha=.2, fill="#FF6666")

ggarrange(T_g_5,T_g_20,T_g_50,T_g_90,T_g_130,T_g_190,s_m_5,s_m_20,s_m_50,s_m_
90,s_m_130,s_m_190, ncol=2, nrow=2)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 719 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 719 rows containing non-finite values (`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 719 rows containing non-finite values (`stat_bin()`).
## Removed 719 rows containing non-finite values (`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 719 rows containing non-finite values (`stat_bin()`).
## Removed 719 rows containing non-finite values (`stat_density()`).
```

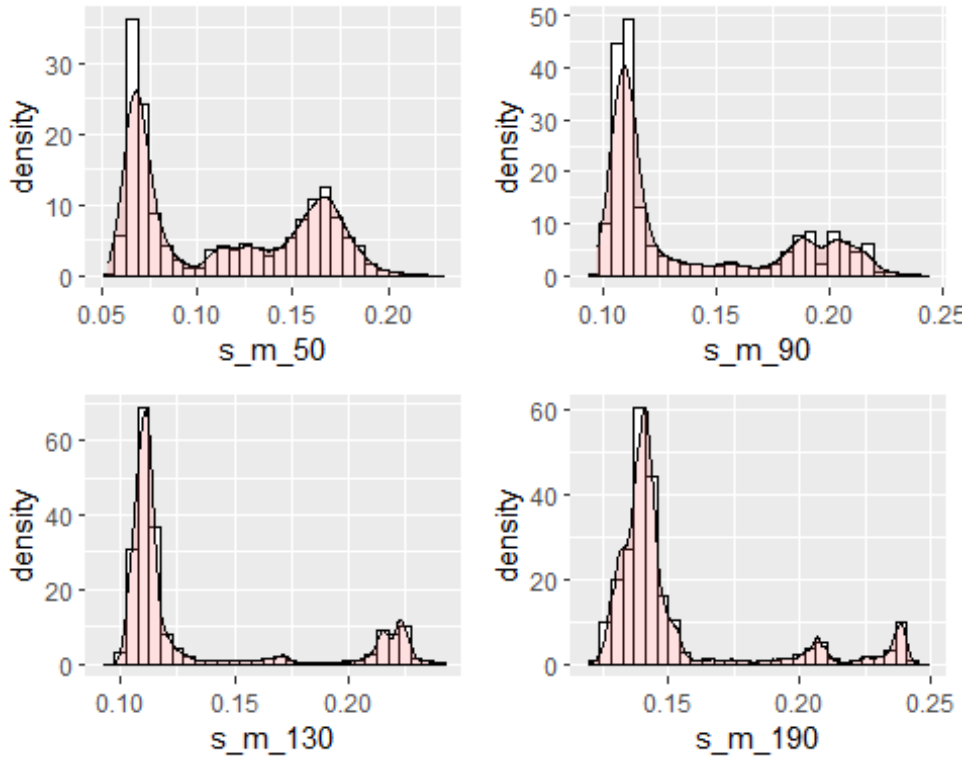
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 719 rows containing non-finite values (`stat_bin()`).
## Removed 719 rows containing non-finite values (`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 719 rows containing non-finite values (`stat_bin()`).
## Removed 719 rows containing non-finite values (`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 719 rows containing non-finite values (`stat_bin()`).
## Removed 719 rows containing non-finite values (`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 719 rows containing non-finite values (`stat_bin()`).
## Removed 719 rows containing non-finite values (`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 719 rows containing non-finite values (`stat_bin()`).
## Removed 719 rows containing non-finite values (`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 719 rows containing non-finite values (`stat_bin()`).
## Removed 719 rows containing non-finite values (`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 719 rows containing non-finite values (`stat_bin()`).
## Removed 719 rows containing non-finite values (`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 719 rows containing non-finite values (`stat_bin()`).
## Removed 719 rows containing non-finite values (`stat_density()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 719 rows containing non-finite values (`stat_bin()`).
## Removed 719 rows containing non-finite values (`stat_density()`).

## $`1`
```




```
##
## $`2`
```



```
##
## $`3`
```




```
##
## attr(,"class")
## [1] "list"      "ggarrange"

summary(T3_Soil)

##        WY            Year          Month             Day              Hour
##  Min.   :2011   Min.   :2010   Min.   : 1.000   Min.   : 1.00   Min.   :
0.00
##  1st Qu.:2012   1st Qu.:2011   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.75
##  Median :2012   Median :2012   Median : 7.000   Median :16.00   Median
:11.50
##  Mean   :2012   Mean   :2012   Mean   : 6.523   Mean   :15.73   Mean
:11.50
##  3rd Qu.:2013   3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.25
##  Max.   :2014   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.00
##
##      Minute      T_g_5            T_g_20          T_g_50
##  Min.   :0   Min.   :-10.700   Min.   :-8.000   Min.   :-2.500
##  1st Qu.:0   1st Qu.:  0.400   1st Qu.: 0.800   1st Qu.: 2.100
##  Median :0   Median :  7.000   Median : 7.300   Median : 7.500
```

```
##  Mean   :0   Mean   :  9.353   Mean   : 8.762   Mean   : 8.451
##  3rd Qu.:0   3rd Qu.: 17.900   3rd Qu.:16.700   3rd Qu.:14.900
##  Max.   :0   Max.   : 39.100   Max.   :27.500   Max.   :20.700
##              NA's   :719       NA's   :719      NA's   :719
##      T_g_90          T_g_130          T_g_190          s_m_5
##  Min.   : 1.200   Min.   : 2.700   Min.   : 3.80   Min.   :0.0600
##  1st Qu.: 3.400   1st Qu.: 4.400   1st Qu.: 5.60   1st Qu.:0.0840
##  Median : 7.800   Median : 7.800   Median : 8.30   Median :0.1170
##  Mean   : 8.431   Mean   : 8.312   Mean   : 8.44   Mean   :0.1452
##  3rd Qu.:13.300   3rd Qu.:12.000   3rd Qu.:11.10   3rd Qu.:0.2040
##  Max.   :17.500   Max.   :15.200   Max.   :13.50   Max.   :0.3970
##  NA's   :719      NA's   :719      NA's   :719     NA's   :719
##      s_m_20           s_m_50           s_m_90          s_m_130
##  Min.   :0.0280   Min.   :0.0530   Min.   :0.0970   Min.   :0.0970
##  1st Qu.:0.0440   1st Qu.:0.0690   1st Qu.:0.1080   1st Qu.:0.1090
##  Median :0.0920   Median :0.1070   Median :0.1130   Median :0.1120
##  Mean   :0.1156   Mean   :0.1127   Mean   :0.1373   Mean   :0.1328
##  3rd Qu.:0.1910   3rd Qu.:0.1600   3rd Qu.:0.1800   3rd Qu.:0.1260
##  Max.   :0.3010   Max.   :0.2260   Max.   :0.2430   Max.   :0.2420
##  NA's   :719      NA's   :719      NA's   :719      NA's   :719
##     s_m_190
##  Min.   :0.1200
##  1st Qu.:0.1370
##  Median :0.1410
##  Mean   :0.1522
##  3rd Qu.:0.1480
##  Max.   :0.2460
##  NA's   :719
```

```
T4_Soil <- read.table('rc.tg_.dc_.jd-jdt2_stm.txt', header = TRUE, sep = ",")
T4_Soil <- subset(T4_Soil, select = -c(Date_time))
head(T4_Soil)
```

```
##     WY Year Month Day Hour Minute T_g_5 T_g_20 T_g_50 T_g_75 T_g_100
s_m_5
## 1 2011 2010    10   1    0      0  11.5   12.6   12.7   12.8    12.4
0.03625
## 2 2011 2010    10   1    1      0  11.3   12.7   12.7   12.7    12.4
0.03725
## 3 2011 2010    10   1    2      0  11.2   12.6   12.7   12.8    12.4
0.03400
## 4 2011 2010    10   1    3      0  11.0   12.3   12.7   12.7    12.4
0.03600
## 5 2011 2010    10   1    4      0  10.7   12.4   12.8   12.8    12.4
0.03325
## 6 2011 2010    10   1    5      0  10.6   12.3   12.8   12.8    12.4
0.03650
##   s_m_20 s_m_50 s_m_75     s_m_100
## 1  0.041  0.036  0.021  0.00418024
## 2  0.047  0.042  0.027  0.00297032
```

```
## 3  0.039  0.034  0.019  0.00236536
## 4  0.041  0.036  0.021  0.00176040
## 5  0.039  0.034  0.019 -0.00065944
## 6  0.039  0.034  0.019 -0.00005448
```

```
summary(T4_Soil)
```

```
##        WY            Year          Month             Day              Hour
##  Min.   :2011   Min.   :2010   Min.   : 1.000   Min.   : 1.00   Min.   :
0.00
##  1st Qu.:2012   1st Qu.:2011   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.75
##  Median :2012   Median :2012   Median : 7.000   Median :16.00   Median
:11.50
##  Mean   :2012   Mean   :2012   Mean   : 6.523   Mean   :15.73   Mean
:11.50
##  3rd Qu.:2013   3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.25
##  Max.   :2014   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.00
##      Minute      T_g_5             T_g_20              T_g_50
##  Min.   :0   Min.   :-9999.0   Min.   :-9999.000   Min.   :-9999.000
##  1st Qu.:0   1st Qu.:   -0.3   1st Qu.:    0.100   1st Qu.:    0.900
##  Median :0   Median :    3.5   Median :    5.100   Median :    5.800
##  Mean   :0   Mean   : -201.3   Mean   :    6.949   Mean   :    6.653
##  3rd Qu.:0   3rd Qu.:   14.1   3rd Qu.:   14.600   3rd Qu.:   14.300
##  Max.   :0   Max.   :   32.7   Max.   :   23.300   Max.   :   20.300
##      T_g_75            T_g_100              s_m_5             s_m_20
##  Min.   :-9999.00   Min.   :-9999.000   Min.   :0.03000   Min.   :0.0370
##  1st Qu.:    1.60   1st Qu.:    2.000   1st Qu.:0.05125   1st Qu.:0.0590
##  Median :    6.40   Median :    6.500   Median :0.10300   Median :0.1160
##  Mean   :    7.34   Mean   :    7.213   Mean   :0.10904   Mean   :0.1193
##  3rd Qu.:   13.70   3rd Qu.:   12.800   3rd Qu.:0.15550   3rd Qu.:0.1820
##  Max.   :   18.80   Max.   :   17.200   Max.   :0.45350   Max.   :0.2870
##      s_m_50            s_m_75            s_m_100
##  Min.   :0.03100   Min.   :0.01500   Min.   :-0.002172
##  1st Qu.:0.05400   1st Qu.:0.03800   1st Qu.: 0.019446
##  Median :0.09700   Median :0.05400   Median : 0.027916
##  Mean   :0.09853   Mean   :0.07662   Mean   : 0.052734
##  3rd Qu.:0.14000   3rd Qu.:0.11700   3rd Qu.: 0.100000
##  Max.   :0.21600   Max.   :0.24800   Max.   : 0.187400
```

```
#check if NA's Exist
list_na <- colnames(T4_Soil)[ apply(T4_Soil, 2, anyNA) ]
list_na
```

```
## character(0)
```

```
#check If missing values -9999 exist
any(T4_Soil==-9999)
```

```
## [1] TRUE

# replace -9999 with Na's
T4_Soil <- na_if(T4_Soil, -9999)

#check if NA's Exist
list_na <- colnames(T4_Soil)[ apply(T4_Soil, 2, anyNA) ]
list_na

## [1] "T_g_5"   "T_g_20"  "T_g_50"  "T_g_75"  "T_g_100"

T5_Soil <- read.table('rc.tg_.dc_.jd-jdt2b_stm.txt', header = TRUE, sep =
",")
T5_Soil <- subset(T5_Soil, select = -c(Date_time))
head(T5_Soil)

##     WY Year Month Day Hour Minute T_g_5 T_g_20 T_g_35 T_g_50 T_g_75 s_m_5
## 1 2011 2010    10   1    0      0 -9999  -9999  -9999  -9999  -9999 -9999
## 2 2011 2010    10   1    1      0 -9999  -9999  -9999  -9999  -9999 -9999
## 3 2011 2010    10   1    2      0 -9999  -9999  -9999  -9999  -9999 -9999
## 4 2011 2010    10   1    3      0 -9999  -9999  -9999  -9999  -9999 -9999
## 5 2011 2010    10   1    4      0 -9999  -9999  -9999  -9999  -9999 -9999
## 6 2011 2010    10   1    5      0 -9999  -9999  -9999  -9999  -9999 -9999
##   s_m_20 s_m_35 s_m_50 s_m_75
## 1  -9999  -9999  -9999  -9999
## 2  -9999  -9999  -9999  -9999
## 3  -9999  -9999  -9999  -9999
## 4  -9999  -9999  -9999  -9999
## 5  -9999  -9999  -9999  -9999
## 6  -9999  -9999  -9999  -9999

summary(T5_Soil)

##        WY            Year          Month             Day              Hour
##  Min.   :2011   Min.   :2010   Min.   : 1.000   Min.   : 1.00   Min.   :
0.00
##  1st Qu.:2012   1st Qu.:2011   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.75
##  Median :2012   Median :2012   Median : 7.000   Median :16.00   Median
:11.50
##  Mean   :2012   Mean   :2012   Mean   : 6.523   Mean   :15.73   Mean
:11.50
##  3rd Qu.:2013   3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.25
##  Max.   :2014   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.00
##      Minute      T_g_5            T_g_20           T_g_35
##  Min.   :0   Min.   :-9999.0   Min.   :-9999.0   Min.   :-9999.0
##  1st Qu.:0   1st Qu.:    1.6   1st Qu.:    2.9   1st Qu.:    3.1
##  Median :0   Median :    9.5   Median :   10.2   Median :   10.2
##  Mean   :0   Mean   : -618.0   Mean   : -618.1   Mean   : -618.2
```

```
##  3rd Qu.:0   3rd Qu.:   20.3   3rd Qu.:   20.5   3rd Qu.:   19.9
##  Max.   :0   Max.   :   39.1   Max.   :   29.2   Max.   :   26.9
##      T_g_50            T_g_75            s_m_5              s_m_20
##  Min.   :-9999.00   Min.   :-9999.0   Min.   :-9999.000   Min.   :-
9999.000
##  1st Qu.:    3.40   1st Qu.:    4.0   1st Qu.:    0.064   1st Qu.:
0.159
##  Median :   10.15   Median :    9.9   Median :    0.113   Median :
0.180
##  Mean   : -618.45   Mean   : -618.7   Mean   : -629.516   Mean   : -
629.428
##  3rd Qu.:   18.90   3rd Qu.:   17.7   3rd Qu.:    0.195   3rd Qu.:
0.303
##  Max.   :   25.10   Max.   :   23.5   Max.   :    0.321   Max.   :
0.384
##      s_m_35              s_m_50              s_m_75
##  Min.   :-9999.000   Min.   :-9999.000   Min.   :-9999.000
##  1st Qu.:    0.148   1st Qu.:    0.156   1st Qu.:    0.139
##  Median :    0.167   Median :    0.175   Median :    0.158
##  Mean   : -629.442   Mean   : -629.436   Mean   : -629.447
##  3rd Qu.:    0.293   3rd Qu.:    0.295   3rd Qu.:    0.294
##  Max.   :    0.350   Max.   :    0.352   Max.   :    0.336
```

```
#check if NA's Exist
list_na <- colnames(T5_Soil)[ apply(T5_Soil, 2, anyNA) ]
list_na
```

```
## character(0)
```

```
#check If missing values -9999 exist
any(T5_Soil==-9999)
```

```
## [1] TRUE
```

```
# replace -9999 with Na's
T5_Soil <- na_if(T5_Soil, -9999)

#check if NA's Exist
list_na <- colnames(T5_Soil)[ apply(T5_Soil, 2, anyNA) ]
list_na
```

```
##  [1] "T_g_5"  "T_g_20" "T_g_35" "T_g_50" "T_g_75" "s_m_5"  "s_m_20"
"s_m_35"
##  [9] "s_m_50" "s_m_75"
```

```
T6_Soil <- read.table('rc.tg_.dc_.jd-jdt3_stm.txt', header = TRUE, sep = ",")
T6_Soil <- subset(T6_Soil, select = -c(Date_time))
head(T6_Soil)
```

```
##     WY Year Month Day Hour Minute T_g_5 T_g_20 T_g_50 T_g_75 T_g_100 s_m_5
## 1 2011 2010    10   1    0      0  11.3   12.4   11.7   11.2    11.5 0.043
## 2 2011 2010    10   1    1      0  11.1   12.3   11.7   11.2    11.3 0.039
```

```
## 3 2011 2010    10   1    2      0  10.7   12.1   11.7   11.2    11.3 0.043
## 4 2011 2010    10   1    3      0  10.4   12.0   11.7   11.2    11.3 0.035
## 5 2011 2010    10   1    4      0  10.2   11.9   11.7   11.2    11.5 0.043
## 6 2011 2010    10   1    5      0   9.8   11.9   11.7   11.2    11.3 0.035
##   s_m_20 s_m_50 s_m_75 s_m_100
## 1  0.063  0.083  0.073   0.078
## 2  0.059  0.079  0.069   0.074
## 3  0.063  0.083  0.073   0.078
## 4  0.055  0.075  0.065   0.070
## 5  0.063  0.083  0.073   0.078
## 6  0.055  0.075  0.065   0.070

summary(T6_Soil)

##        WY            Year          Month              Day              Hour
##  Min.   :2011   Min.   :2010   Min.   : 1.000   Min.   : 1.00   Min.   :
0.00
##  1st Qu.:2012   1st Qu.:2011   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.75
##  Median :2012   Median :2012   Median : 7.000   Median :16.00   Median
:11.50
##  Mean   :2012   Mean   :2012   Mean   : 6.523   Mean   :15.73   Mean
:11.50
##  3rd Qu.:2013   3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.25
##  Max.   :2014   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.00
##      Minute      T_g_5             T_g_20          T_g_50
##  Min.   :0   Min.   :-11.700   Min.   :-9.70   Min.   :-4.100
##  1st Qu.:0   1st Qu.: -0.100   1st Qu.: 0.20   1st Qu.: 1.300
##  Median :0   Median :  4.300   Median : 5.00   Median : 5.800
##  Mean   :0   Mean   :  7.865   Mean   : 7.38   Mean   : 7.362
##  3rd Qu.:0   3rd Qu.: 15.400   3rd Qu.:14.90   3rd Qu.:13.500
##  Max.   :0   Max.   : 40.100   Max.   :24.40   Max.   :18.800
##      T_g_75          T_g_100           s_m_5            s_m_20
##  Min.   :-1.500   Min.   : 0.300   Min.   :0.0320   Min.   :0.032
##  1st Qu.: 1.800   1st Qu.: 2.600   1st Qu.:0.0690   1st Qu.:0.082
##  Median : 6.200   Median : 6.400   Median :0.1270   Median :0.124
##  Mean   : 7.058   Mean   : 7.216   Mean   :0.1382   Mean   :0.137
##  3rd Qu.:12.400   3rd Qu.:11.700   3rd Qu.:0.1960   3rd Qu.:0.193
##  Max.   :16.500   Max.   :15.400   Max.   :0.4690   Max.   :0.350
##      s_m_50           s_m_75          s_m_100
##  Min.   :0.0520   Min.   :0.0420   Min.   :0.0670
##  1st Qu.:0.0970   1st Qu.:0.0850   1st Qu.:0.0830
##  Median :0.1310   Median :0.1090   Median :0.0890
##  Mean   :0.1518   Mean   :0.1339   Mean   :0.1163
##  3rd Qu.:0.2090   3rd Qu.:0.1940   3rd Qu.:0.1700
##  Max.   :0.3040   Max.   :0.2440   Max.   :0.2130
```

```
#check if NA's Exist
list_na <- colnames(T6_Soil)[ apply(T6_Soil, 2, anyNA) ]
list_na

## character(0)

#check If missing values -9999 exist
any(T6_Soil==-9999)

## [1] FALSE

# replace -9999 with Na's
T6_Soil <- na_if(T6_Soil, -9999)

#check if NA's Exist
list_na <- colnames(T6_Soil)[ apply(T6_Soil, 2, anyNA) ]
list_na

## character(0)

T7_Soil <- read.table('rc.tg_.dc_.jd-jdt3b_stm.txt', header = TRUE, sep =
",")
T7_Soil <- subset(T7_Soil, select = -c(Date_time))
head(T7_Soil)

##     WY Year Month Day Hour Minute T_g_5 T_g_20 T_g_35 T_g_50 s_m_5 s_m_20
## 1 2011 2010    10   1    0      0 -9999  -9999  -9999  -9999 -9999  -9999
## 2 2011 2010    10   1    1      0 -9999  -9999  -9999  -9999 -9999  -9999
## 3 2011 2010    10   1    2      0 -9999  -9999  -9999  -9999 -9999  -9999
## 4 2011 2010    10   1    3      0 -9999  -9999  -9999  -9999 -9999  -9999
## 5 2011 2010    10   1    4      0 -9999  -9999  -9999  -9999 -9999  -9999
## 6 2011 2010    10   1    5      0 -9999  -9999  -9999  -9999 -9999  -9999
##   s_m_35 s_m_50
## 1  -9999  -9999
## 2  -9999  -9999
## 3  -9999  -9999
## 4  -9999  -9999
## 5  -9999  -9999
## 6  -9999  -9999

summary(T7_Soil)

##        WY            Year          Month             Day              Hour
##  Min.   :2011   Min.   :2010   Min.   : 1.000   Min.   : 1.00   Min.   :
0.00
##  1st Qu.:2012   1st Qu.:2011   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.75
##  Median :2012   Median :2012   Median : 7.000   Median :16.00   Median
:11.50
##  Mean   :2012   Mean   :2012   Mean   : 6.523   Mean   :15.73   Mean
:11.50
##  3rd Qu.:2013   3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd
```

```
Qu.:17.25
##  Max.   :2014   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.00
##      Minute      T_g_5              T_g_20            T_g_35
##  Min.   :0   Min.   :-9999.00   Min.   :-9999.0   Min.   :-9999.0
##  1st Qu.:0   1st Qu.:    2.05   1st Qu.:    3.2   1st Qu.:    3.8
##  Median :0   Median :    9.65   Median :   10.4   Median :   10.4
##  Mean   :0   Mean   : -617.80   Mean   : -617.7   Mean   : -617.9
##  3rd Qu.:0   3rd Qu.:   20.40   3rd Qu.:   20.9   3rd Qu.:   19.9
##  Max.   :0   Max.   :   41.45   Max.   :   29.2   Max.   :   26.4
##      T_g_50            s_m_5               s_m_20             s_m_35
##  Min.   :-9999.0   Min.   :-9999.000   Min.   :-9999.000   Min.   :-
9999.000
##  1st Qu.:    3.9   1st Qu.:    0.052   1st Qu.:    0.108   1st Qu.:
0.106
##  Median :    9.9   Median :    0.112   Median :    0.134   Median :
0.120
##  Mean   : -618.5   Mean   : -629.532   Mean   : -629.498   Mean   : -
629.500
##  3rd Qu.:   18.6   3rd Qu.:    0.168   3rd Qu.:    0.198   3rd Qu.:
0.199
##  Max.   :   24.4   Max.   :    0.277   Max.   :    0.280   Max.   :
0.290
##      s_m_50
##  Min.   :-9999.000
##  1st Qu.:    0.086
##  Median :    0.102
##  Mean   : -629.523
##  3rd Qu.:    0.173
##  Max.   :    0.288
```

```
#check if NA's Exist
list_na <- colnames(T7_Soil)[ apply(T7_Soil, 2, anyNA) ]
list_na
```

```
## character(0)
```

```
#check If missing values -9999 exist
any(T7_Soil==-9999)
```

```
## [1] TRUE
```

```
# replace -9999 with Na's
T7_Soil <- na_if(T7_Soil, -9999)

#check if NA's Exist
list_na <- colnames(T7_Soil)[ apply(T7_Soil, 2, anyNA) ]
list_na
```

```
## [1] "T_g_5"  "T_g_20" "T_g_35" "T_g_50" "s_m_5"  "s_m_20" "s_m_35"
"s_m_50"
```

```
T8_Soil <- read.table('rc.tg_.dc_.jd-jdt4_stm.txt', header = TRUE, sep = ",")
T8_Soil <- subset(T8_Soil, select = -c(Date_time))
head(T8_Soil)

##     WY Year Month Day Hour Minute T_g_5 T_g_20 T_g_50 T_g_75 T_g_100
s_m_5
## 1 2011 2010    10   1    0      0  10.8   12.1   11.0   11.1    10.7
0.0305
## 2 2011 2010    10   1    1      0  10.6   12.0   11.2   11.0    10.8
0.0315
## 3 2011 2010    10   1    2      0  10.2   12.0   11.0   11.0    10.8
0.0290
## 4 2011 2010    10   1    3      0   9.9   12.0   11.0   11.1    10.8
0.0290
## 5 2011 2010    10   1    4      0   9.7   12.0   11.0   11.1    10.8
0.0315
## 6 2011 2010    10   1    5      0   9.5   12.0   11.0   11.0    10.8
0.0300
##   s_m_20 s_m_50 s_m_75 s_m_100
## 1  0.036  0.016  0.017   0.021
## 2  0.031  0.019  0.017   0.021
## 3  0.031  0.019  0.021   0.021
## 4  0.031  0.022  0.017   0.021
## 5  0.036  0.019  0.021   0.021
## 6  0.036  0.022  0.017   0.018

summary(T8_Soil)

##        WY            Year          Month             Day              Hour
##  Min.   :2011   Min.   :2010   Min.   : 1.000   Min.   : 1.00   Min.   :
0.00
##  1st Qu.:2012   1st Qu.:2011   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.75
##  Median :2012   Median :2012   Median : 7.000   Median :16.00   Median
:11.50
##  Mean   :2012   Mean   :2012   Mean   : 6.523   Mean   :15.73   Mean
:11.50
##  3rd Qu.:2013   3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.25
##  Max.   :2014   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.00
##
##      Minute      T_g_5             T_g_20           T_g_50
##  Min.   :0   Min.   :-10.900   Min.   :-7.400   Min.   :-3.400
##  1st Qu.:0   1st Qu.: -0.200   1st Qu.: 0.675   1st Qu.: 1.100
##  Median :0   Median :  2.400   Median : 4.000   Median : 4.400
##  Mean   :0   Mean   :  6.403   Mean   : 6.590   Mean   : 6.063
##  3rd Qu.:0   3rd Qu.: 12.800   3rd Qu.:13.000   3rd Qu.:11.300
##  Max.   :0   Max.   : 32.700   Max.   :21.900   Max.   :17.000
##              NA's   :2560      NA's   :2560     NA's   :2560
```

```
##      T_g_75          T_g_100           s_m_5           s_m_20
##  Min.   :-0.400   Min.   : 1.200   Min.   :0.0200   Min.   :0.0270
##  1st Qu.: 2.200   1st Qu.: 2.700   1st Qu.:0.0395   1st Qu.:0.0550
##  Median : 5.200   Median : 5.500   Median :0.0855   Median :0.1210
##  Mean   : 6.476   Mean   : 6.398   Mean   :0.0839   Mean   :0.1155
##  3rd Qu.:10.800   3rd Qu.: 9.900   3rd Qu.:0.1160   3rd Qu.:0.1650
##  Max.   :15.200   Max.   :13.800   Max.   :0.4530   Max.   :0.2780
##  NA's   :2560     NA's   :2560     NA's   :2560     NA's   :2560
##      s_m_50           s_m_75          s_m_100
##  Min.   :0.0110   Min.   :0.0170   Min.   :0.0120
##  1st Qu.:0.0290   1st Qu.:0.0320   1st Qu.:0.0260
##  Median :0.1240   Median :0.0985   Median :0.0680
##  Mean   :0.1023   Mean   :0.0889   Mean   :0.0672
##  3rd Qu.:0.1600   3rd Qu.:0.1430   3rd Qu.:0.1080
##  Max.   :0.2310   Max.   :0.2050   Max.   :0.1480
##  NA's   :2560     NA's   :2560     NA's   :2560
```

```
#check if NA's Exist
list_na <- colnames(T8_Soil)[ apply(T8_Soil, 2, anyNA) ]
list_na
```

```
##  [1] "T_g_5"   "T_g_20"  "T_g_50"  "T_g_75"  "T_g_100" "s_m_5"   "s_m_20"
##  [8] "s_m_50"  "s_m_75"  "s_m_100"
```

```
#check If missing values -9999 exist
any(T8_Soil==-9999)
```

```
## [1] NA
```

```
# replace -9999 with Na's
T8_Soil <- na_if(T8_Soil, -9999)

#check if NA's Exist
list_na <- colnames(T8_Soil)[ apply(T8_Soil, 2, anyNA) ]
list_na
```

```
##  [1] "T_g_5"   "T_g_20"  "T_g_50"  "T_g_75"  "T_g_100" "s_m_5"   "s_m_20"
##  [8] "s_m_50"  "s_m_75"  "s_m_100"
```

```
T9_Soil <- read.table('rc.tg_.dc_.jd-jdt4b_stm.txt', header = TRUE, sep =
",")
T9_Soil <- subset(T9_Soil, select = -c(Date_time))
colnames(T9_Soil)[c(9)] <- c("T_g_35")
head(T9_Soil)
```

```
##     WY Year Month Day Hour Minute T_g_5 T_g_20 T_g_35 T_g_50 s_m_5 s_m_20
## 1 2011 2010    10   1    0      0 -9999  -9999  -9999  -9999 -9999  -9999
## 2 2011 2010    10   1    1      0 -9999  -9999  -9999  -9999 -9999  -9999
## 3 2011 2010    10   1    2      0 -9999  -9999  -9999  -9999 -9999  -9999
## 4 2011 2010    10   1    3      0 -9999  -9999  -9999  -9999 -9999  -9999
## 5 2011 2010    10   1    4      0 -9999  -9999  -9999  -9999 -9999  -9999
## 6 2011 2010    10   1    5      0 -9999  -9999  -9999  -9999 -9999  -9999
```

```
##   s_m_35 s_m_50
## 1  -9999  -9999
## 2  -9999  -9999
## 3  -9999  -9999
## 4  -9999  -9999
## 5  -9999  -9999
## 6  -9999  -9999
```

```
summary(T9_Soil)
```

```
##        WY            Year          Month             Day              Hour
##  Min.   :2011   Min.   :2010   Min.   : 1.000   Min.   : 1.00   Min.   :
0.00
##  1st Qu.:2012   1st Qu.:2011   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.75
##  Median :2012   Median :2012   Median : 7.000   Median :16.00   Median
:11.50
##  Mean   :2012   Mean   :2012   Mean   : 6.523   Mean   :15.73   Mean
:11.50
##  3rd Qu.:2013   3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.25
##  Max.   :2014   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.00
##      Minute      T_g_5             T_g_20            T_g_35
##  Min.   :0   Min.   :-9999.0   Min.   :-9999.0   Min.   :-9999.0
##  1st Qu.:0   1st Qu.:    2.3   1st Qu.:    3.3   1st Qu.:    3.7
##  Median :0   Median :    9.8   Median :   10.1   Median :    9.9
##  Mean   :0   Mean   : -617.6   Mean   : -618.0   Mean   : -618.3
##  3rd Qu.:0   3rd Qu.:   20.5   3rd Qu.:   20.5   3rd Qu.:   19.3
##  Max.   :0   Max.   :   46.8   Max.   :   29.5   Max.   :   26.4
##      T_g_50            s_m_5               s_m_20              s_m_35
##  Min.   :-9999.0   Min.   :-9999.000   Min.   :-9999.000   Min.   :-
9999.000
##  1st Qu.:    3.9   1st Qu.:    0.038   1st Qu.:    0.091   1st Qu.:
0.116
##  Median :    9.7   Median :    0.097   Median :    0.127   Median :
0.152
##  Mean   : -618.6   Mean   : -629.547   Mean   : -629.506   Mean   : -
629.482
##  3rd Qu.:   18.2   3rd Qu.:    0.148   3rd Qu.:    0.189   3rd Qu.:
0.221
##  Max.   :   24.7   Max.   :    0.288   Max.   :    0.281   Max.   :
0.281
##      s_m_50
##  Min.   :-9999.000
##  1st Qu.:    0.084
##  Median :    0.101
##  Mean   : -629.527
##  3rd Qu.:    0.161
##  Max.   :    0.228
```

```r
#check if NA's Exist
list_na <- colnames(T9_Soil)[ apply(T9_Soil, 2, anyNA) ]
list_na
```

```
## character(0)
```

```r
#check If missing values -9999 exist
any(T9_Soil==-9999)
```

```
## [1] TRUE
```

```r
# replace -9999 with Na's
T9_Soil <- na_if(T9_Soil, -9999)

#check if NA's Exist
list_na <- colnames(T9_Soil)[ apply(T9_Soil, 2, anyNA) ]
list_na
```

```
## [1] "T_g_5"  "T_g_20" "T_g_35" "T_g_50" "s_m_5"  "s_m_20" "s_m_35"
"s_m_50"
```

```r
# ==================== Handling missing values ============================
#install.packages('tidyr')
#remove.packages('tidyr')
library(dplyr)
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.2.3
```

```r
#replace NA values in all numeric columns with their respective medians

T1_Soil <- T1_Soil %>% mutate(across(where(is.numeric), ~replace_na(.,
median(., na.rm=TRUE))))

T2_Soil <- T2_Soil %>% mutate(across(where(is.numeric), ~replace_na(.,
median(., na.rm=TRUE))))

T3_Soil <- T3_Soil %>% mutate(across(where(is.numeric), ~replace_na(.,
median(., na.rm=TRUE))))

T4_Soil <- T4_Soil %>% mutate(across(where(is.numeric), ~replace_na(.,
median(., na.rm=TRUE))))

T5_Soil <- T5_Soil %>% mutate(across(where(is.numeric), ~replace_na(.,
median(., na.rm=TRUE))))

T6_Soil <- T6_Soil %>% mutate(across(where(is.numeric), ~replace_na(.,
median(., na.rm=TRUE))))

T7_Soil <- T7_Soil %>% mutate(across(where(is.numeric), ~replace_na(.,
median(., na.rm=TRUE))))
```

```
T8_Soil <- T8_Soil %>% mutate(across(where(is.numeric), ~replace_na(.,
median(., na.rm=TRUE))))

T9_Soil <- T9_Soil %>% mutate(across(where(is.numeric), ~replace_na(.,
median(., na.rm=TRUE))))

# use identical(newT1, T1) to check if two diff data frames are same or not

#Create a subset dataset using the grouping by featuers
mergeData = T1_Soil
mergeData_Sub = subset(mergeData, select =
c("WY","Year","Month","Day","Hour","Minute"))
summary(mergeData_Sub)
```

```
##        WY            Year          Month             Day              Hour
##  Min.   :2011   Min.   :2010   Min.   : 1.000   Min.   : 1.00   Min.   :
0.00
##  1st Qu.:2012   1st Qu.:2011   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.75
##  Median :2012   Median :2012   Median : 7.000   Median :16.00   Median
:11.50
##  Mean   :2012   Mean   :2012   Mean   : 6.523   Mean   :15.73   Mean
:11.50
##  3rd Qu.:2013   3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.25
##  Max.   :2014   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.00
##      Minute
##  Min.   :0
##  1st Qu.:0
##  Median :0
##  Mean   :0
##  3rd Qu.:0
##  Max.   :0
```

We will be going to handle/merge each field manually for all the 9 sites

```
# =============== T_g_5, s_m_5 Available in all sites ===============

mergeData_Sub['T_g_5']  =  (T1_Soil['T_g_5'] + T2_Soil['T_g_5'] +
T3_Soil['T_g_5'] + T4_Soil['T_g_5'] + T5_Soil['T_g_5'] + T6_Soil['T_g_5'] +
T7_Soil['T_g_5'] + T8_Soil['T_g_5'] + T9_Soil['T_g_5'] ) / 9
mergeData_Sub['s_m_5']  =  (T1_Soil['s_m_5'] + T2_Soil['s_m_5'] +
T3_Soil['s_m_5'] + T4_Soil['s_m_5'] + T5_Soil['s_m_5'] + T6_Soil['s_m_5'] +
T7_Soil['s_m_5'] + T8_Soil['s_m_5'] + T9_Soil['s_m_5'] ) / 9

# =============== T_g_20, s_m_20 Available in all sites ===============

mergeData_Sub['T_g_20']  =  (T1_Soil['T_g_20'] + T2_Soil['T_g_20'] +
```

```
T3_Soil['T_g_20'] + T4_Soil['T_g_20'] + T5_Soil['T_g_20'] + T6_Soil['T_g_20']
+ T7_Soil['T_g_20'] + T8_Soil['T_g_20'] + T9_Soil['T_g_20'] ) / 9
mergeData_Sub['s_m_20']  =  (T1_Soil['s_m_20'] + T2_Soil['s_m_20'] +
T3_Soil['s_m_20'] + T4_Soil['s_m_20'] + T5_Soil['s_m_20'] + T6_Soil['s_m_20']
+ T7_Soil['s_m_20'] + T8_Soil['s_m_20'] + T9_Soil['s_m_20'] ) / 9

# ============== T_g_35, s_m_35 Available in T2_Soil, T5_Soil, T7_Soil,
T9_Soil sites ==============

mergeData_Sub['T_g_35']  =  ( T2_Soil['T_g_35'] + T5_Soil['T_g_35'] +
T7_Soil['T_g_35'] + T9_Soil['T_g_35'] ) / 4
mergeData_Sub['s_m_35']  =  ( T2_Soil['s_m_35'] + T5_Soil['s_m_35'] +
T7_Soil['s_m_35'] + T9_Soil['s_m_35'] ) / 4

# ============== T_g_50, s_m_50 Available in all sites ==============

mergeData_Sub['T_g_50']  =  (T1_Soil['T_g_50'] + T2_Soil['T_g_50'] +
T3_Soil['T_g_50'] + T4_Soil['T_g_50'] + T5_Soil['T_g_50'] + T6_Soil['T_g_50']
+ T7_Soil['T_g_50'] + T8_Soil['T_g_50'] + T9_Soil['T_g_50'] ) / 9
mergeData_Sub['s_m_50']  =  (T1_Soil['s_m_50'] + T2_Soil['s_m_50'] +
T3_Soil['s_m_50'] + T4_Soil['s_m_50'] + T5_Soil['s_m_50'] + T6_Soil['s_m_50']
+ T7_Soil['s_m_50'] + T8_Soil['s_m_50'] + T9_Soil['s_m_50'] ) / 9

# ============== T_g_75, s_m_75 Available in T1_Soil, T4_Soil, T5_Soil,
T6_Soil, T8_Soil sites ==============

mergeData_Sub['T_g_75']  =  ( T1_Soil['T_g_75'] + T4_Soil['T_g_75'] +
T5_Soil['T_g_75'] + T6_Soil['T_g_75'] + T8_Soil['T_g_75'] ) / 5
mergeData_Sub['s_m_75']  =  ( T1_Soil['s_m_75'] + T4_Soil['s_m_75'] +
T5_Soil['s_m_75'] + T6_Soil['s_m_75'] + T8_Soil['s_m_75'] ) / 5

# ============== T_g_90, s_m_90 Available in T1_Soil, T3_Soil sites
==============

mergeData_Sub['T_g_90']  =  ( T1_Soil['T_g_90'] + T3_Soil['T_g_90'] ) / 2
mergeData_Sub['s_m_90']  =  ( T1_Soil['s_m_90'] + T3_Soil['s_m_90'] ) / 2

# ============== T_g_100, s_m_100 Available in T4_Soil, T6_Soil, T8_Soil
sites ==============

mergeData_Sub['T_g_100']  =  ( T4_Soil['T_g_100'] + T6_Soil['T_g_100'] +
T8_Soil['T_g_100'] ) / 3
mergeData_Sub['s_m_100']  =  ( T4_Soil['s_m_100'] + T6_Soil['s_m_100'] +
T8_Soil['s_m_100'] ) / 3

# ============== T_g_130, s_m_130 Available in T3_Soil site ==============

mergeData_Sub['T_g_130']  =  ( T3_Soil['T_g_130'] )
mergeData_Sub['s_m_130']  =  ( T3_Soil['s_m_130'] )
```

```
# =============== T_g_190, s_m_190 Available in T3_Soil site ===============

mergeData_Sub['T_g_190']  =  ( T3_Soil['T_g_190'] )
mergeData_Sub['s_m_190']  =  ( T3_Soil['s_m_190'] )

Final_Soil = mergeData_Sub
summary(Final_Soil)

##        WY            Year          Month             Day              Hour
##  Min.   :2011   Min.   :2010   Min.   : 1.000   Min.   : 1.00   Min.   :
0.00
##  1st Qu.:2012   1st Qu.:2011   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.75
##  Median :2012   Median :2012   Median : 7.000   Median :16.00   Median
:11.50
##  Mean   :2012   Mean   :2012   Mean   : 6.523   Mean   :15.73   Mean
:11.50
##  3rd Qu.:2013   3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.25
##  Max.   :2014   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.00
##      Minute      T_g_5            s_m_5            T_g_20
##  Min.   :0   Min.   :-6.872   Min.   :0.04139   Min.   :-4.022
##  1st Qu.:0   1st Qu.: 1.328   1st Qu.:0.07964   1st Qu.: 1.833
##  Median :0   Median : 6.964   Median :0.13774   Median : 7.867
##  Mean   :0   Mean   : 9.221   Mean   :0.13480   Mean   : 9.273
##  3rd Qu.:0   3rd Qu.:16.311   3rd Qu.:0.17876   3rd Qu.:16.733
##  Max.   :0   Max.   :34.211   Max.   :0.34720   Max.   :24.444
##      s_m_20           T_g_35           s_m_35           T_g_50
##  Min.   :0.06919   Min.   : 0.675   Min.   :0.0955   Min.   :-0.5394
##  1st Qu.:0.10411   1st Qu.: 4.025   1st Qu.:0.1250   1st Qu.: 2.6067
##  Median :0.15711   Median :10.425   Median :0.1665   Median : 8.2092
##  Mean   :0.15935   Mean   :11.299   Mean   :0.1770   Mean   : 8.9447
##  3rd Qu.:0.21144   3rd Qu.:18.350   3rd Qu.:0.2370   3rd Qu.:15.1414
##  Max.   :0.29844   Max.   :24.750   Max.   :0.2870   Max.   :20.2533
##      s_m_50           T_g_75           s_m_75           T_g_90
##  Min.   :0.07862   Min.   : 0.4558   Min.   :0.0926   Min.   : 1.073
##  1st Qu.:0.10671   1st Qu.: 3.0298   1st Qu.:0.1148   1st Qu.: 3.640
##  Median :0.15106   Median : 7.3498   Median :0.1496   Median : 7.585
##  Mean   :0.15312   Mean   : 7.9400   Mean   :0.1581   Mean   : 7.855
##  3rd Qu.:0.20267   3rd Qu.:12.6498   3rd Qu.:0.2044   3rd Qu.:11.855
##  Max.   :0.26011   Max.   :17.3420   Max.   :0.2464   Max.   :15.240
##      s_m_90          T_g_100          s_m_100          T_g_130
##  Min.   :0.0970   Min.   : 0.3667   Min.   :0.02879   Min.   : 2.700
##  1st Qu.:0.1295   1st Qu.: 2.5000   1st Qu.:0.04365   1st Qu.: 4.500
##  Median :0.1595   Median : 6.3333   Median :0.06181   Median : 7.800
##  Mean   :0.1847   Mean   : 7.0156   Mean   :0.07875   Mean   : 8.302
##  3rd Qu.:0.2445   3rd Qu.:11.2667   3rd Qu.:0.12513   3rd Qu.:12.000
##  Max.   :0.3210   Max.   :15.4333   Max.   :0.16733   Max.   :15.200
##     s_m_130          T_g_190          s_m_190
```

```
##  Min.   :0.0970   Min.   : 3.800   Min.   :0.120
##  1st Qu.:0.1090   1st Qu.: 5.700   1st Qu.:0.137
##  Median :0.1120   Median : 8.300   Median :0.141
##  Mean   :0.1323   Mean   : 8.437   Mean   :0.152
##  3rd Qu.:0.1240   3rd Qu.:11.100   3rd Qu.:0.148
##  Max.   :0.2420   Max.   :13.500   Max.   :0.246
```

#download the SnowDepth datasets which is merged into excel for further analysis.

```
library("writexl")
write_xlsx(Final_Soil,"Soil_merged.xlsx")

#Merging df weather Snow Soil Precipitation into 1
weather_Snow_Soil_PPt_merged <-weather_data_merged %>%
    left_join(Snow_depth, by=c("WY","Year","Month","Day","Hour","Minute"))
%>%
    left_join(Final_Soil, by=c("WY","Year","Month","Day","Hour","Minute"))
%>%
    left_join(Precipitation_merged,
by=c("WY","Year","Month","Day","Hour","Minute"))
head(weather_Snow_Soil_PPt_merged)

## # A tibble: 6 × 34
## # Groups:   WY, Year, Month, Day, Hour [6]
##      WY  Year Month   Day  Hour Minute   T_a    RH   e_a     T_d   S_i
w_s
##   <int> <int> <int> <int> <int>  <int> <dbl> <dbl> <dbl>   <dbl> <dbl>
<dbl>
## 1  2004  2003    10     1     0      0  16.4 0.29   536  -1.57       0
1.03
## 2  2004  2003    10     1     1      0  16.1 0.303  548. -1.3        0
1.17
## 3  2004  2003    10     1     2      0  14.9 0.333  561. -1.03       0  1
## 4  2004  2003    10     1     3      0  14.4 0.357  578. -0.7        0
0.8
## 5  2004  2003    10     1     4      0  14.6 0.363  599. -0.233      0
1.07
## 6  2004  2003    10     1     5      0  14.8 0.363  606. -0.0667     0
1.03
## # … with 22 more variables: w_d <dbl>, z_s <dbl>, T_g_5 <dbl>, s_m_5
<dbl>,
##   T_g_20 <dbl>, s_m_20 <dbl>, T_g_35 <dbl>, s_m_35 <dbl>, T_g_50 <dbl>,
##   s_m_50 <dbl>, T_g_75 <dbl>, s_m_75 <dbl>, T_g_90 <dbl>, s_m_90 <dbl>,
##   T_g_100 <dbl>, s_m_100 <dbl>, T_g_130 <dbl>, s_m_130 <dbl>, T_g_190
<dbl>,
##   s_m_190 <dbl>, ppt_a <dbl>, perc_snow <dbl>

library("writexl")
write_xlsx(weather_Snow_Soil_PPt_merged,"All_4_merged.xlsx")
```

```
#check if NA's Exist
list_na <- colnames(weather_Snow_Soil_PPt_merged)[
apply(weather_Snow_Soil_PPt_merged, 2, anyNA) ]
list_na
```

```
##  [1] "S_i"       "w_s"       "w_d"       "z_s"       "T_g_5"     "s_m_5"
##  [7] "T_g_20"    "s_m_20"    "T_g_35"    "s_m_35"    "T_g_50"    "s_m_50"
## [13] "T_g_75"    "s_m_75"    "T_g_90"    "s_m_90"    "T_g_100"   "s_m_100"
## [19] "T_g_130"   "s_m_130"   "T_g_190"   "s_m_190"   "ppt_a"
"perc_snow"
```

```
#Replace all NA's with 0 since those stations had not begain recording at
that time/year.
weather_Snow_Soil_PPt_merged <- replace(weather_Snow_Soil_PPt_merged,
is.na(weather_Snow_Soil_PPt_merged), 0)

#check if NA's Exist
list_na <- colnames(weather_Snow_Soil_PPt_merged)[
apply(weather_Snow_Soil_PPt_merged, 2, anyNA) ]
list_na
```

```
## character(0)
```

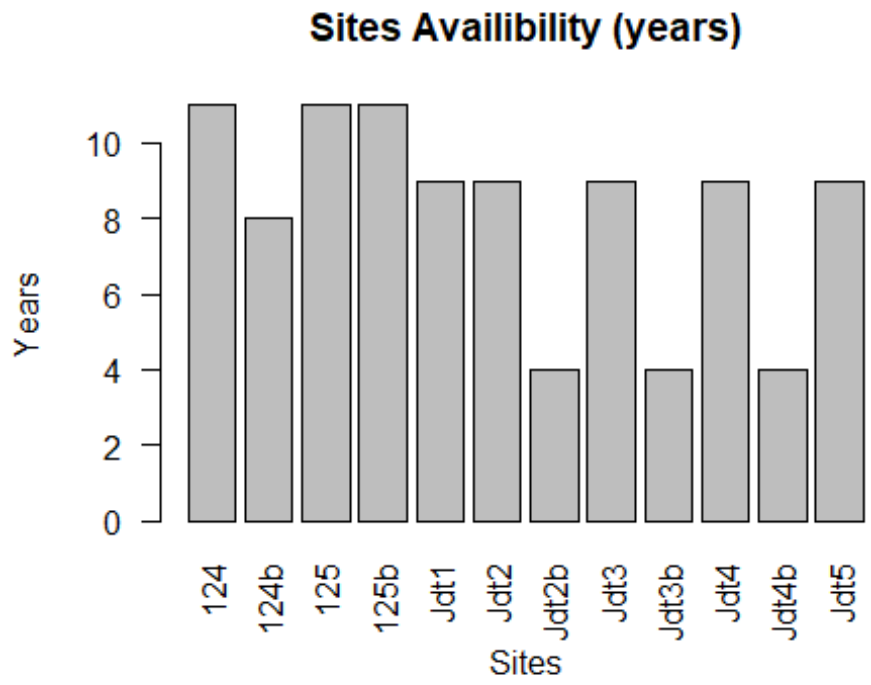
```
#Remove Minute feature as it only contains value 0
weather_Snow_Soil_PPt_merged <- subset(weather_Snow_Soil_PPt_merged, select =
-c(Minute))


summary(weather_Snow_Soil_PPt_merged)
```

```
##        WY            Year          Month             Day              Hour
##  Min.   :2004   Min.   :2003   Min.   : 1.000   Min.   : 1.00   Min.   :
0.0
##  1st Qu.:2006   1st Qu.:2006   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.0
##  Median :2009   Median :2009   Median : 7.000   Median :16.00   Median
:11.0
##  Mean   :2009   Mean   :2009   Mean   : 6.523   Mean   :15.73   Mean
:11.5
##  3rd Qu.:2012   3rd Qu.:2011   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.0
##  Max.   :2015   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.0
##       T_a                RH               e_a               T_d
##  Min.   :-16.792   Min.   :0.06333   Min.   :  61.17   Min.   :-25.3583
##  1st Qu.:  1.725   1st Qu.:0.37500   1st Qu.: 410.33   1st Qu.: -4.9687
##  Median :  6.642   Median :0.53333   Median : 522.42   Median : -1.9667
##  Mean   :  7.758   Mean   :0.53987   Mean   : 548.29   Mean   : -2.0857
##  3rd Qu.: 13.633   3rd Qu.:0.69917   3rd Qu.: 652.75   3rd Qu.:  0.8167
##  Max.   : 34.717   Max.   :1.00000   Max.   :1716.75   Max.   : 15.1167
##       S_i               w_s              w_d              z_s
```

```
##  Min.   :   0.00   Min.   :0.0000   Min.   :  0.00   Min.   : 0.000
##  1st Qu.:   0.00   1st Qu.:0.0000   1st Qu.:  0.00   1st Qu.: 0.000
##  Median :   0.00   Median :0.0000   Median :  0.00   Median : 0.000
##  Mean   :  34.41   Mean   :0.4425   Mean   : 37.56   Mean   : 4.047
##  3rd Qu.:   0.00   3rd Qu.:0.0000   3rd Qu.:  0.00   3rd Qu.: 4.364
##  Max.   :1040.33   Max.   :9.8667   Max.   :359.33   Max.   :42.091
##      T_g_5             s_m_5             T_g_20           s_m_20
##  Min.   :-6.872   Min.   :0.00000   Min.   :-4.022   Min.   :0.00000
##  1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.: 0.000   1st Qu.:0.00000
##  Median : 0.000   Median :0.00000   Median : 0.000   Median :0.00000
##  Mean   : 3.353   Mean   :0.04901   Mean   : 3.372   Mean   :0.05794
##  3rd Qu.: 2.339   3rd Qu.:0.08813   3rd Qu.: 2.800   3rd Qu.:0.11233
##  Max.   :34.211   Max.   :0.34720   Max.   :24.444   Max.   :0.29844
##      T_g_35            s_m_35            T_g_50            s_m_50
##  Min.   : 0.000   Min.   :0.00000   Min.   :-0.5394   Min.   :0.00000
##  1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.: 0.0000   1st Qu.:0.00000
##  Median : 0.000   Median :0.00000   Median : 0.0000   Median :0.00000
##  Mean   : 4.108   Mean   :0.06436   Mean   : 3.2524   Mean   :0.05568
##  3rd Qu.: 5.500   3rd Qu.:0.12950   3rd Qu.: 3.5490   3rd Qu.:0.11300
##  Max.   :24.750   Max.   :0.28700   Max.   :20.2533   Max.   :0.26011
##      T_g_75            s_m_75            T_g_90           s_m_90
##  Min.   : 0.000   Min.   :0.0000   Min.   : 0.000   Min.   :0.00000
##  1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.: 0.000   1st Qu.:0.00000
##  Median : 0.000   Median :0.0000   Median : 0.000   Median :0.00000
##  Mean   : 2.887   Mean   :0.0575   Mean   : 2.856   Mean   :0.06718
##  3rd Qu.: 3.588   3rd Qu.:0.1182   3rd Qu.: 4.386   3rd Qu.:0.13350
##  Max.   :17.342   Max.   :0.2464   Max.   :15.240   Max.   :0.32100
##     T_g_100           s_m_100           T_g_130           s_m_130
##  Min.   : 0.000   Min.   :0.00000   Min.   : 0.000   Min.   :0.00000
##  1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.: 0.000   1st Qu.:0.00000
##  Median : 0.000   Median :0.00000   Median : 0.000   Median :0.00000
##  Mean   : 2.551   Mean   :0.02864   Mean   : 3.019   Mean   :0.04812
##  3rd Qu.: 3.200   3rd Qu.:0.04616   3rd Qu.: 5.200   3rd Qu.:0.11000
##  Max.   :15.433   Max.   :0.16733   Max.   :15.200   Max.   :0.24200
##     T_g_190           s_m_190             ppt_a            perc_snow
##  Min.   : 0.000   Min.   :0.00000   Min.   : 0.00000   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.: 0.00000   1st Qu.:0.0500
##  Median : 0.000   Median :0.00000   Median : 0.00000   Median :1.0000
##  Mean   : 3.068   Mean   :0.05527   Mean   : 0.06945   Mean   :0.6661
##  3rd Qu.: 6.300   3rd Qu.:0.13800   3rd Qu.: 0.00000   3rd Qu.:1.0000
##  Max.   :13.500   Max.   :0.24600   Max.   :16.33333   Max.   :1.0000
```

```
library("writexl")
write_xlsx(weather_Snow_Soil_PPt_merged,"All_4_merged_cleaned.xlsx")

#install.packages("tabplot", dependencies = TRUE)
require(ggplot2)
#install.packages('tabplot')
#library(tabplot)
```

```
#Making barplot to see how many years each station recorded data for

locations =
c('125b','125b','125b','125b','125b','125b','125b','125b','125b','125b','125b
','125','125','125','125','125','125','125','125','125','125','125','Jdt1','J
dt1','Jdt1','Jdt1','Jdt1','Jdt1','Jdt1','Jdt1','Jdt1','Jdt2b','Jdt2b','Jdt2b'
,'Jdt2b','Jdt2','Jdt2','Jdt2','Jdt2','Jdt2','Jdt2','Jdt2','Jdt2','Jdt2','Jdt3
','Jdt3','Jdt3','Jdt3','Jdt3','Jdt3','Jdt3','Jdt3','Jdt3','Jdt3b','Jdt3b','Jd
t3b','Jdt3b','Jdt4b','Jdt4b','Jdt4b','Jdt4b','Jdt4','Jdt4','Jdt4','Jdt4','Jdt
4','Jdt4','Jdt4','Jdt4','Jdt4','Jdt5','Jdt5','Jdt5','Jdt5','Jdt5','Jdt5','Jdt
5','Jdt5','Jdt5','124b','124b','124b','124b','124b','124b','124b','124b','124
','124','124','124','124','124','124','124','124','124','124')
Sites <- data.frame(locations)
x <- table(Sites$locations)
barplot(x, main="Sites Availibility (years)", xlab="Sites", ylab="Years",
las=2)
```



#Relationship Between Numerical Variables

```
#Correlation Matrix
#(take all features except Minute because since it's only value is 0, it
shows NA in correlation with other variables which will disrupt correlation
plot later)

#Ignore standard deviation warning using suppressWarnings function
suppressWarnings({corr <- round(cor(weather_Snow_Soil_PPt_merged), 1)})
corr
```

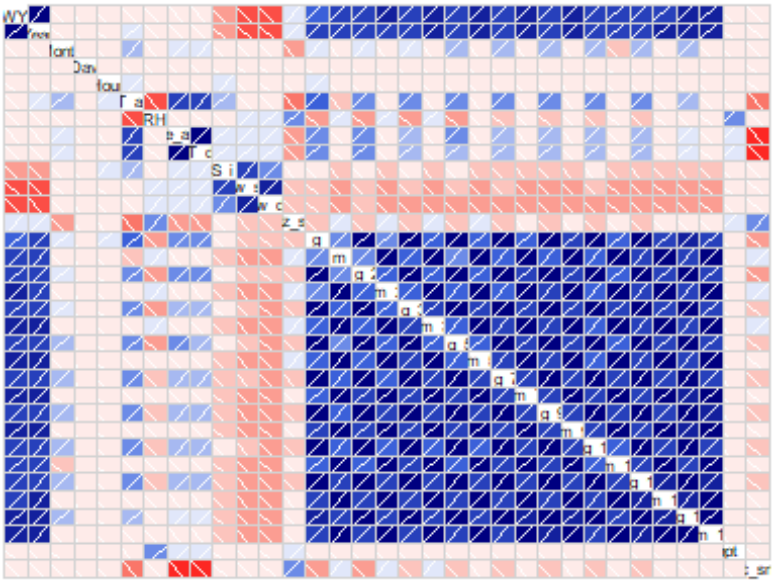
```
##              WY Year Month Day Hour  T_a   RH  e_a  T_d  S_i  w_s  w_d  z_s
## WY          1.0  1.0   0.0   0  0.0  0.0 -0.1  0.0 -0.1 -0.4 -0.6 -0.6  0.1
## Year        1.0  1.0  -0.1   0  0.0  0.1 -0.1  0.0  0.0 -0.3 -0.6 -0.6  0.1
## Month       0.0 -0.1   1.0   0  0.0  0.2 -0.1  0.1  0.1  0.0  0.0  0.0 -0.4
## Day         0.0  0.0   0.0   1  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
## Hour        0.0  0.0   0.0   0  1.0  0.1 -0.1  0.0  0.0  0.1  0.0  0.0  0.0
## T_a         0.0  0.1   0.2   0  0.1  1.0 -0.7  0.6  0.6  0.2  0.0  0.0 -0.5
## RH         -0.1 -0.1  -0.1   0 -0.1 -0.7  1.0  0.0  0.0 -0.1  0.1  0.1  0.3
## e_a         0.0  0.0   0.1   0  0.0  0.6  0.0  1.0  1.0  0.1  0.1  0.1 -0.3
## T_d        -0.1  0.0   0.1   0  0.0  0.6  0.0  1.0  1.0  0.1  0.1  0.1 -0.3
## S_i        -0.4 -0.3   0.0   0  0.1  0.2 -0.1  0.1  0.1  1.0  0.6  0.4 -0.1
## w_s        -0.6 -0.6   0.0   0  0.0  0.0  0.1  0.1  0.1  0.6  1.0  0.8 -0.2
## w_d        -0.6 -0.6   0.0   0  0.0  0.0  0.1  0.1  0.1  0.4  0.8  1.0 -0.2
## z_s         0.1  0.1  -0.4   0  0.0 -0.5  0.3 -0.3 -0.3 -0.1 -0.2 -0.2  1.0
## T_g_5       0.5  0.6   0.1   0  0.1  0.5 -0.3  0.3  0.3 -0.1 -0.2 -0.2 -0.2
## s_m_5       0.7  0.7  -0.1   0  0.0 -0.2  0.1 -0.1 -0.1 -0.2 -0.3 -0.3  0.1
## T_g_20      0.6  0.6   0.1   0  0.0  0.4 -0.3  0.3  0.3 -0.1 -0.2 -0.2 -0.2
## s_m_20      0.8  0.7  -0.1   0  0.0 -0.1  0.1 -0.1 -0.1 -0.2 -0.3 -0.3  0.1
## T_g_35      0.7  0.7   0.1   0  0.0  0.4 -0.3  0.2  0.2 -0.1 -0.2 -0.3 -0.2
## s_m_35      0.8  0.8  -0.1   0  0.0 -0.1  0.1 -0.1 -0.1 -0.2 -0.3 -0.3  0.1
## T_g_50      0.6  0.6   0.2   0  0.0  0.4 -0.3  0.3  0.2 -0.1 -0.2 -0.3 -0.2
## s_m_50      0.8  0.8  -0.1   0  0.0 -0.1  0.0 -0.1 -0.1 -0.2 -0.3 -0.3  0.1
## T_g_75      0.7  0.7   0.2   0  0.0  0.4 -0.2  0.2  0.2 -0.2 -0.2 -0.3 -0.2
## s_m_75      0.8  0.8  -0.1   0  0.0  0.0  0.0  0.0 -0.1 -0.2 -0.3 -0.3  0.0
## T_g_90      0.7  0.7   0.2   0  0.0  0.3 -0.2  0.2  0.2 -0.2 -0.3 -0.3 -0.2
## s_m_90      0.8  0.8  -0.1   0  0.0  0.0  0.0  0.0  0.0 -0.2 -0.3 -0.3  0.0
## T_g_100     0.6  0.6   0.2   0  0.0  0.4 -0.2  0.2  0.2 -0.1 -0.2 -0.3 -0.2
## s_m_100     0.7  0.7  -0.2   0  0.0  0.0  0.0  0.0  0.0 -0.2 -0.3 -0.3  0.0
## T_g_130     0.7  0.7   0.2   0  0.0  0.3 -0.2  0.2  0.2 -0.2 -0.3 -0.3 -0.2
## s_m_130     0.7  0.8  -0.1   0  0.0  0.0  0.0  0.0  0.0 -0.2 -0.3 -0.3  0.0
## T_g_190     0.8  0.8   0.2   0  0.0  0.2 -0.1  0.1  0.1 -0.2 -0.3 -0.3 -0.1
## s_m_190     0.8  0.8   0.0   0  0.0  0.0  0.0  0.0  0.0 -0.2 -0.3 -0.3  0.0
## ppt_a       0.0  0.0   0.0   0  0.0 -0.1  0.3  0.1  0.1  0.0  0.0  0.0  0.1
## perc_snow   0.0  0.0  -0.1   0  0.0 -0.5  0.0 -0.8 -0.8 -0.1  0.0  0.0  0.3
##            T_g_5 s_m_5 T_g_20 s_m_20 T_g_35 s_m_35 T_g_50 s_m_50 T_g_75
s_m_75
## WY           0.5   0.7    0.6    0.8    0.7    0.8    0.6    0.8    0.7
0.8
## Year         0.6   0.7    0.6    0.7    0.7    0.8    0.6    0.8    0.7
0.8
## Month        0.1  -0.1    0.1   -0.1    0.1   -0.1    0.2   -0.1    0.2   -
0.1
## Day          0.0   0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0
0.0
## Hour         0.1   0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0
0.0
## T_a          0.5  -0.2    0.4   -0.1    0.4   -0.1    0.4   -0.1    0.4
0.0
## RH          -0.3   0.1   -0.3    0.1   -0.3    0.1   -0.3    0.0   -0.2
0.0
```

```
## e_a         0.3  -0.1    0.3   -0.1    0.2   -0.1    0.3   -0.1    0.2
0.0
## T_d         0.3  -0.1    0.3   -0.1    0.2   -0.1    0.2   -0.1    0.2   -
0.1
## S_i        -0.1  -0.2   -0.1   -0.2   -0.1   -0.2   -0.1   -0.2   -0.2   -
0.2
## w_s        -0.2  -0.3   -0.2   -0.3   -0.2   -0.3   -0.2   -0.3   -0.2   -
0.3
## w_d        -0.2  -0.3   -0.2   -0.3   -0.3   -0.3   -0.3   -0.3   -0.3   -
0.3
## z_s        -0.2   0.1   -0.2    0.1   -0.2    0.1   -0.2    0.1   -0.2
0.0
## T_g_5       1.0   0.4    1.0    0.4    0.9    0.5    1.0    0.5    0.9
0.5
## s_m_5       0.4   1.0    0.4    1.0    0.5    1.0    0.4    0.9    0.5
0.9
## T_g_20      1.0   0.4    1.0    0.5    1.0    0.5    1.0    0.5    1.0
0.6
## s_m_20      0.4   1.0    0.5    1.0    0.5    1.0    0.5    1.0    0.5
1.0
## T_g_35      0.9   0.5    1.0    0.5    1.0    0.6    1.0    0.6    1.0
0.6
## s_m_35      0.5   1.0    0.5    1.0    0.6    1.0    0.5    1.0    0.6
1.0
## T_g_50      1.0   0.4    1.0    0.5    1.0    0.5    1.0    0.5    1.0
0.6
## s_m_50      0.5   0.9    0.5    1.0    0.6    1.0    0.5    1.0    0.6
1.0
## T_g_75      0.9   0.5    1.0    0.5    1.0    0.6    1.0    0.6    1.0
0.6
## s_m_75      0.5   0.9    0.6    1.0    0.6    1.0    0.6    1.0    0.6
1.0
## T_g_90      0.9   0.5    1.0    0.6    1.0    0.6    1.0    0.6    1.0
0.7
## s_m_90      0.6   0.9    0.6    0.9    0.7    0.9    0.6    1.0    0.7
1.0
## T_g_100     0.9   0.5    1.0    0.5    1.0    0.5    1.0    0.5    1.0
0.6
## s_m_100     0.5   0.9    0.5    0.9    0.6    0.9    0.5    0.9    0.6
1.0
## T_g_130     0.9   0.6    0.9    0.6    1.0    0.7    1.0    0.7    1.0
0.7
## s_m_130     0.6   0.9    0.6    0.9    0.7    0.9    0.7    0.9    0.7
0.9
## T_g_190     0.8   0.7    0.9    0.7    0.9    0.7    0.9    0.7    0.9
0.8
## s_m_190     0.6   0.9    0.7    0.9    0.7    0.9    0.7    0.9    0.7
0.9
## ppt_a       0.0   0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0
0.0
```

```
## perc_snow  -0.3   0.1   -0.3    0.1   -0.2    0.1   -0.2    0.0   -0.2
0.0
##           T_g_90 s_m_90 T_g_100 s_m_100 T_g_130 s_m_130 T_g_190 s_m_190
ppt_a
## WY           0.7    0.8     0.6     0.7     0.7     0.7     0.8     0.8
0.0
## Year         0.7    0.8     0.6     0.7     0.7     0.8     0.8     0.8
0.0
## Month        0.2   -0.1     0.2    -0.2     0.2    -0.1     0.2     0.0
0.0
## Day          0.0    0.0     0.0     0.0     0.0     0.0     0.0     0.0
0.0
## Hour         0.0    0.0     0.0     0.0     0.0     0.0     0.0     0.0
0.0
## T_a          0.3    0.0     0.4     0.0     0.3     0.0     0.2     0.0  -
0.1
## RH          -0.2    0.0    -0.2     0.0    -0.2     0.0    -0.1     0.0
0.3
## e_a          0.2    0.0     0.2     0.0     0.2     0.0     0.1     0.0
0.1
## T_d          0.2    0.0     0.2     0.0     0.2     0.0     0.1     0.0
0.1
## S_i         -0.2   -0.2    -0.1    -0.2    -0.2    -0.2    -0.2    -0.2
0.0
## w_s         -0.3   -0.3    -0.2    -0.3    -0.3    -0.3    -0.3    -0.3
0.0
## w_d         -0.3   -0.3    -0.3    -0.3    -0.3    -0.3    -0.3    -0.3
0.0
## z_s         -0.2    0.0    -0.2     0.0    -0.2     0.0    -0.1     0.0
0.1
## T_g_5        0.9    0.6     0.9     0.5     0.9     0.6     0.8     0.6
0.0
## s_m_5        0.5    0.9     0.5     0.9     0.6     0.9     0.7     0.9
0.0
## T_g_20       1.0    0.6     1.0     0.5     0.9     0.6     0.9     0.7
0.0
## s_m_20       0.6    0.9     0.5     0.9     0.6     0.9     0.7     0.9
0.0
## T_g_35       1.0    0.7     1.0     0.6     1.0     0.7     0.9     0.7
0.0
## s_m_35       0.6    0.9     0.5     0.9     0.7     0.9     0.7     0.9
0.0
## T_g_50       1.0    0.6     1.0     0.5     1.0     0.7     0.9     0.7
0.0
## s_m_50       0.6    1.0     0.5     0.9     0.7     0.9     0.7     0.9
0.0
## T_g_75       1.0    0.7     1.0     0.6     1.0     0.7     0.9     0.7
0.0
## s_m_75       0.7    1.0     0.6     1.0     0.7     0.9     0.8     0.9
0.0
```

```
## T_g_90       1.0    0.7     1.0     0.6     1.0     0.7     1.0     0.8
0.0
## s_m_90       0.7    1.0     0.6     1.0     0.7     1.0     0.8     1.0
0.0
## T_g_100      1.0    0.6     1.0     0.5     1.0     0.7     0.9     0.7
0.0
## s_m_100      0.6    1.0     0.5     1.0     0.6     0.9     0.7     0.9
0.0
## T_g_130      1.0    0.7     1.0     0.6     1.0     0.8     1.0     0.8
0.0
## s_m_130      0.7    1.0     0.7     0.9     0.8     1.0     0.8     1.0
0.0
## T_g_190      1.0    0.8     0.9     0.7     1.0     0.8     1.0     0.9
0.0
## s_m_190      0.8    1.0     0.7     0.9     0.8     1.0     0.9     1.0
0.0
## ppt_a        0.0    0.0     0.0     0.0     0.0     0.0     0.0     0.0
1.0
## perc_snow   -0.2    0.0    -0.2     0.0    -0.2     0.0    -0.1     0.0  -
0.1
##           perc_snow
## WY              0.0
## Year            0.0
## Month          -0.1
## Day             0.0
## Hour            0.0
## T_a            -0.5
## RH              0.0
## e_a            -0.8
## T_d            -0.8
## S_i            -0.1
## w_s             0.0
## w_d             0.0
## z_s             0.3
## T_g_5          -0.3
## s_m_5           0.1
## T_g_20         -0.3
## s_m_20          0.1
## T_g_35         -0.2
## s_m_35          0.1
## T_g_50         -0.2
## s_m_50          0.0
## T_g_75         -0.2
## s_m_75          0.0
## T_g_90         -0.2
## s_m_90          0.0
## T_g_100        -0.2
## s_m_100         0.0
## T_g_130        -0.2
## s_m_130         0.0
```

```
## T_g_190         -0.1
## s_m_190          0.0
## ppt_a           -0.1
## perc_snow        1.0

correlation <- data.frame(corr)
correlation <- as.data.frame(correlation)
write_xlsx(correlation,"correlation.xlsx")

#install.packages('corrgram')
library(corrgram)

## Warning: package 'corrgram' was built under R version 4.2.3

#Visualizing using a Correlogram
corrgram(corr)
```



#Before handling outliers:

#########Multiple Linear Regression on the combined with correlation dataset##

```
#Fitting multiple linear regression model using snow depth as response
lm.fits <- lm(z_s ~. , data = weather_Snow_Soil_PPt_merged)
summary(lm.fits)

##
## Call:
## lm(formula = z_s ~ ., data = weather_Snow_Soil_PPt_merged)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.8031  -3.2496  -0.7048   2.1891  29.2952
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.898e+03  3.147e+01 -60.316  < 2e-16 ***
## WY           3.613e+00  9.372e-02  38.550  < 2e-16 ***
## Year        -2.667e+00  9.348e-02 -28.535  < 2e-16 ***
## Month       -1.088e+00  1.143e-02 -95.244  < 2e-16 ***
## Day          4.289e-03  2.070e-03   2.072 0.038285 *
## Hour         1.785e-02  2.739e-03   6.517 7.20e-11 ***
## T_a         -6.682e-02  1.180e-02  -5.662 1.50e-08 ***
## RH           7.572e+00  3.532e-01  21.437  < 2e-16 ***
## e_a          9.559e-03  4.494e-04  21.268  < 2e-16 ***
## T_d         -5.273e-01  2.327e-02 -22.663  < 2e-16 ***
## S_i          4.407e-03  1.733e-04  25.432  < 2e-16 ***
## w_s         -3.968e-01  3.315e-02 -11.972  < 2e-16 ***
## w_d         -7.052e-03  4.207e-04 -16.764  < 2e-16 ***
## T_g_5        3.124e-01  1.375e-02  22.724  < 2e-16 ***
## s_m_5       -7.247e+00  1.666e+00  -4.350 1.36e-05 ***
## T_g_20      -8.281e-02  3.977e-02  -2.082 0.037324 *
## s_m_20      -2.565e+01  2.594e+00  -9.889  < 2e-16 ***
## T_g_35       5.551e-01  4.685e-02  11.849  < 2e-16 ***
## s_m_35       5.280e+01  2.560e+00  20.626  < 2e-16 ***
## T_g_50      -1.036e+00  1.221e-01  -8.489  < 2e-16 ***
## s_m_50       6.663e+01  3.507e+00  19.000  < 2e-16 ***
## T_g_75       3.077e+00  8.426e-02  36.513  < 2e-16 ***
## s_m_75      -1.548e+02  3.547e+00 -43.660  < 2e-16 ***
## T_g_90      -4.556e+00  1.075e-01 -42.382  < 2e-16 ***
## s_m_90      -5.899e+01  1.357e+00 -43.461  < 2e-16 ***
## T_g_100     -5.712e-02  7.181e-02  -0.795 0.426347
## s_m_100      1.274e+02  3.108e+00  40.991  < 2e-16 ***
## T_g_130      5.063e-01  1.428e-01   3.546 0.000391 ***
## s_m_130     -3.107e+01  1.641e+00 -18.929  < 2e-16 ***
## T_g_190      1.096e+00  1.066e-01  10.281  < 2e-16 ***
## s_m_190      5.546e+01  1.857e+00  29.870  < 2e-16 ***
## ppt_a        2.219e-01  5.267e-02   4.213 2.52e-05 ***
## perc_snow    2.505e+00  7.530e-02  33.266  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.64 on 96400 degrees of freedom
## Multiple R-squared:  0.4774, Adjusted R-squared:  0.4772
## F-statistic:  2752 on 32 and 96400 DF,  p-value: < 2.2e-16

#Confidence Interval
confint(lm.fits)
```

```
##                      2.5 %        97.5 %
## (Intercept) -1.959747e+03 -1.836391e+03
## WY           3.429064e+00  3.796427e+00
## Year        -2.850694e+00 -2.484250e+00
## Month       -1.110584e+00 -1.065798e+00
## Day          2.315304e-04  8.347494e-03
## Hour         1.248095e-02  2.321664e-02
## T_a         -8.995348e-02 -4.369296e-02
## RH           6.879414e+00  8.263944e+00
## e_a          8.678013e-03  1.043981e-02
## T_d         -5.728747e-01 -4.816715e-01
## S_i          4.067603e-03  4.746922e-03
## w_s         -4.618165e-01 -3.318757e-01
## w_d         -7.876450e-03 -6.227495e-03
## T_g_5        2.854661e-01  3.393591e-01
## s_m_5       -1.051248e+01 -3.981576e+00
## T_g_20      -1.607658e-01 -4.861980e-03
## s_m_20      -3.073523e+01 -2.056732e+01
## T_g_35       4.632903e-01  6.469330e-01
## s_m_35       4.777867e+01  5.781226e+01
## T_g_50      -1.275681e+00 -7.971112e-01
## s_m_50       5.975650e+01  7.350343e+01
## T_g_75       2.911478e+00  3.241775e+00
## s_m_75      -1.617918e+02 -1.478895e+02
## T_g_90      -4.767057e+00 -4.345629e+00
## s_m_90      -6.164940e+01 -5.632881e+01
## T_g_100     -1.978596e-01  8.362077e-02
## s_m_100      1.213061e+02  1.334893e+02
## T_g_130      2.264354e-01  7.861246e-01
## s_m_130     -3.428600e+01 -2.785201e+01
## T_g_190      8.871484e-01  1.305059e+00
## s_m_190      5.182061e+01  5.909885e+01
## ppt_a        1.186687e-01  3.251293e-01
## perc_snow    2.357320e+00  2.652494e+00
```

```
#Creating our own function for MSE and RMSE Calculations
MSE <- mean(lm.fits$residuals^2)
RMSE <- sqrt(MSE)

cat("Mean Square Error: ", MSE)
```

```
## Mean Square Error:  31.79896
```

```
cat(", Root Mean Square Error: ", RMSE)
```

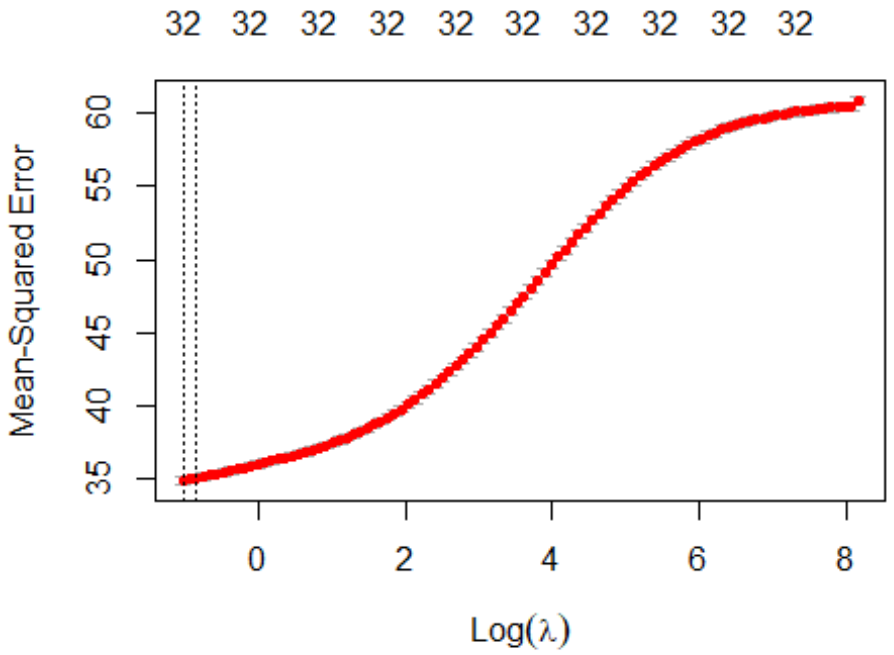
```
## , Root Mean Square Error:  5.639057
```

```
#Compute Error Rate using RSE - Error Rate is RSE divided by mean of response variable
error <- sigma(lm.fits)/mean(weather_Snow_Soil_PPt_merged$z_s)
cat("\nError rate: ", error)
```

```
##
## Error rate:  1.393717
```

#########Ridge Regression on the combined with correlation dataset###################################

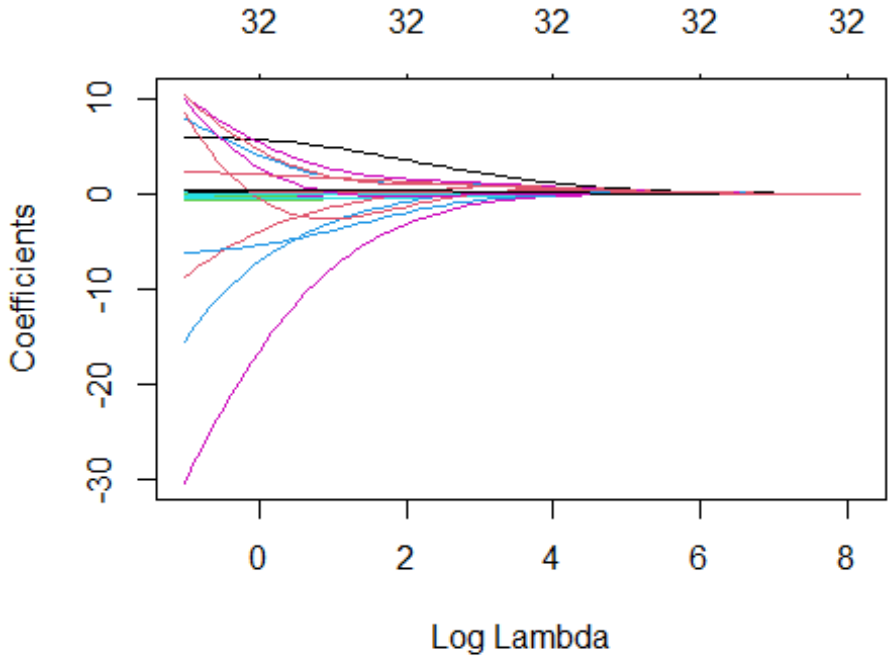
```
library(glmnet)

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loaded glmnet 4.1-6

library(mgcv)

## Loading required package: nlme

##
## Attaching package: 'nlme'

## The following object is masked from 'package:dplyr':
##
##     collapse

## This is mgcv 1.8-41. For overview type 'help("mgcv-package")'.

library(visreg)

#Ridge
#pass x matrix and y vector:
x <- model.matrix(z_s ~ ., data=weather_Snow_Soil_PPt_merged )[, -1]
y <- weather_Snow_Soil_PPt_merged$z_s

model <- glmnet(x, y, alpha = 0)

#find optimal lambda value
ridge.mod <- cv.glmnet(x, y, alpha = 0)
plot(ridge.mod)
```




```
min_lambda_ridge <- ridge.mod$lambda.min
cat("Minimum value of Lambda for ridge: ", min_lambda_ridge,"\n")

## Minimum value of Lambda for ridge:  0.3545392

ridge.mod2 <- glmnet(x, y, alpha = 0, lambda = min_lambda_ridge)

coef(ridge.mod2)

## 33 x 1 sparse Matrix of class "dgCMatrix"
##                        s0
## (Intercept) -1.181672e+03
## WY           4.031501e-01
## Year         1.876251e-01
## Month       -7.219317e-01
## Day          6.744209e-04
## Hour         3.195443e-02
## T_a         -1.641354e-01
## RH           5.917024e+00
## e_a          2.132477e-03
## T_d         -2.043004e-01
## S_i          3.806279e-03
## w_s         -4.540041e-01
## w_d         -1.005171e-02
## T_g_5        1.545860e-01
## s_m_5       -8.604764e+00
## T_g_20      -2.333829e-02
```

```
## s_m_20        8.220570e+00
## T_g_35       -9.548487e-02
## s_m_35        1.074603e+01
## T_g_50       -8.099965e-02
## s_m_50        1.083329e+01
## T_g_75        7.031872e-02
## s_m_75       -1.543293e+01
## T_g_90       -2.026089e-01
## s_m_90       -3.104350e+01
## T_g_100       1.119208e-01
## s_m_100       7.392533e+00
## T_g_130      -6.779499e-02
## s_m_130      -6.342890e+00
## T_g_190       7.722809e-02
## s_m_190       1.016684e+01
## ppt_a         3.932585e-01
## perc_snow     2.345991e+00
```

```r
#produce Ridge trace plot
plot(model, xvar = "lambda")
```

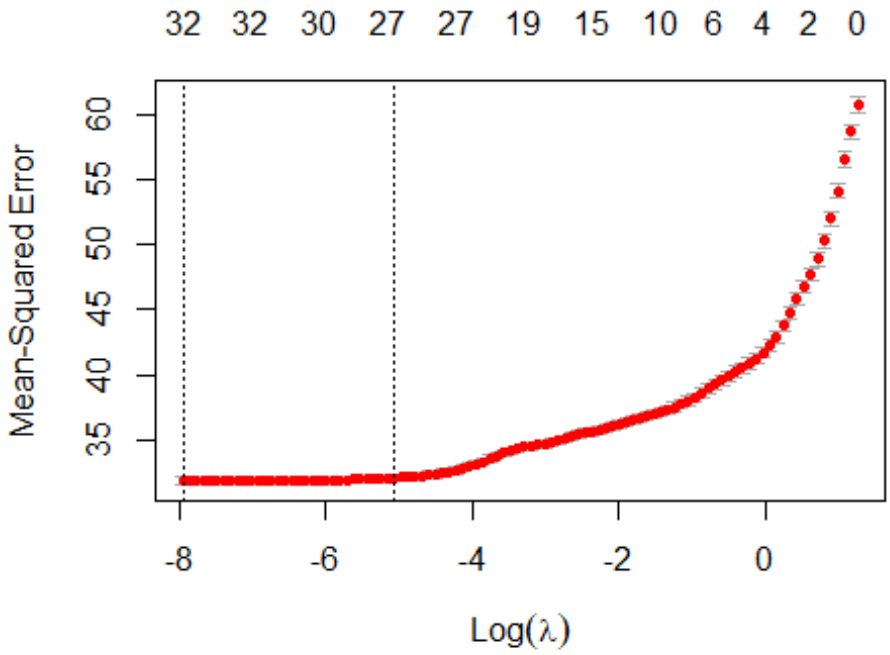



```r
#use fitted best model to make predictions on train data

y_pred_ridge <- predict(ridge.mod2, s = min_lambda_ridge, newx=x)

mse_ridge <- mean((y - y_pred_ridge)^2)
rmse_ridge <- sqrt(mse_ridge)
```

```
RSS_ridge <- sum((y - y_pred_ridge)^2)
TSS_ridge <- (sum((y - mean(y))^2))
rsquared_ridge <- 1-(RSS_ridge/TSS_ridge)

cat("Mean Square Error Ridge: ", mse_ridge)

## Mean Square Error Ridge:  34.85472

cat("\nRoot Mean Square Error Ridge: ", rmse_ridge)

##
## Root Mean Square Error Ridge:  5.903789

cat("\nR^2 Ridge: ", rsquared_ridge)

##
## R^2 Ridge:  0.4272046
```

#########Lasso Regression on the combined with correlation dataset####################################

```
#Lasso

#pass x matrix and y vector:
x <- model.matrix(z_s ~ ., data=weather_Snow_Soil_PPt_merged )[, -1]
y <- weather_Snow_Soil_PPt_merged$z_s

lasso.mod <- cv.glmnet(x, y, alpha = 1)
#lasso.mod <- cv.glmnet(x, y, alpha = 1)

min_lambda_lasso <- lasso.mod$lambda.min
cat("Minimum value of Lambda: ", min_lambda_lasso,"\n")

## Minimum value of Lambda:  0.0003545392

#produce plot of test MSE by lambda value
plot(lasso.mod)
```




```
lasso.mod2 <- glmnet(x, y, alpha = 1, lambda = min_lambda_lasso)

coef(lasso.mod2)
```

```
## 33 x 1 sparse Matrix of class "dgCMatrix"
##                          s0
## (Intercept) -1.934955e+03
## WY           3.575754e+00
## Year        -2.612038e+00
## Month       -1.081935e+00
## Day          4.203899e-03
## Hour         1.935769e-02
## T_a         -7.002714e-02
## RH           7.506882e+00
## e_a          9.366381e-03
## T_d         -5.177004e-01
## S_i          4.423670e-03
## w_s         -3.901256e-01
## w_d         -6.851396e-03
## T_g_5        3.163196e-01
## s_m_5       -6.629003e+00
## T_g_20      -2.179148e-01
## s_m_20      -2.389271e+01
## T_g_35       5.523899e-01
## s_m_35       5.634055e+01
## T_g_50      -5.441189e-01
## s_m_50       6.248502e+01
```

```
## T_g_75       2.845577e+00
## s_m_75      -1.620422e+02
## T_g_90      -5.115150e+00
## s_m_90      -6.020086e+01
## T_g_100     -7.245004e-02
## s_m_100      1.347527e+02
## T_g_130      9.773420e-01
## s_m_130     -2.980825e+01
## T_g_190      1.000962e+00
## s_m_190      5.705120e+01
## ppt_a        2.163887e-01
## perc_snow    2.500357e+00
```

```
#use fitted best model to make predictions on train data

y_pred_lasso <- predict(lasso.mod2, s = min_lambda_lasso, newx=x)

mse_lasso <- mean((y - y_pred_lasso)^2)
rmse_lasso <- sqrt(mse_lasso)
RSS_lasso <- sum((y - y_pred_lasso)^2)
TSS_lasso <- (sum((y - mean(y))^2))
rsquared_lasso <- 1-(RSS_lasso/TSS_lasso)

cat("Mean Square Error Lasso: ", mse_lasso)

## Mean Square Error Lasso:  31.81502

cat("\n Root Mean Square Error Lasso: ", rmse_lasso)

##
##  Root Mean Square Error Lasso:  5.64048

cat("\n R^2 Lasso: ", rsquared_lasso)

##
##  R^2 Lasso:  0.4771584
```

######################Handling correlation$data##

```
#Handling the correlation by merging all the T_g features into one and all
s_m features into one
weather_Snow_Soil_PPt_merged_cor<-weather_Snow_Soil_PPt_merged
weather_Snow_Soil_PPt_merged_cor$T_g <-
rowMeans(weather_Snow_Soil_PPt_merged_cor[ ,
c("T_g_5","T_g_20","T_g_35","T_g_50","T_g_75","T_g_90","T_g_100","T_g_130","T
_g_190")])

weather_Snow_Soil_PPt_merged_cor <- subset(weather_Snow_Soil_PPt_merged_cor,
select = -
c(T_g_5,T_g_20,T_g_35,T_g_50,T_g_75,T_g_90,T_g_100,T_g_130,T_g_190))
```

```
weather_Snow_Soil_PPt_merged_cor$s_m <-
rowMeans(weather_Snow_Soil_PPt_merged_cor[ ,
c("s_m_5","s_m_20","s_m_35","s_m_50","s_m_75","s_m_90","s_m_100","s_m_130","s
_m_190")])

weather_Snow_Soil_PPt_merged_cor <- subset(weather_Snow_Soil_PPt_merged_cor,
select = -
c(s_m_5,s_m_20,s_m_35,s_m_50,s_m_75,s_m_90,s_m_100,s_m_130,s_m_190))

summary(weather_Snow_Soil_PPt_merged_cor)
```

```
##        WY            Year          Month             Day              Hour
##  Min.   :2004   Min.   :2003   Min.   : 1.000   Min.   : 1.00   Min.   :
0.0
##  1st Qu.:2006   1st Qu.:2006   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.0
##  Median :2009   Median :2009   Median : 7.000   Median :16.00   Median
:11.0
##  Mean   :2009   Mean   :2009   Mean   : 6.523   Mean   :15.73   Mean
:11.5
##  3rd Qu.:2012   3rd Qu.:2011   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.0
##  Max.   :2015   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.0
##       T_a                RH               e_a               T_d
##  Min.   :-16.792   Min.   :0.06333   Min.   :  61.17   Min.   :-25.3583
##  1st Qu.:  1.725   1st Qu.:0.37500   1st Qu.: 410.33   1st Qu.: -4.9687
##  Median :  6.642   Median :0.53333   Median : 522.42   Median : -1.9667
##  Mean   :  7.758   Mean   :0.53987   Mean   : 548.29   Mean   : -2.0857
##  3rd Qu.: 13.633   3rd Qu.:0.69917   3rd Qu.: 652.75   3rd Qu.:  0.8167
##  Max.   : 34.717   Max.   :1.00000   Max.   :1716.75   Max.   : 15.1167
##       S_i               w_s              w_d              z_s
##  Min.   :   0.00   Min.   :0.0000   Min.   :  0.00   Min.   : 0.000
##  1st Qu.:   0.00   1st Qu.:0.0000   1st Qu.:  0.00   1st Qu.: 0.000
##  Median :   0.00   Median :0.0000   Median :  0.00   Median : 0.000
##  Mean   :  34.41   Mean   :0.4425   Mean   : 37.56   Mean   : 4.047
##  3rd Qu.:   0.00   3rd Qu.:0.0000   3rd Qu.:  0.00   3rd Qu.: 4.364
##  Max.   :1040.33   Max.   :9.8667   Max.   :359.33   Max.   :42.091
##      ppt_a            perc_snow           T_g              s_m
##  Min.   : 0.00000   Min.   :0.0000   Min.   : 0.000   Min.   :0.00000
##  1st Qu.: 0.00000   1st Qu.:0.0500   1st Qu.: 0.000   1st Qu.:0.00000
##  Median : 0.00000   Median :1.0000   Median : 0.000   Median :0.00000
##  Mean   : 0.06945   Mean   :0.6661   Mean   : 3.163   Mean   :0.05374
##  3rd Qu.: 0.00000   3rd Qu.:1.0000   3rd Qu.: 4.092   3rd Qu.:0.11656
##  Max.   :16.33333   Max.   :1.0000   Max.   :19.282   Max.   :0.24837
```

```
head(weather_Snow_Soil_PPt_merged_cor)
```

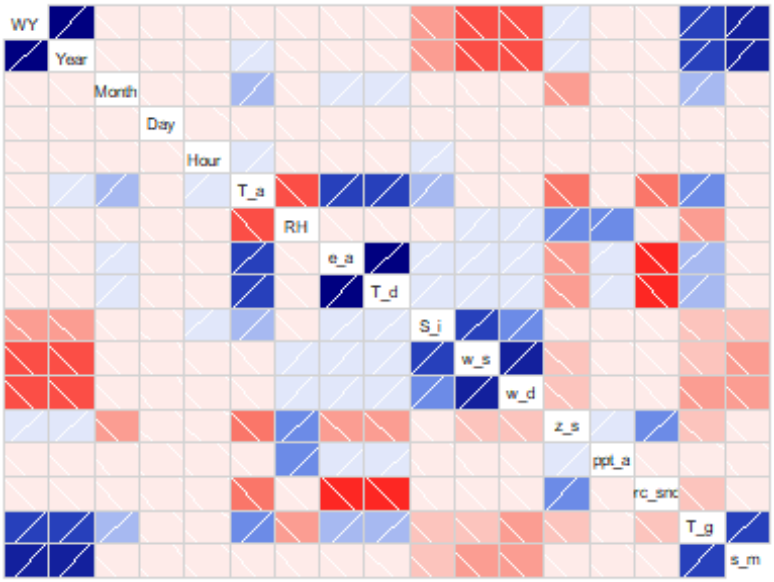
```
## # A tibble: 6 × 17
## # Groups:   WY, Year, Month, Day, Hour [6]
##      WY  Year Month   Day  Hour   T_a    RH   e_a     T_d   S_i   w_s
w_d
##   <int> <int> <int> <int> <int> <dbl> <dbl> <dbl>   <dbl> <dbl> <dbl>
<dbl>
## 1  2004  2003    10     1     0  16.4 0.29   536  -1.57       0  1.03 214.
## 2  2004  2003    10     1     1  16.1 0.303  548. -1.3        0  1.17 199.
## 3  2004  2003    10     1     2  14.9 0.333  561. -1.03       0  1    182.
## 4  2004  2003    10     1     3  14.4 0.357  578. -0.7        0  0.8
41.0
## 5  2004  2003    10     1     4  14.6 0.363  599. -0.233      0  1.07 256.
## 6  2004  2003    10     1     5  14.8 0.363  606. -0.0667     0  1.03 127.
## # … with 5 more variables: z_s <dbl>, ppt_a <dbl>, perc_snow <dbl>, T_g
<dbl>,
## #   s_m <dbl>
```

```
#Correlation Matrix
#(take all features except Minute because since it's only value is 0, it
shows NA in correlation with other variables which will disrupt correlation
plot later)

#Ignore standard deviation warning using suppressWarnings function
suppressWarnings({corr <- round(cor(weather_Snow_Soil_PPt_merged_cor), 1)})
corr
```

```
##             WY Year Month Day Hour  T_a   RH  e_a  T_d  S_i  w_s  w_d  z_s
## WY         1.0  1.0   0.0   0  0.0  0.0 -0.1  0.0 -0.1 -0.4 -0.6 -0.6  0.1
## Year       1.0  1.0  -0.1   0  0.0  0.1 -0.1  0.0  0.0 -0.3 -0.6 -0.6  0.1
## Month      0.0 -0.1   1.0   0  0.0  0.2 -0.1  0.1  0.1  0.0  0.0  0.0 -0.4
## Day        0.0  0.0   0.0   1  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
## Hour       0.0  0.0   0.0   0  1.0  0.1 -0.1  0.0  0.0  0.1  0.0  0.0  0.0
## T_a        0.0  0.1   0.2   0  0.1  1.0 -0.7  0.6  0.6  0.2  0.0  0.0 -0.5
## RH        -0.1 -0.1  -0.1   0 -0.1 -0.7  1.0  0.0  0.0 -0.1  0.1  0.1  0.3
## e_a        0.0  0.0   0.1   0  0.0  0.6  0.0  1.0  1.0  0.1  0.1  0.1 -0.3
## T_d       -0.1  0.0   0.1   0  0.0  0.6  0.0  1.0  1.0  0.1  0.1  0.1 -0.3
## S_i       -0.4 -0.3   0.0   0  0.1  0.2 -0.1  0.1  0.1  1.0  0.6  0.4 -0.1
## w_s       -0.6 -0.6   0.0   0  0.0  0.0  0.1  0.1  0.1  0.6  1.0  0.8 -0.2
## w_d       -0.6 -0.6   0.0   0  0.0  0.0  0.1  0.1  0.1  0.4  0.8  1.0 -0.2
## z_s        0.1  0.1  -0.4   0  0.0 -0.5  0.3 -0.3 -0.3 -0.1 -0.2 -0.2  1.0
## ppt_a      0.0  0.0   0.0   0  0.0 -0.1  0.3  0.1  0.1  0.0  0.0  0.0  0.1
## perc_snow  0.0  0.0  -0.1   0  0.0 -0.5  0.0 -0.8 -0.8 -0.1  0.0  0.0  0.3
## T_g        0.7  0.7   0.2   0  0.0  0.4 -0.3  0.2  0.2 -0.2 -0.2 -0.3 -0.2
## s_m        0.8  0.8  -0.1   0  0.0 -0.1  0.0  0.0 -0.1 -0.2 -0.3 -0.3  0.0
##            ppt_a perc_snow  T_g  s_m
## WY           0.0       0.0  0.7  0.8
## Year         0.0       0.0  0.7  0.8
## Month        0.0      -0.1  0.2 -0.1
## Day          0.0       0.0  0.0  0.0
## Hour         0.0       0.0  0.0  0.0
```

```
## T_a         -0.1       -0.5  0.4 -0.1
## RH           0.3        0.0 -0.3  0.0
## e_a          0.1       -0.8  0.2  0.0
## T_d          0.1       -0.8  0.2 -0.1
## S_i          0.0       -0.1 -0.2 -0.2
## w_s          0.0        0.0 -0.2 -0.3
## w_d          0.0        0.0 -0.3 -0.3
## z_s          0.1        0.3 -0.2  0.0
## ppt_a        1.0       -0.1  0.0  0.0
## perc_snow   -0.1        1.0 -0.2  0.0
## T_g          0.0       -0.2  1.0  0.6
## s_m          0.0        0.0  0.6  1.0
```

```
#install.packages('corrgram')
library(corrgram)

#Visualizing using a Correlogram
corrgram(corr)
```



#########Multiple Linear Regression on the combined cleaned correlation handled dataset#####################################

```
lm.fits1 <- lm(z_s ~. , data = weather_Snow_Soil_PPt_merged_cor)
summary(lm.fits1)
```

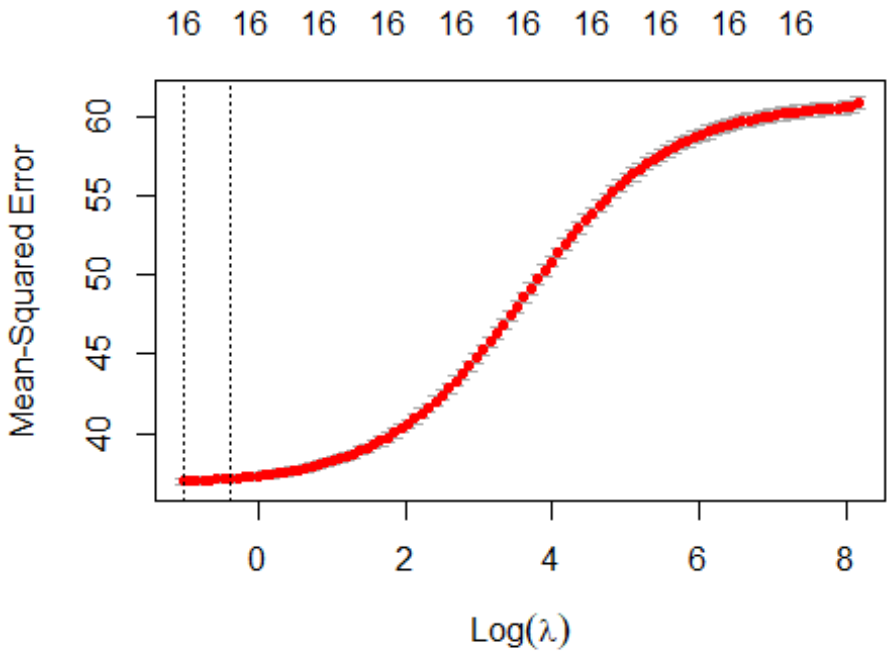
```
##
## Call:
## lm(formula = z_s ~ ., data = weather_Snow_Soil_PPt_merged_cor)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.3127  -3.7086  -0.9456   2.1263  29.9204
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.351e+03  2.843e+01 -47.528  < 2e-16 ***
## WY           3.868e+00  9.478e-02  40.810  < 2e-16 ***
## Year        -3.197e+00  9.503e-02 -33.646  < 2e-16 ***
## Month       -1.099e+00  1.175e-02 -93.541  < 2e-16 ***
## Day         -2.481e-03  2.207e-03  -1.124    0.261
## Hour         3.395e-02  2.848e-03  11.921  < 2e-16 ***
## T_a          4.889e-02  1.190e-02   4.107 4.01e-05 ***
## RH           1.058e+01  3.645e-01  29.023  < 2e-16 ***
## e_a          1.220e-02  4.708e-04  25.909  < 2e-16 ***
## T_d         -7.647e-01  2.391e-02 -31.985  < 2e-16 ***
## S_i          3.966e-03  1.828e-04  21.700  < 2e-16 ***
## w_s         -4.925e-01  3.533e-02 -13.939  < 2e-16 ***
## w_d         -1.005e-02  4.407e-04 -22.792  < 2e-16 ***
## ppt_a        2.326e-01  5.618e-02   4.141 3.46e-05 ***
## perc_snow    2.653e+00  7.997e-02  33.174  < 2e-16 ***
## T_g         -1.078e-01  6.382e-03 -16.888  < 2e-16 ***
## s_m         -2.428e+01  4.730e-01 -51.322  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.027 on 96416 degrees of freedom
## Multiple R-squared:  0.4032, Adjusted R-squared:  0.4031
## F-statistic:  4072 on 16 and 96416 DF,  p-value: < 2.2e-16

#Confidence Interval
confint(lm.fits1)

##                     2.5 %        97.5 %
## (Intercept) -1.407033e+03 -1.295581e+03
## WY           3.682171e+00  4.053704e+00
## Year        -3.383550e+00 -3.011045e+00
## Month       -1.122352e+00 -1.076283e+00
## Day         -6.805954e-03  1.844741e-03
## Hour         2.837041e-02  3.953512e-02
## T_a          2.555740e-02  7.221568e-02
## RH           9.864404e+00  1.129322e+01
## e_a          1.127535e-02  1.312093e-02
## T_d         -8.115169e-01 -7.178014e-01
## S_i          3.607698e-03  4.324109e-03
## w_s         -5.617329e-01 -4.232357e-01
## w_d         -1.090896e-02 -9.181300e-03
## ppt_a        1.225243e-01  3.427296e-01
## perc_snow    2.496199e+00  2.809679e+00
```

```
## T_g          -1.202905e-01 -9.527268e-02
## s_m          -2.520251e+01 -2.334834e+01

#Creating our own function for MSE and RMSE Calculations
MSE1 <- mean(lm.fits1$residuals^2)
RMSE1 <- sqrt(MSE1)

cat("Mean Square Error: ", MSE1)

## Mean Square Error:  36.31291

cat(", Root Mean Square Error: ", RMSE1)

## , Root Mean Square Error:  6.026019

#Compute Error Rate using RSE - Error Rate is RSE divided by mean of response
variable
error1 <- sigma(lm.fits1)/mean(weather_Snow_Soil_PPt_merged$z_s)
cat("\nError rate: ", error1)

##
## Error rate:  1.489233
```

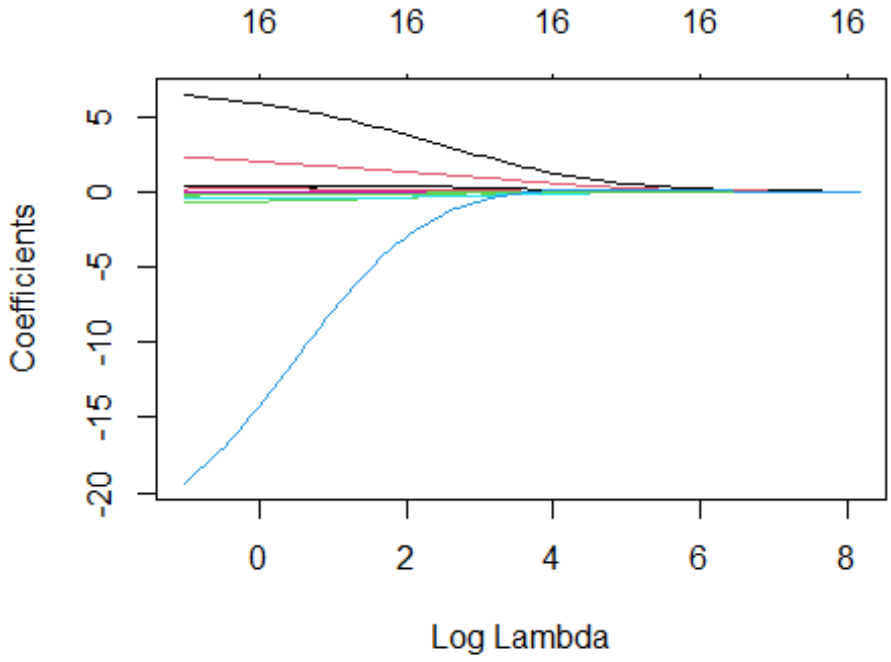
#########Ridge Regression on the combined cleaned correlation handled dataset#####################################

```
library(glmnet)
library(mgcv)
library(visreg)

#Ridge
#pass x matrix and y vector:
x <- model.matrix(z_s ~ ., data=weather_Snow_Soil_PPt_merged_cor )[, -1]
y <- weather_Snow_Soil_PPt_merged_cor$z_s

model <- glmnet(x, y, alpha = 0)

#find optimal lambda value
ridge.mod <- cv.glmnet(x, y, alpha = 0)
plot(ridge.mod)
```




```
min_lambda_ridge <- ridge.mod$lambda.min
cat("Minimum value of Lambda for ridge: ", min_lambda_ridge,"\n")

## Minimum value of Lambda for ridge:  0.3545392

ridge.mod2 <- glmnet(x, y, alpha = 0, lambda = min_lambda_ridge)

coef(ridge.mod2)

## 17 x 1 sparse Matrix of class "dgCMatrix"
##                        s0
## (Intercept) -1.165173e+03
## WY           4.125043e-01
## Year         1.696454e-01
## Month       -6.989822e-01
## Day         -2.529319e-03
## Hour         4.210481e-02
## T_a         -1.476095e-01
## RH           6.454802e+00
## e_a          2.302727e-03
## T_d         -2.462423e-01
## S_i          3.845244e-03
## w_s         -4.590228e-01
## w_d         -1.008087e-02
## ppt_a        3.880648e-01
## perc_snow    2.319382e+00
```

```
## T_g          -1.128515e-01
## s_m          -1.940122e+01

#produce Ridge trace plot
plot(model, xvar = "lambda")
```

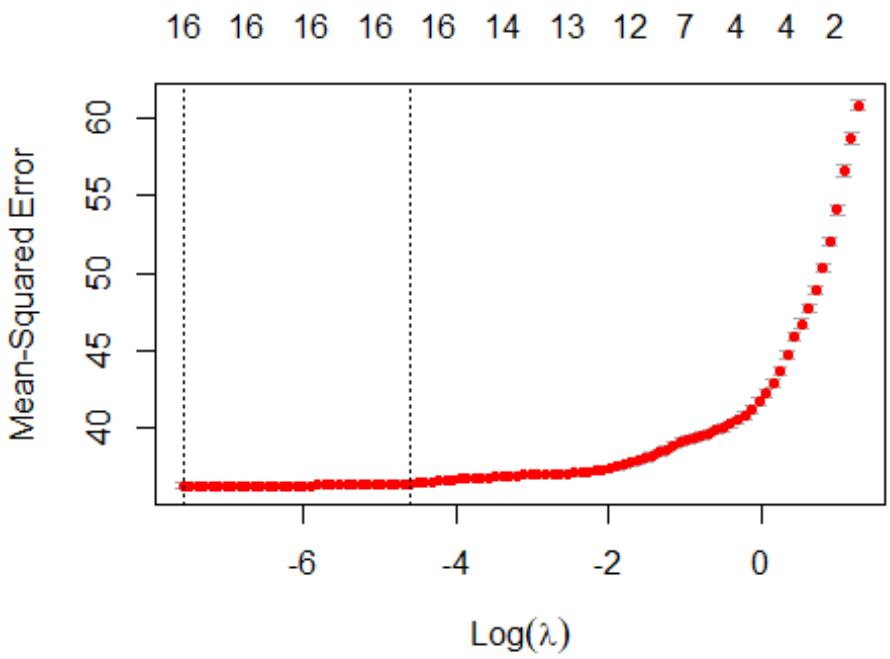



```
#use fitted best model to make predictions on train data

y_pred_ridge <- predict(ridge.mod2, s = min_lambda_ridge, newx=x)

mse_ridge <- mean((y - y_pred_ridge)^2)
rmse_ridge <- sqrt(mse_ridge)
RSS_ridge <- sum((y - y_pred_ridge)^2)
TSS_ridge <- (sum((y - mean(y))^2))
rsquared_ridge <- 1-(RSS_ridge/TSS_ridge)

cat("Mean Square Error Ridge: ", mse_ridge)

## Mean Square Error Ridge:  36.99648

cat("\nRoot Mean Square Error Ridge: ", rmse_ridge)

##
## Root Mean Square Error Ridge:  6.082473

cat("\nR^2 Ridge: ", rsquared_ridge)
```

```
##
## R^2 Ridge:  0.3920074
```

#########Lasso Regression on the combined cleaned correlation handled dataset#####################################

```
#Lasso

#pass x matrix and y vector:
x <- model.matrix(z_s ~ ., data=weather_Snow_Soil_PPt_merged_cor )[, -1]
y <- weather_Snow_Soil_PPt_merged_cor$z_s

lasso.mod <- cv.glmnet(x, y, alpha = 1)
#lasso.mod <- cv.glmnet(x, y, alpha = 1)

min_lambda_lasso <- lasso.mod$lambda.min
cat("Minimum value of Lambda: ", min_lambda_lasso,"\n")

## Minimum value of Lambda:  0.0005143757

#produce plot of test MSE by lambda value
plot(lasso.mod)
```




```
lasso.mod2 <- glmnet(x, y, alpha = 1, lambda = min_lambda_lasso)

coef(lasso.mod2)
```

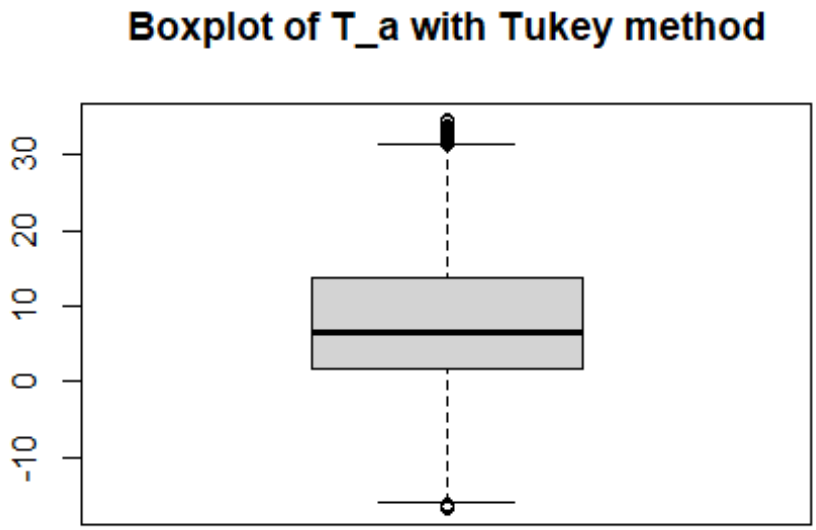
```
## 17 x 1 sparse Matrix of class "dgCMatrix"
##                          s0
## (Intercept) -1.348088e+03
## WY           3.593224e+00
## Year        -2.924061e+00
## Month       -1.069848e+00
## Day         -2.377104e-03
## Hour         3.475393e-02
## T_a          3.586687e-02
## RH           1.036357e+01
## e_a          1.193791e-02
## T_d         -7.465853e-01
## S_i          3.985671e-03
## w_s         -4.876813e-01
## w_d         -1.007540e-02
## ppt_a        2.396109e-01
## perc_snow    2.661558e+00
## T_g         -1.078209e-01
## s_m         -2.417660e+01
```

```
#use fitted best model to make predictions on train data

y_pred_lasso <- predict(lasso.mod2, s = min_lambda_lasso, newx=x)

mse_lasso <- mean((y - y_pred_lasso)^2)
rmse_lasso <- sqrt(mse_lasso)
RSS_lasso <- sum((y - y_pred_lasso)^2)
TSS_lasso <- (sum((y - mean(y))^2))
rsquared_lasso <- 1-(RSS_lasso/TSS_lasso)

cat("Mean Square Error Lasso: ", mse_lasso)

## Mean Square Error Lasso:  36.31616

cat("\n Root Mean Square Error Lasso: ", rmse_lasso)

##
##  Root Mean Square Error Lasso:  6.026289

cat("\n R^2 Lasso: ", rsquared_lasso)

##
##  R^2 Lasso:  0.4031876
```

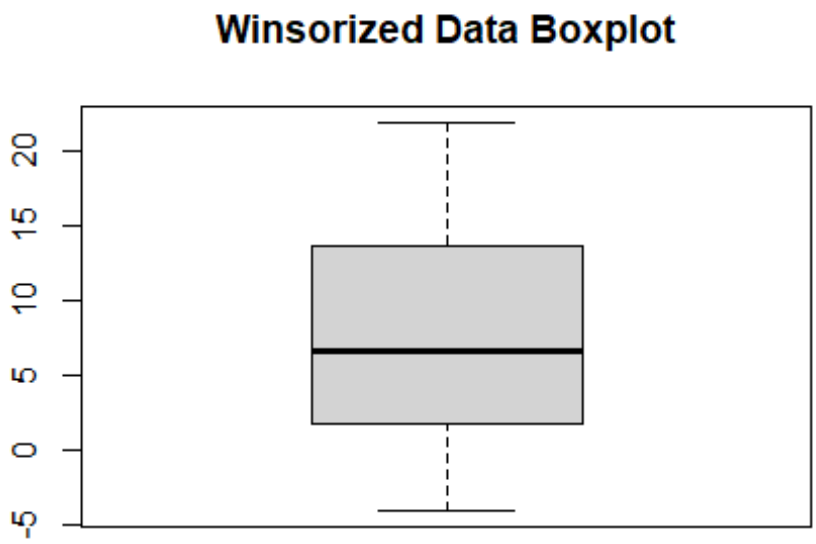
#Checking for outliers using Tukey method, if exists Handling the outliers via winsorizing

```
Q1 <- quantile(weather_Snow_Soil_PPt_merged$T_a, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$T_a, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR
```

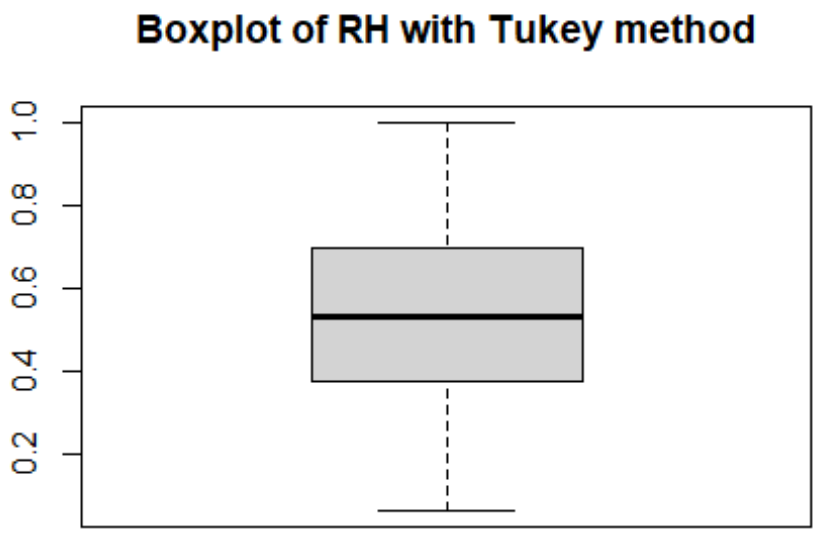
```r
#Locate and visualize the outlier on the boxplot:
outliers <- which(weather_Snow_Soil_PPt_merged$T_a < lower |
weather_Snow_Soil_PPt_merged$T_a > upper)

boxplot(weather_Snow_Soil_PPt_merged$T_a, main = "Boxplot of T_a with Tukey
method")
points(outliers, weather_Snow_Soil_PPt_merged$T_a[outliers], col = "red", pch
= 19)
```



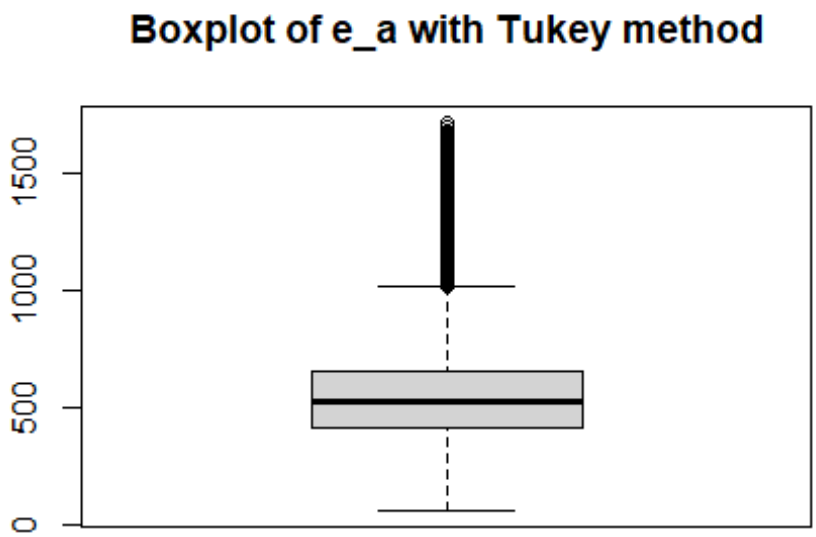

```r
#Using winsorizing technique to handle the extreme outlier values
library(DescTools)

## Warning: package 'DescTools' was built under R version 4.2.3

weather_Snow_Soil_PPt_merged$T_a <-
Winsorize(weather_Snow_Soil_PPt_merged$T_a, probs = c(0.05, 0.95))

boxplot(weather_Snow_Soil_PPt_merged$T_a, main = "Winsorized Data Boxplot")
```



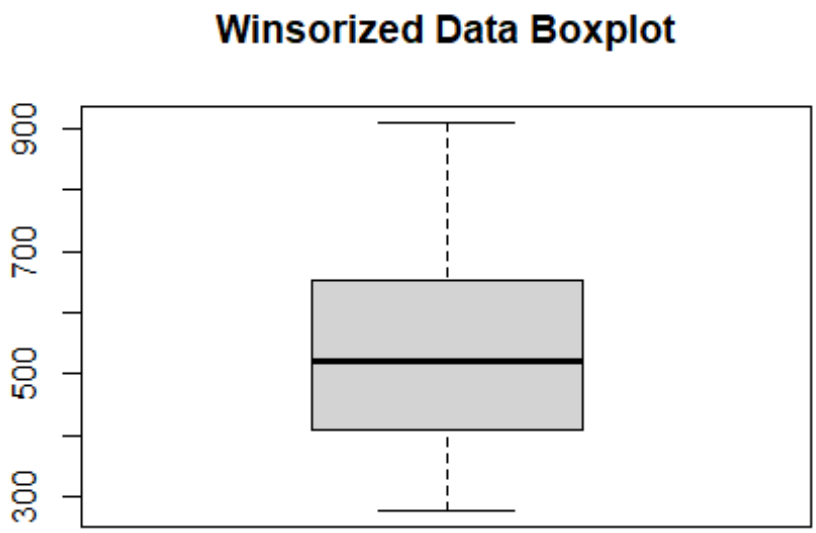

```
Q1 <- quantile(weather_Snow_Soil_PPt_merged$RH, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$RH, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$RH < lower |
weather_Snow_Soil_PPt_merged$RH > upper)

boxplot(weather_Snow_Soil_PPt_merged$RH, main = "Boxplot of RH with Tukey
method")
points(outliers, weather_Snow_Soil_PPt_merged$RH[outliers], col = "red", pch
= 19)
```



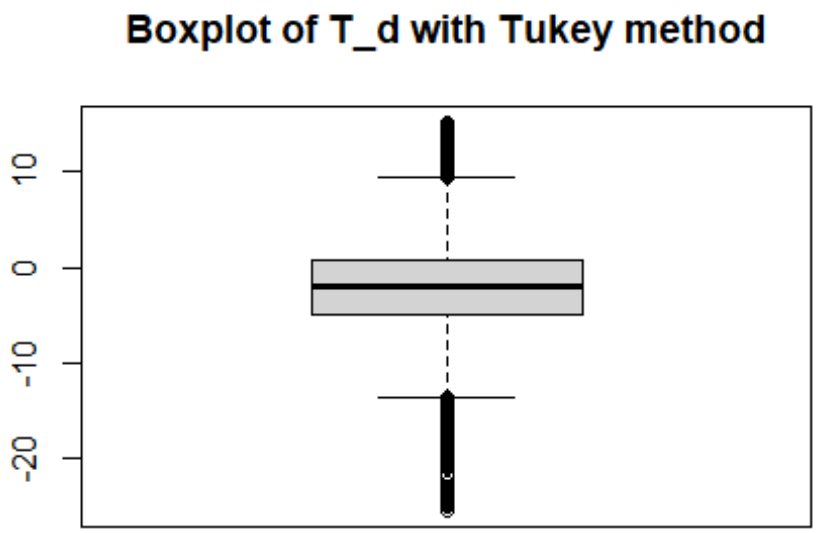

```r
Q1 <- quantile(weather_Snow_Soil_PPt_merged$e_a, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$e_a, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$e_a < lower |
weather_Snow_Soil_PPt_merged$e_a > upper)

boxplot(weather_Snow_Soil_PPt_merged$e_a, main = "Boxplot of e_a with Tukey
method")
points(outliers, weather_Snow_Soil_PPt_merged$e_a[outliers], col = "red", pch
= 19)
```



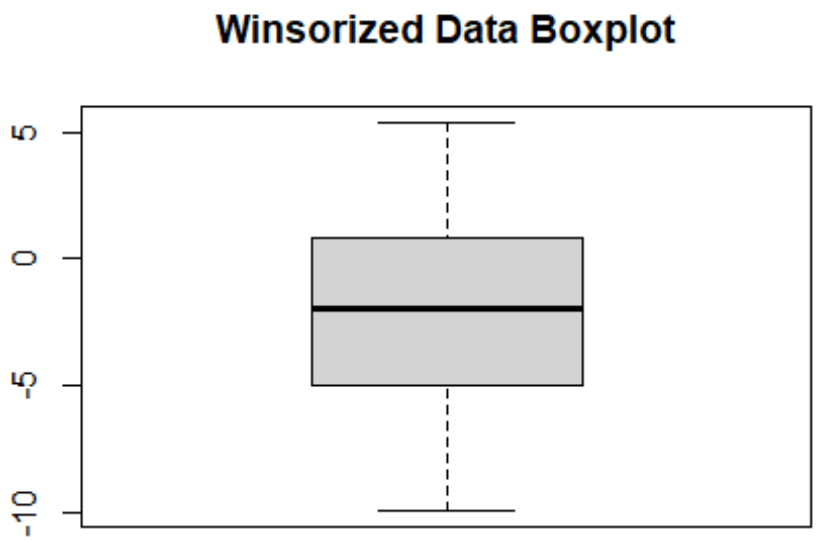

```
library(DescTools)
weather_Snow_Soil_PPt_merged$e_a <-
Winsorize(weather_Snow_Soil_PPt_merged$e_a, probs = c(0.05, 0.95))

boxplot(weather_Snow_Soil_PPt_merged$e_a, main = "Winsorized Data Boxplot")
```



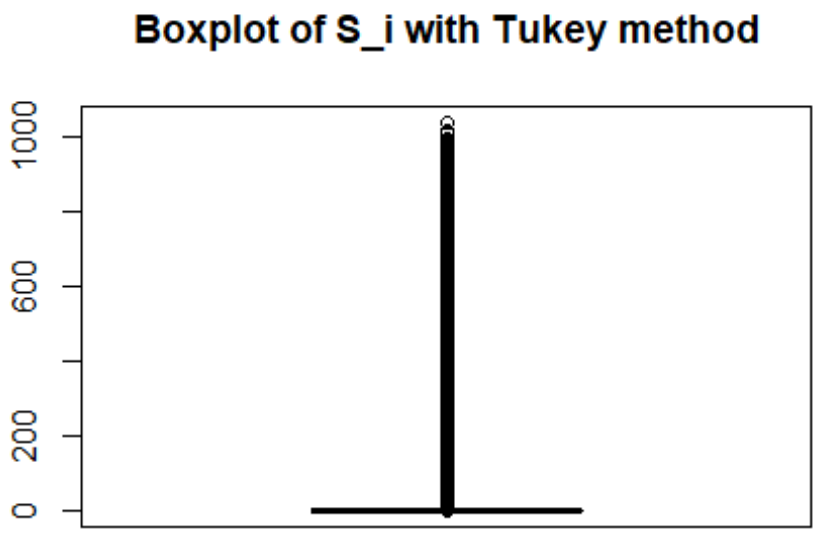

```r
Q1 <- quantile(weather_Snow_Soil_PPt_merged$T_d, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$T_d, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$T_d < lower |
weather_Snow_Soil_PPt_merged$T_d > upper)

boxplot(weather_Snow_Soil_PPt_merged$T_d, main = "Boxplot of T_d with Tukey
method")
points(outliers, weather_Snow_Soil_PPt_merged$T_d[outliers], col = "red", pch
= 19)
```



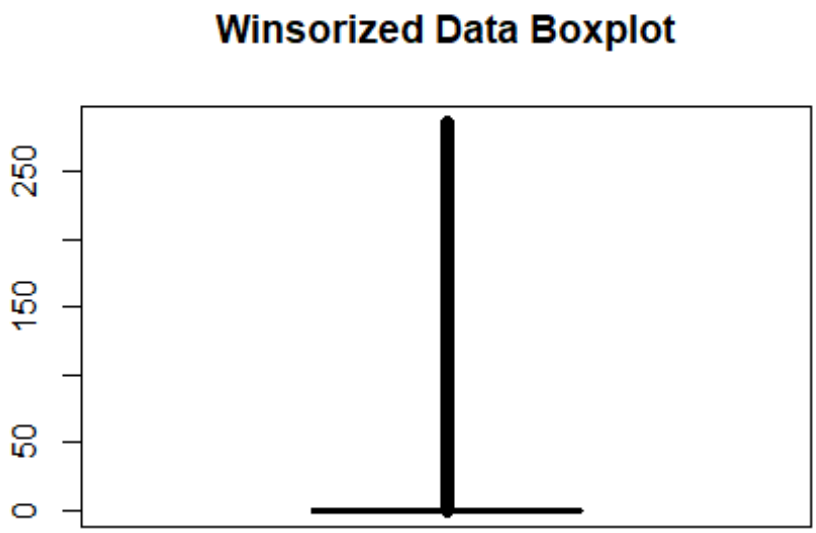

```
library(DescTools)
weather_Snow_Soil_PPt_merged$T_d <-
Winsorize(weather_Snow_Soil_PPt_merged$T_d, probs = c(0.05, 0.95))

boxplot(weather_Snow_Soil_PPt_merged$T_d, main = "Winsorized Data Boxplot")
```



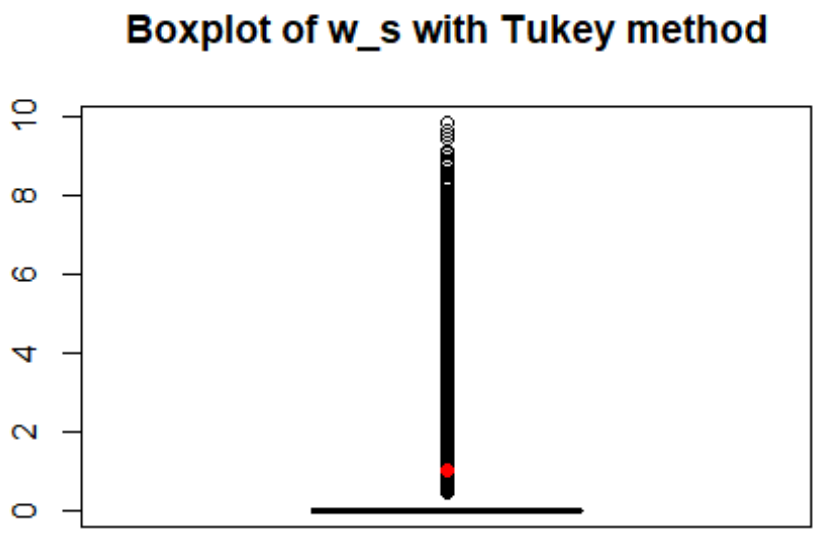

```
Q1 <- quantile(weather_Snow_Soil_PPt_merged$S_i, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$S_i, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$S_i < lower |
weather_Snow_Soil_PPt_merged$S_i > upper)

boxplot(weather_Snow_Soil_PPt_merged$S_i, main = "Boxplot of S_i with Tukey
method")
points(outliers, weather_Snow_Soil_PPt_merged$S_i[outliers], col = "red", pch
= 19)
```



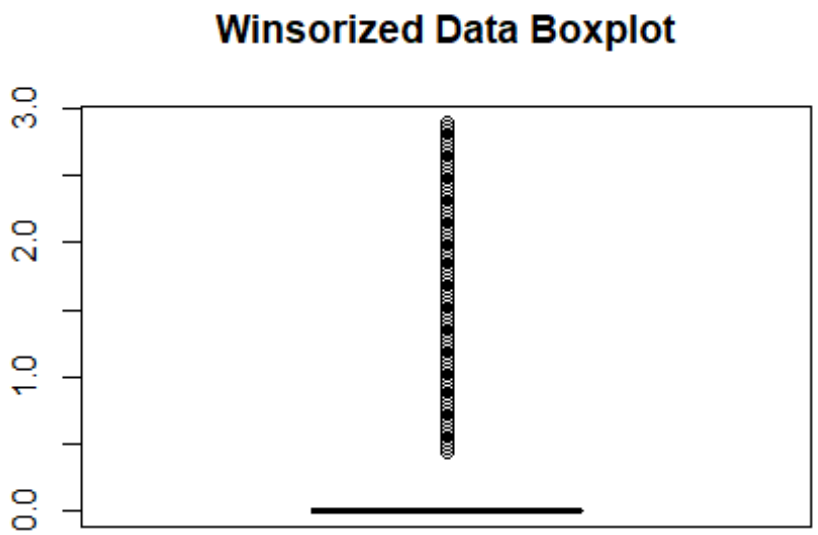

```
library(DescTools)
weather_Snow_Soil_PPt_merged$S_i <-
Winsorize(weather_Snow_Soil_PPt_merged$S_i, probs = c(0.05, 0.95))

boxplot(weather_Snow_Soil_PPt_merged$S_i, main = "Winsorized Data Boxplot")
```



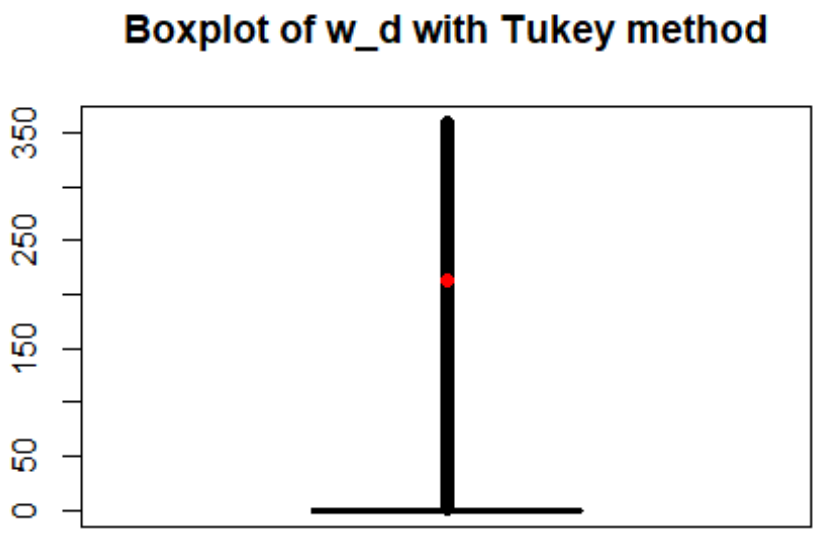

```
Q1 <- quantile(weather_Snow_Soil_PPt_merged$w_s, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$w_s, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$w_s < lower |
weather_Snow_Soil_PPt_merged$w_s > upper)

boxplot(weather_Snow_Soil_PPt_merged$w_s, main = "Boxplot of w_s with Tukey
method")
points(outliers, weather_Snow_Soil_PPt_merged$w_s[outliers], col = "red", pch
= 19)
```



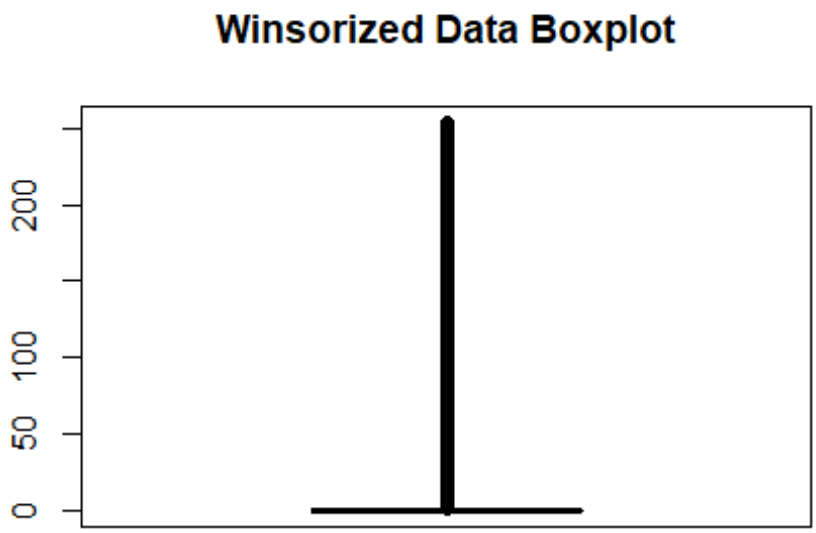

```r
#install.packages('DescTools')
#install.packages('lmom')
library(DescTools)
weather_Snow_Soil_PPt_merged$w_s <-
Winsorize(weather_Snow_Soil_PPt_merged$w_s, probs = c(0.05, 0.95))

boxplot(weather_Snow_Soil_PPt_merged$w_s, main = "Winsorized Data Boxplot")
```



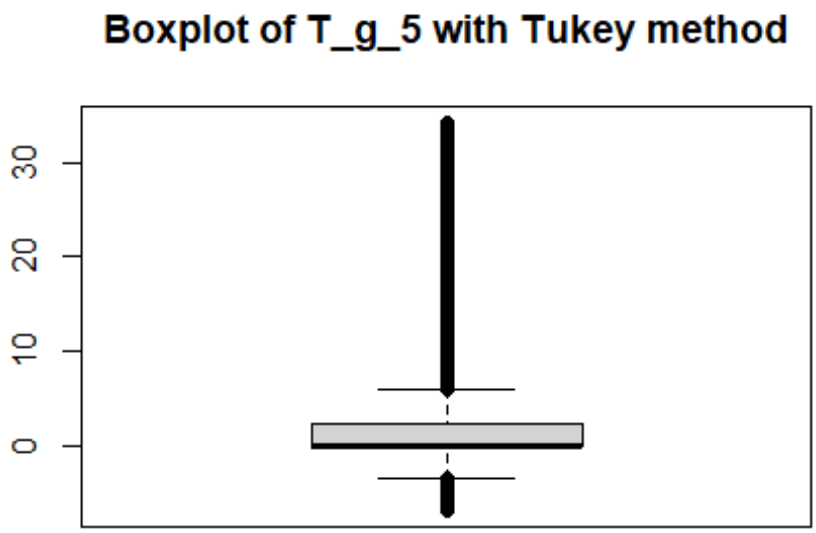

```r
Q1 <- quantile(weather_Snow_Soil_PPt_merged$w_d, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$w_d, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$w_d < lower |
weather_Snow_Soil_PPt_merged$w_d > upper)

boxplot(weather_Snow_Soil_PPt_merged$w_d, main = "Boxplot of w_d with Tukey
method")
points(outliers, weather_Snow_Soil_PPt_merged$w_d[outliers], col = "red", pch
= 19)
```



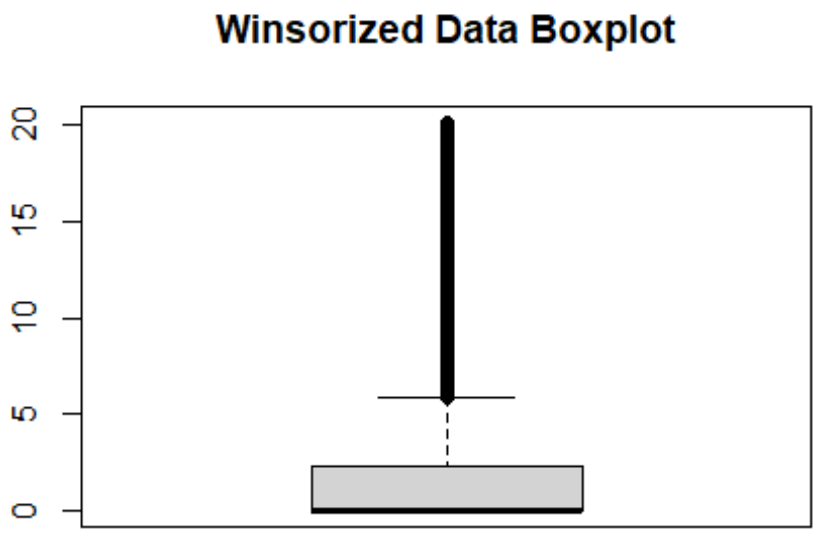

```
library(DescTools)
weather_Snow_Soil_PPt_merged$w_d <-
Winsorize(weather_Snow_Soil_PPt_merged$w_d, probs = c(0.05, 0.95))

boxplot(weather_Snow_Soil_PPt_merged$w_d, main = "Winsorized Data Boxplot")
```



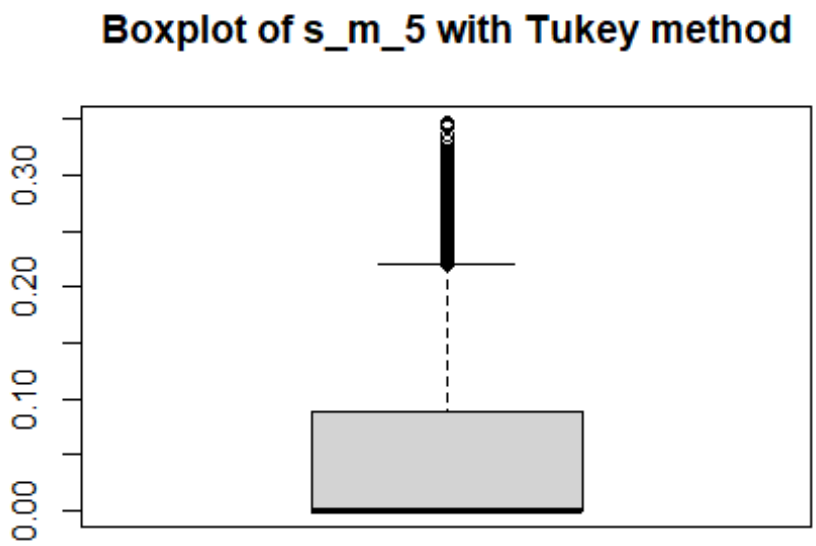

```
Q1 <- quantile(weather_Snow_Soil_PPt_merged$T_g_5, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$T_g_5, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$T_g_5 < lower |
weather_Snow_Soil_PPt_merged$T_g_5 > upper)

boxplot(weather_Snow_Soil_PPt_merged$T_g_5, main = "Boxplot of T_g_5 with
Tukey method")
points(outliers, weather_Snow_Soil_PPt_merged$T_g_5[outliers], col = "red",
pch = 19)
```



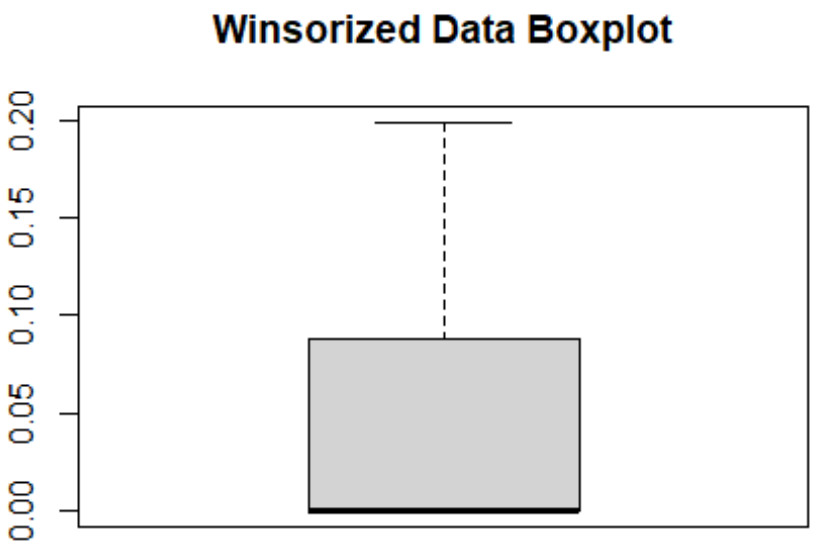

```
library(DescTools)
weather_Snow_Soil_PPt_merged$T_g_5 <-
Winsorize(weather_Snow_Soil_PPt_merged$T_g_5, probs = c(0.05, 0.95))

boxplot(weather_Snow_Soil_PPt_merged$T_g_5, main = "Winsorized Data Boxplot")
```



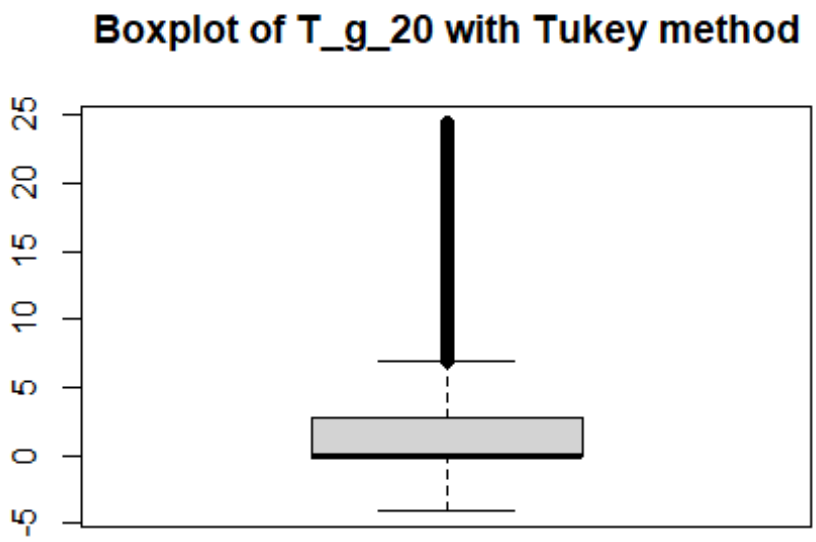

```r
Q1 <- quantile(weather_Snow_Soil_PPt_merged$s_m_5, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$s_m_5, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$s_m_5 < lower |
weather_Snow_Soil_PPt_merged$s_m_5 > upper)

boxplot(weather_Snow_Soil_PPt_merged$s_m_5, main = "Boxplot of s_m_5 with
Tukey method")
points(outliers, weather_Snow_Soil_PPt_merged$s_m_5[outliers], col = "red",
pch = 19)
```



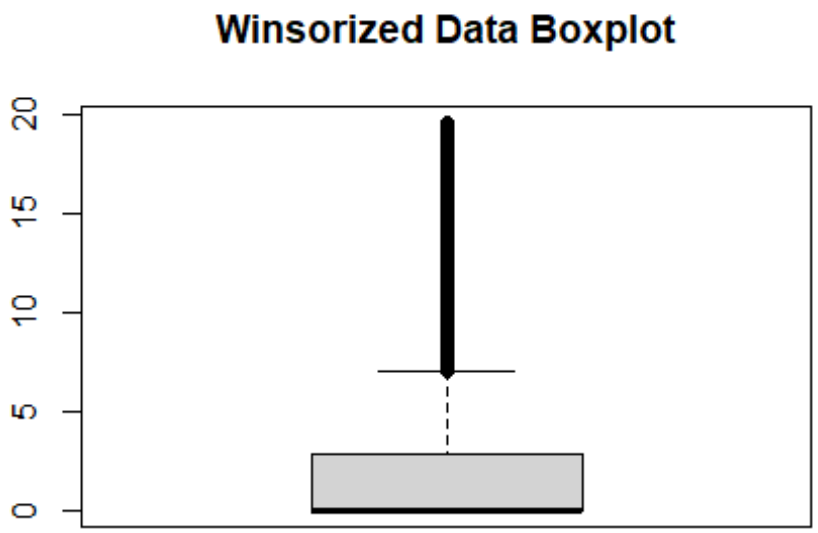

```
library(DescTools)
weather_Snow_Soil_PPt_merged$s_m_5 <-
Winsorize(weather_Snow_Soil_PPt_merged$s_m_5, probs = c(0.05, 0.95))

boxplot(weather_Snow_Soil_PPt_merged$s_m_5, main = "Winsorized Data Boxplot")
```



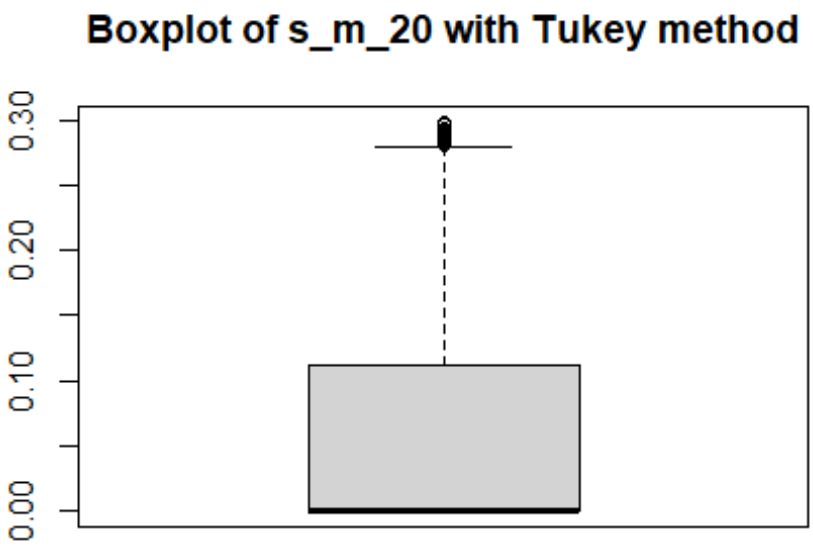

```r
Q1 <- quantile(weather_Snow_Soil_PPt_merged$T_g_20, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$T_g_20, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$T_g_20 < lower |
weather_Snow_Soil_PPt_merged$T_g_20 > upper)

boxplot(weather_Snow_Soil_PPt_merged$T_g_20, main = "Boxplot of T_g_20 with
Tukey method")
points(outliers, weather_Snow_Soil_PPt_merged$T_g_20[outliers], col = "red",
pch = 19)
```



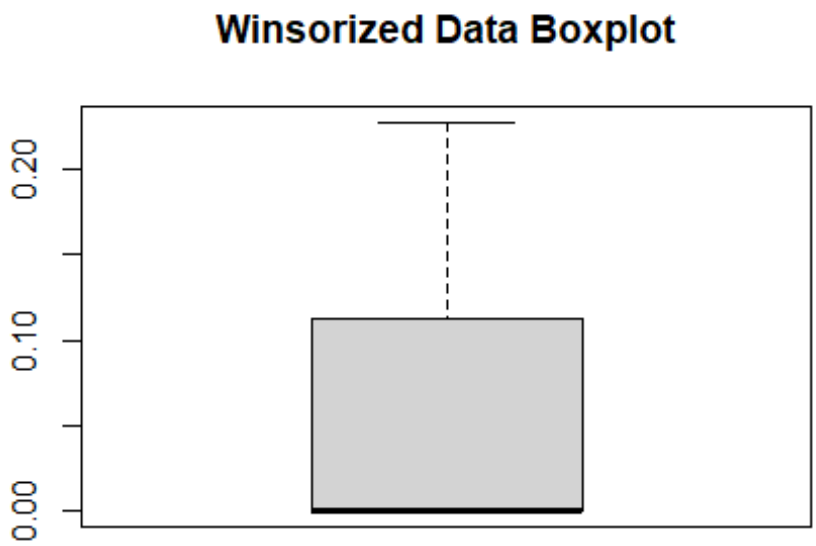

```
library(DescTools)
weather_Snow_Soil_PPt_merged$T_g_20 <-
Winsorize(weather_Snow_Soil_PPt_merged$T_g_20, probs = c(0.05, 0.95))

boxplot(weather_Snow_Soil_PPt_merged$T_g_20, main = "Winsorized Data
Boxplot")
```



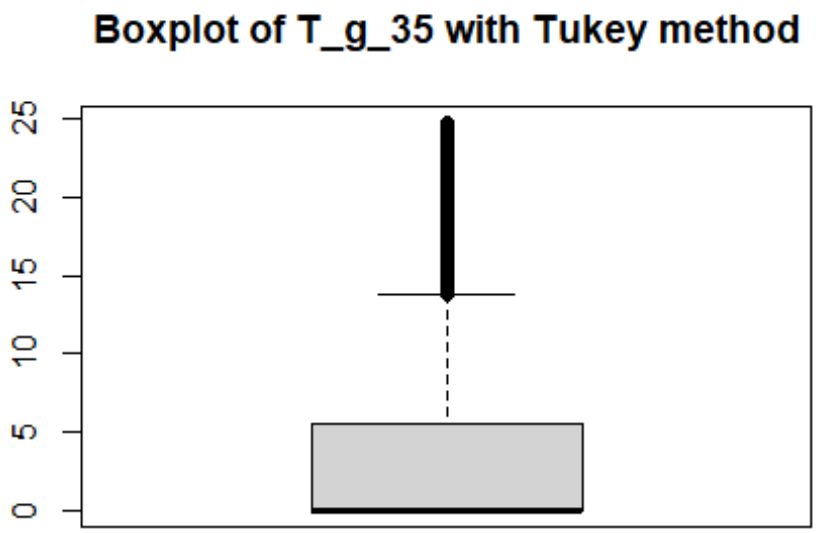

```
Q1 <- quantile(weather_Snow_Soil_PPt_merged$s_m_20, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$s_m_20, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$s_m_20 < lower |
weather_Snow_Soil_PPt_merged$s_m_20 > upper)

boxplot(weather_Snow_Soil_PPt_merged$s_m_20, main = "Boxplot of s_m_20 with
Tukey method")
points(outliers, weather_Snow_Soil_PPt_merged$s_m_20[outliers], col = "red",
pch = 19)
```



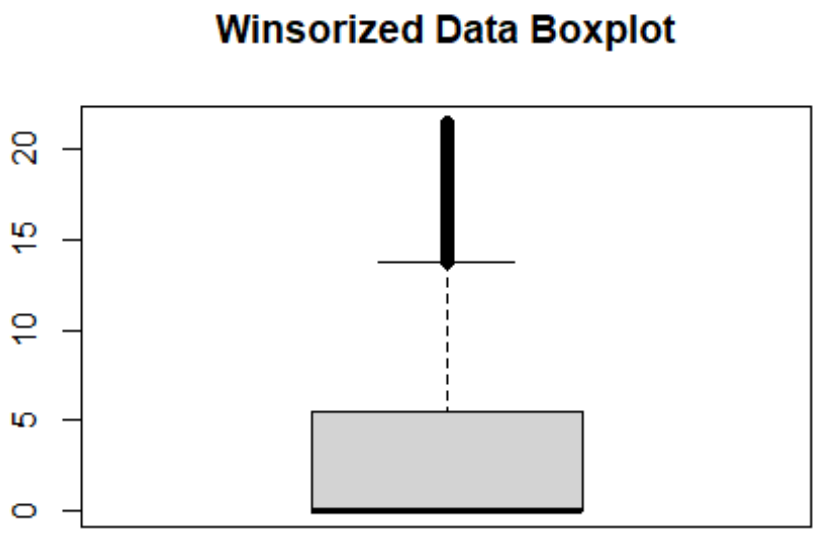

```
library(DescTools)
weather_Snow_Soil_PPt_merged$s_m_20 <-
Winsorize(weather_Snow_Soil_PPt_merged$s_m_20, probs = c(0.05, 0.95))

boxplot(weather_Snow_Soil_PPt_merged$s_m_20, main = "Winsorized Data
Boxplot")
```



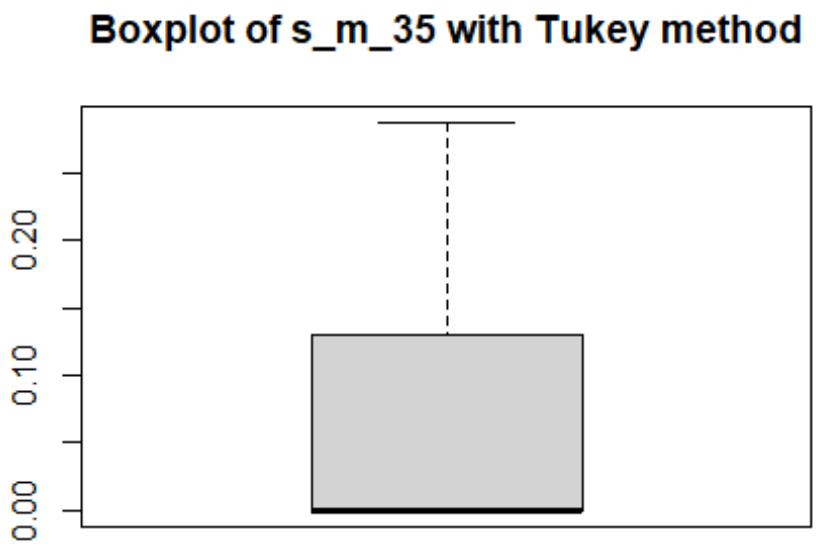

```r
Q1 <- quantile(weather_Snow_Soil_PPt_merged$T_g_35, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$T_g_35, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$T_g_35 < lower |
weather_Snow_Soil_PPt_merged$T_g_35 > upper)

boxplot(weather_Snow_Soil_PPt_merged$T_g_35, main = "Boxplot of T_g_35 with
Tukey method")
points(outliers, weather_Snow_Soil_PPt_merged$T_g_35[outliers], col = "red",
pch = 19)
```



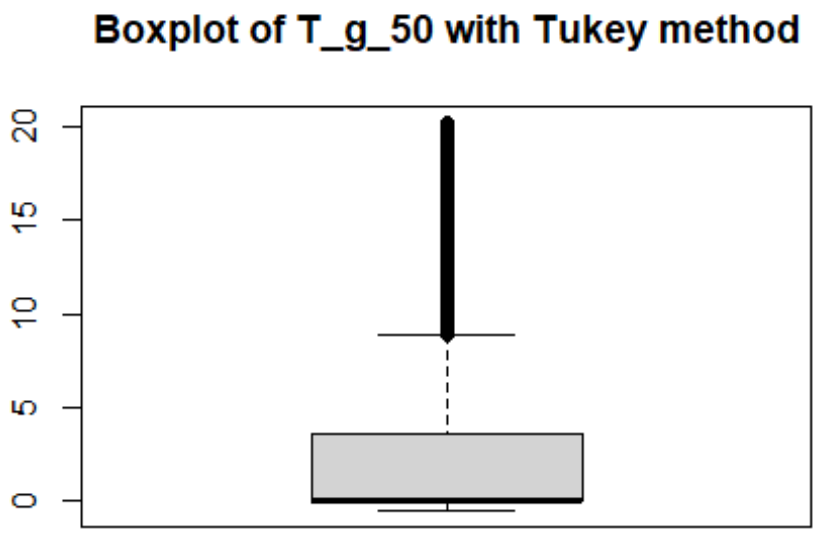

```
library(DescTools)
weather_Snow_Soil_PPt_merged$T_g_35 <-
Winsorize(weather_Snow_Soil_PPt_merged$T_g_35, probs = c(0.05, 0.95))

boxplot(weather_Snow_Soil_PPt_merged$T_g_35, main = "Winsorized Data
Boxplot")
```



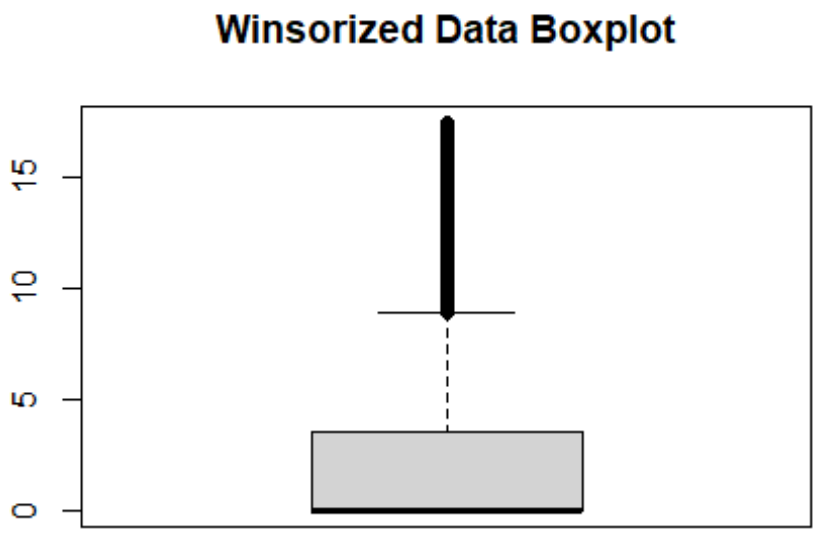

```r
Q1 <- quantile(weather_Snow_Soil_PPt_merged$s_m_35, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$s_m_35, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$s_m_35 < lower |
weather_Snow_Soil_PPt_merged$s_m_35 > upper)

boxplot(weather_Snow_Soil_PPt_merged$s_m_35, main = "Boxplot of s_m_35 with
Tukey method")
points(outliers, weather_Snow_Soil_PPt_merged$s_m_35[outliers], col = "red",
pch = 19)
```



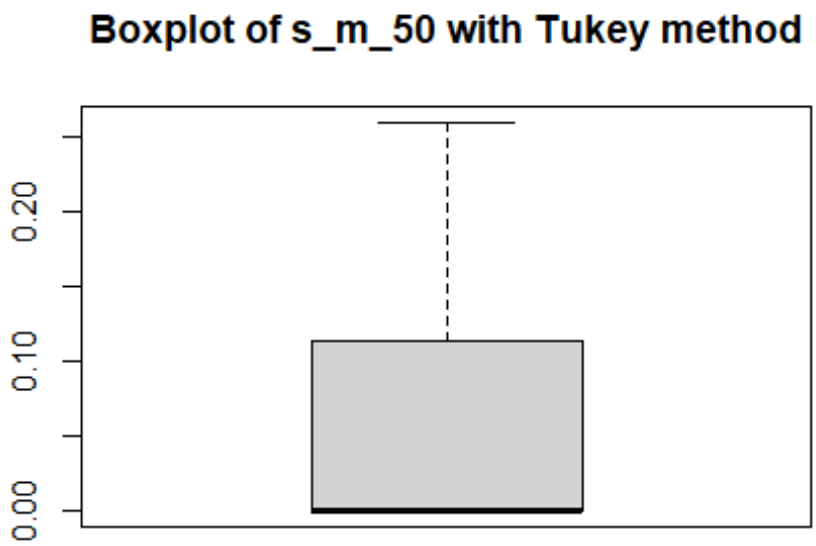

```r
#No outliers

Q1 <- quantile(weather_Snow_Soil_PPt_merged$T_g_50, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$T_g_50, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$T_g_50 < lower |
weather_Snow_Soil_PPt_merged$T_g_50 > upper)

boxplot(weather_Snow_Soil_PPt_merged$T_g_50, main = "Boxplot of T_g_50 with
Tukey method")
points(outliers, weather_Snow_Soil_PPt_merged$T_g_50[outliers], col = "red",
pch = 19)
```



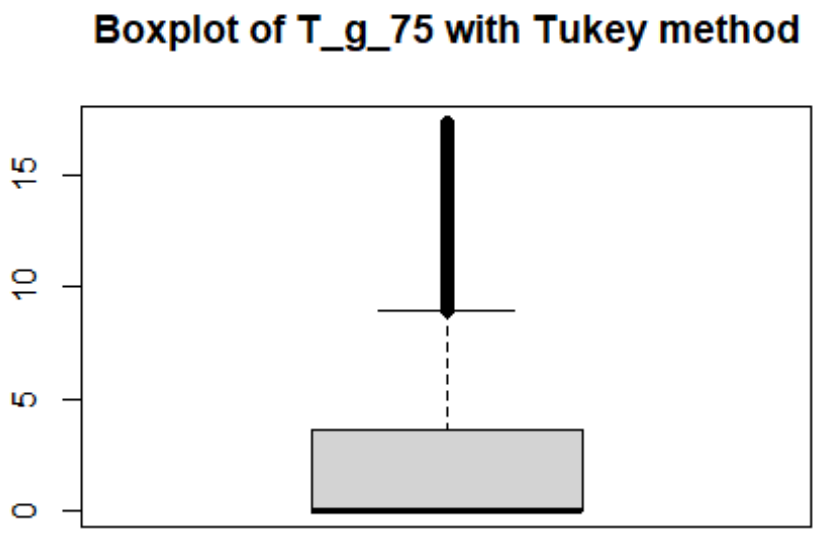

```
library(DescTools)
weather_Snow_Soil_PPt_merged$T_g_50 <-
Winsorize(weather_Snow_Soil_PPt_merged$T_g_50, probs = c(0.05, 0.95))

boxplot(weather_Snow_Soil_PPt_merged$T_g_50, main = "Winsorized Data
Boxplot")
```



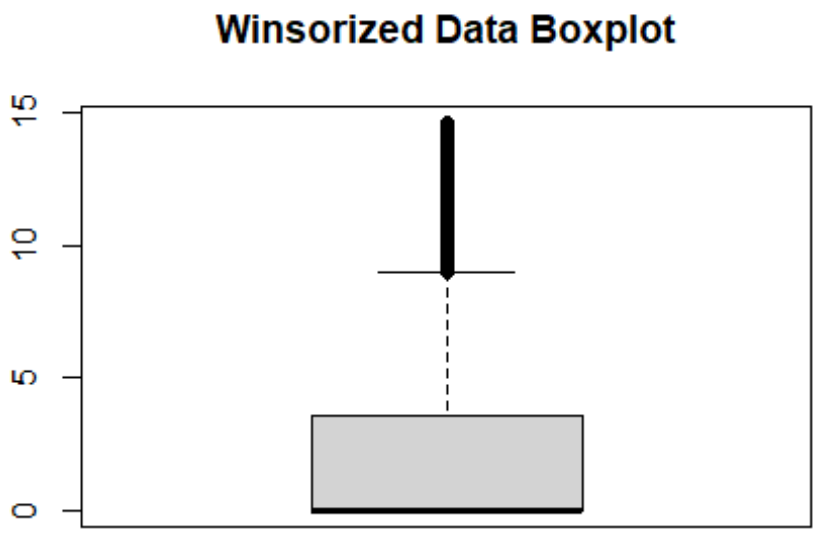

```r
Q1 <- quantile(weather_Snow_Soil_PPt_merged$s_m_50, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$s_m_50, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$s_m_50 < lower |
weather_Snow_Soil_PPt_merged$s_m_50 > upper)

boxplot(weather_Snow_Soil_PPt_merged$s_m_50, main = "Boxplot of s_m_50 with
Tukey method")
points(outliers, weather_Snow_Soil_PPt_merged$s_m_50[outliers], col = "red",
pch = 19)
```



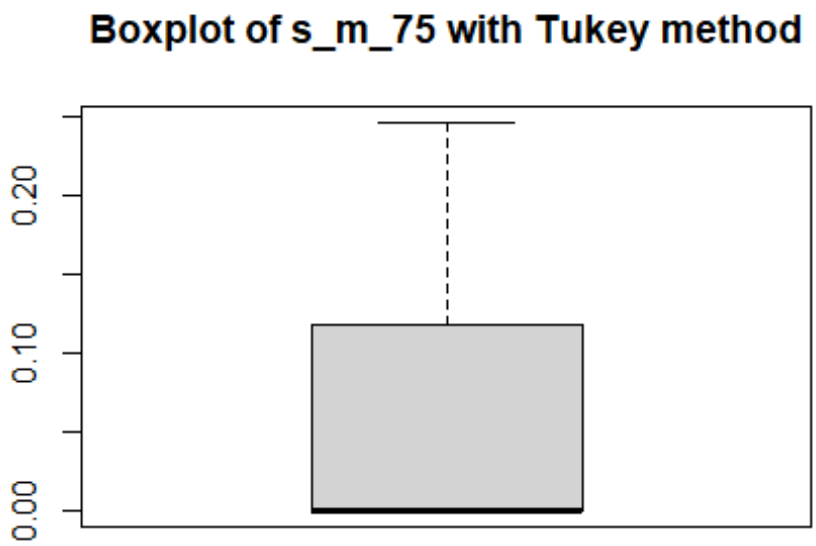

```r
#No Outliers

Q1 <- quantile(weather_Snow_Soil_PPt_merged$T_g_75, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$T_g_75, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$T_g_75 < lower |
weather_Snow_Soil_PPt_merged$T_g_75 > upper)

boxplot(weather_Snow_Soil_PPt_merged$T_g_75, main = "Boxplot of T_g_75 with
Tukey method")
points(outliers, weather_Snow_Soil_PPt_merged$T_g_75[outliers], col = "red",
pch = 19)
```



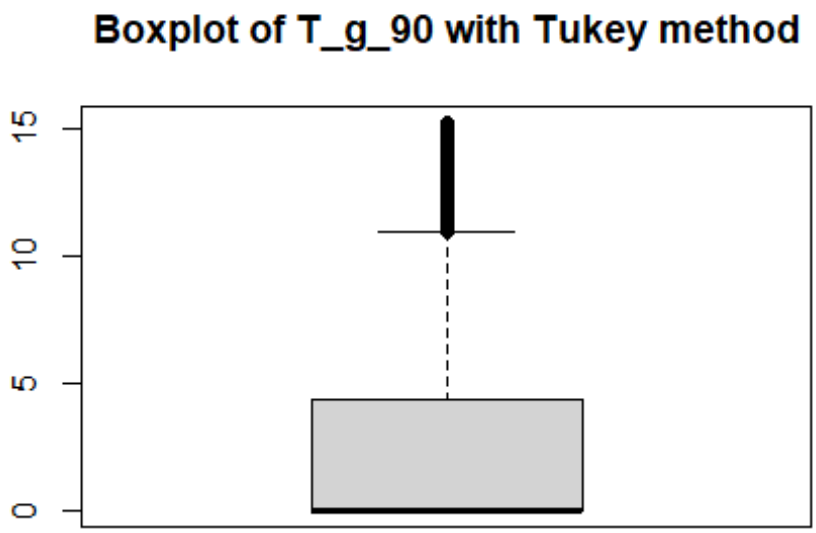

```
library(DescTools)
weather_Snow_Soil_PPt_merged$T_g_75 <-
Winsorize(weather_Snow_Soil_PPt_merged$T_g_75, probs = c(0.05, 0.95))

boxplot(weather_Snow_Soil_PPt_merged$T_g_75, main = "Winsorized Data
Boxplot")
```



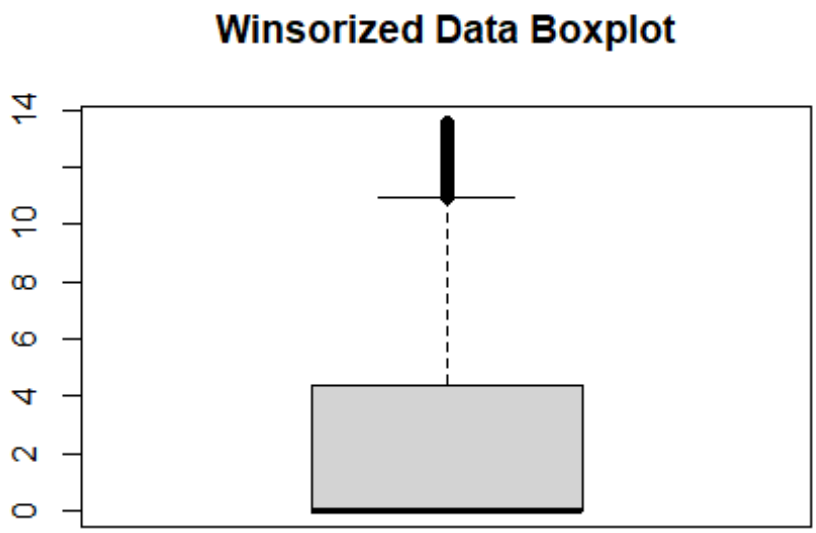

```r
Q1 <- quantile(weather_Snow_Soil_PPt_merged$s_m_75, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$s_m_75, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$s_m_75 < lower |
weather_Snow_Soil_PPt_merged$s_m_75 > upper)

boxplot(weather_Snow_Soil_PPt_merged$s_m_75, main = "Boxplot of s_m_75 with
Tukey method")
points(outliers, weather_Snow_Soil_PPt_merged$s_m_75[outliers], col = "red",
pch = 19)
```



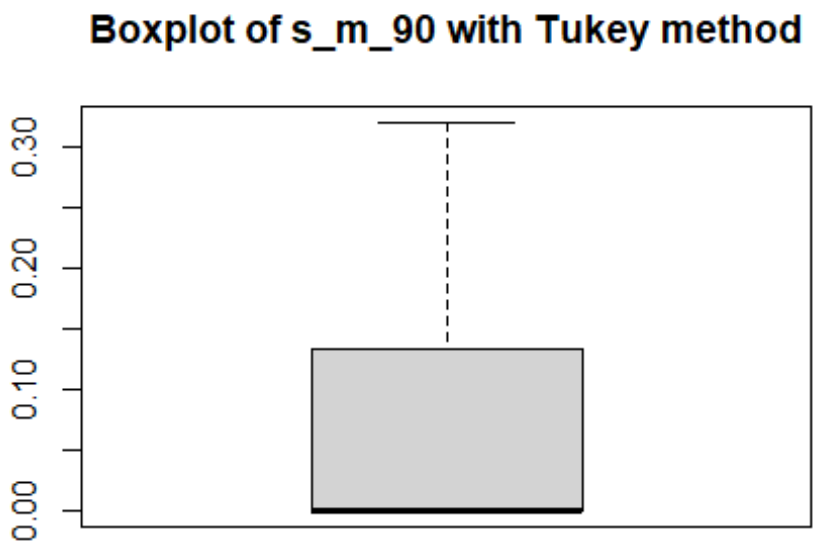

```r
#No Outliers

Q1 <- quantile(weather_Snow_Soil_PPt_merged$T_g_90, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$T_g_90, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$T_g_90 < lower |
weather_Snow_Soil_PPt_merged$T_g_90 > upper)

boxplot(weather_Snow_Soil_PPt_merged$T_g_90, main = "Boxplot of T_g_90 with
Tukey method")
points(outliers, weather_Snow_Soil_PPt_merged$T_g_90[outliers], col = "red",
pch = 19)
```



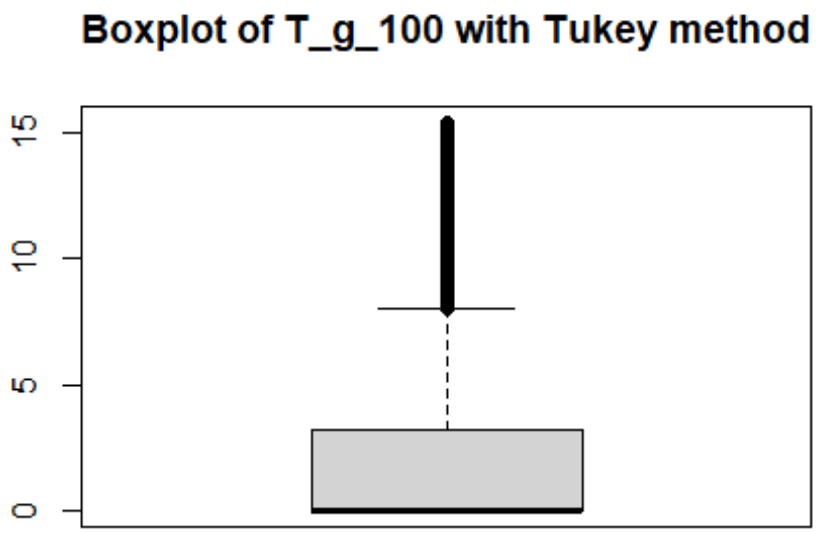

```
library(DescTools)
weather_Snow_Soil_PPt_merged$T_g_90 <-
Winsorize(weather_Snow_Soil_PPt_merged$T_g_90, probs = c(0.05, 0.95))

boxplot(weather_Snow_Soil_PPt_merged$T_g_90, main = "Winsorized Data
Boxplot")
```



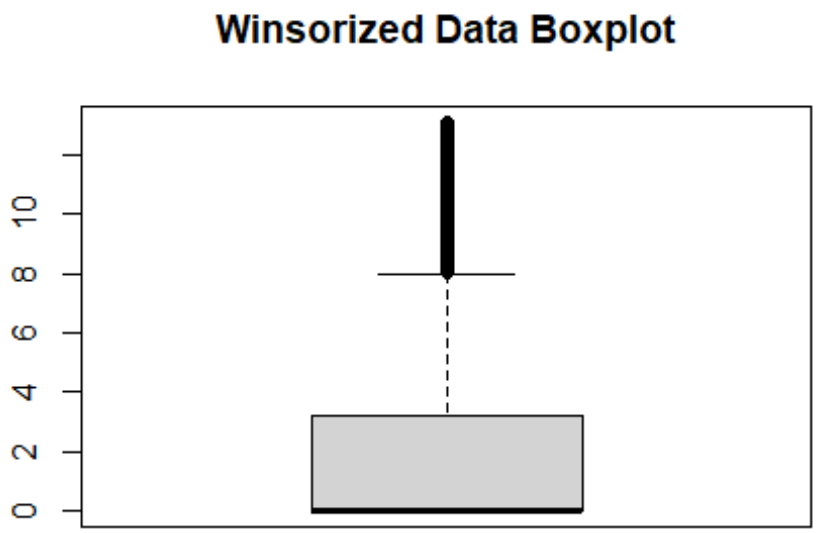

```r
Q1 <- quantile(weather_Snow_Soil_PPt_merged$s_m_90, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$s_m_90, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$s_m_90 < lower |
weather_Snow_Soil_PPt_merged$s_m_90 > upper)

boxplot(weather_Snow_Soil_PPt_merged$s_m_90, main = "Boxplot of s_m_90 with
Tukey method")
points(outliers, weather_Snow_Soil_PPt_merged$s_m_90[outliers], col = "red",
pch = 19)
```



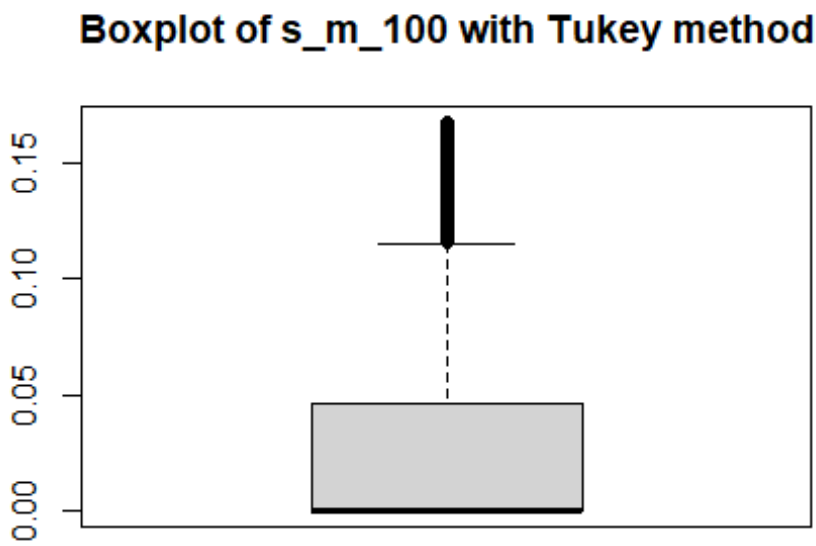

```r
#No Outliers

Q1 <- quantile(weather_Snow_Soil_PPt_merged$T_g_100, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$T_g_100, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$T_g_100 < lower |
weather_Snow_Soil_PPt_merged$T_g_100 > upper)

boxplot(weather_Snow_Soil_PPt_merged$T_g_100, main = "Boxplot of T_g_100 with
Tukey method")
points(outliers, weather_Snow_Soil_PPt_merged$T_g_100[outliers], col = "red",
pch = 19)
```



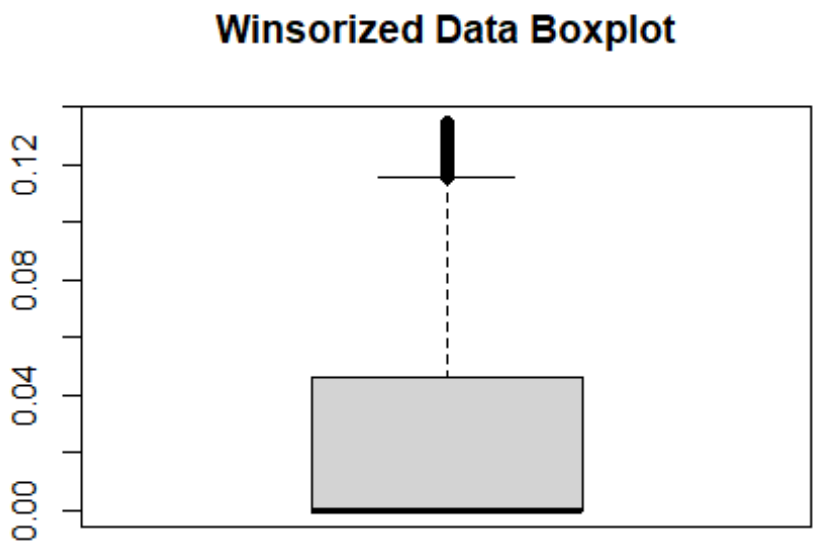

```
library(DescTools)
weather_Snow_Soil_PPt_merged$T_g_100 <-
Winsorize(weather_Snow_Soil_PPt_merged$T_g_100, probs = c(0.05, 0.95))

boxplot(weather_Snow_Soil_PPt_merged$T_g_100, main = "Winsorized Data
Boxplot")
```



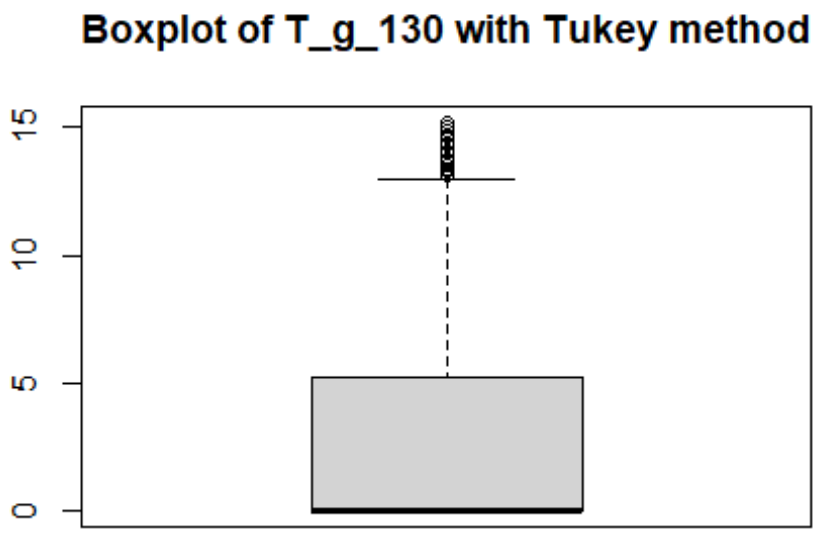

```
Q1 <- quantile(weather_Snow_Soil_PPt_merged$s_m_100, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$s_m_100, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$s_m_100 < lower |
weather_Snow_Soil_PPt_merged$s_m_100 > upper)

boxplot(weather_Snow_Soil_PPt_merged$s_m_100, main = "Boxplot of s_m_100 with
Tukey method")
points(outliers, weather_Snow_Soil_PPt_merged$s_m_100[outliers], col = "red",
pch = 19)
```



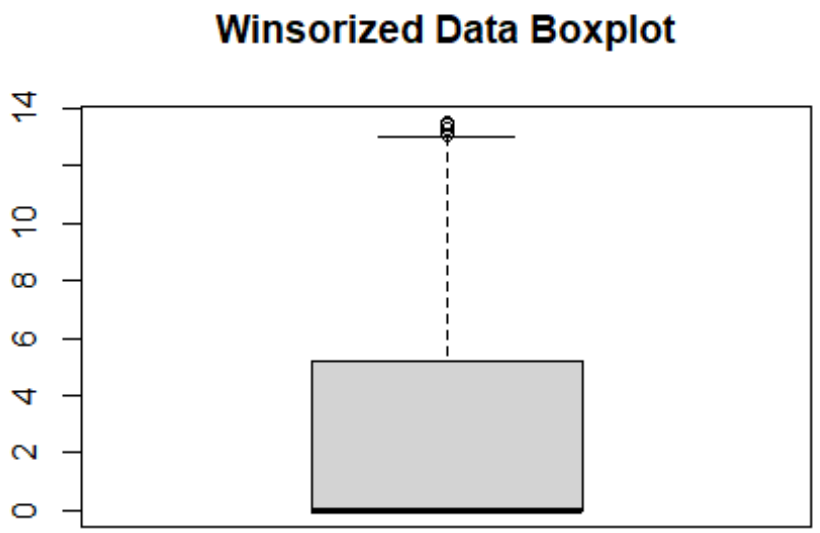

```
library(DescTools)
weather_Snow_Soil_PPt_merged$s_m_100 <-
Winsorize(weather_Snow_Soil_PPt_merged$s_m_100, probs = c(0.05, 0.95))

boxplot(weather_Snow_Soil_PPt_merged$s_m_100, main = "Winsorized Data
Boxplot")
```



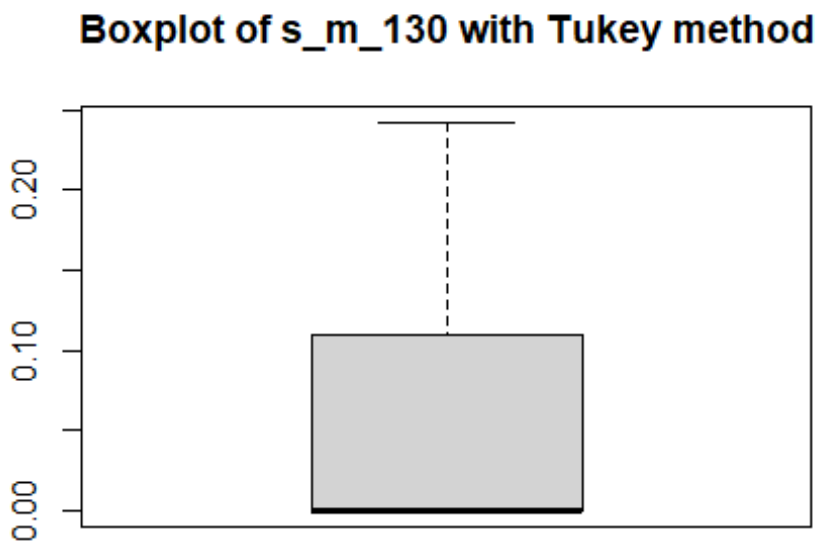

```r
Q1 <- quantile(weather_Snow_Soil_PPt_merged$T_g_130, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$T_g_130, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$T_g_130 < lower |
weather_Snow_Soil_PPt_merged$T_g_130 > upper)

boxplot(weather_Snow_Soil_PPt_merged$T_g_130, main = "Boxplot of T_g_130 with
Tukey method")
points(outliers, weather_Snow_Soil_PPt_merged$T_g_130[outliers], col = "red",
pch = 19)
```



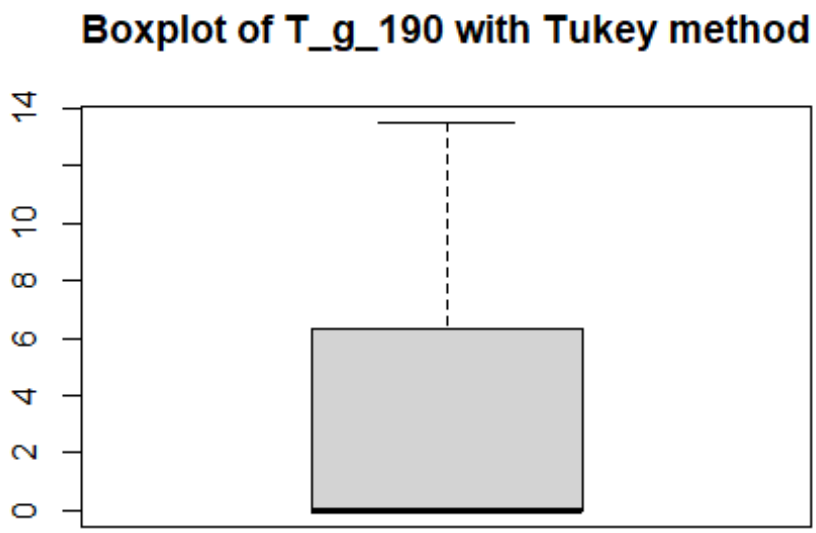

```
library(DescTools)
weather_Snow_Soil_PPt_merged$T_g_130 <-
Winsorize(weather_Snow_Soil_PPt_merged$T_g_130, probs = c(0.05, 0.95))

boxplot(weather_Snow_Soil_PPt_merged$T_g_130, main = "Winsorized Data
Boxplot")
```



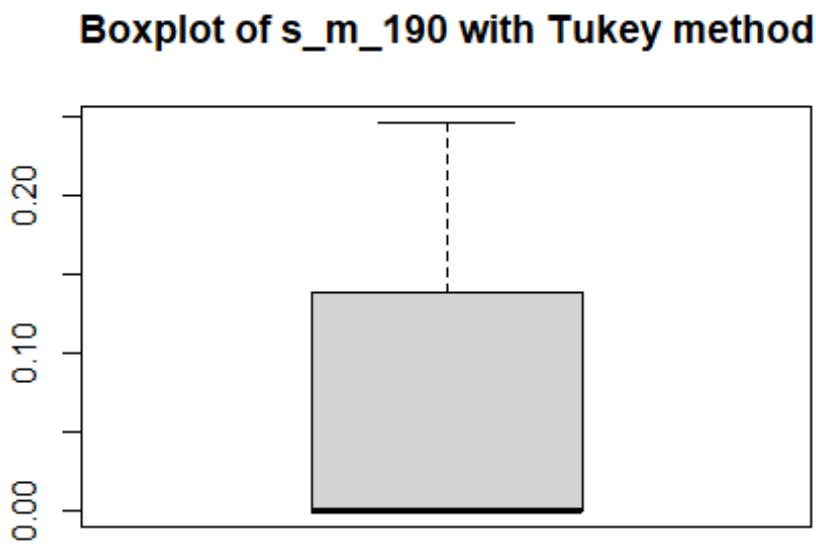

```
Q1 <- quantile(weather_Snow_Soil_PPt_merged$s_m_130, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$s_m_130, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$s_m_130 < lower |
weather_Snow_Soil_PPt_merged$s_m_130 > upper)

boxplot(weather_Snow_Soil_PPt_merged$s_m_130, main = "Boxplot of s_m_130 with
Tukey method")
points(outliers, weather_Snow_Soil_PPt_merged$s_m_130[outliers], col = "red",
pch = 19)
```



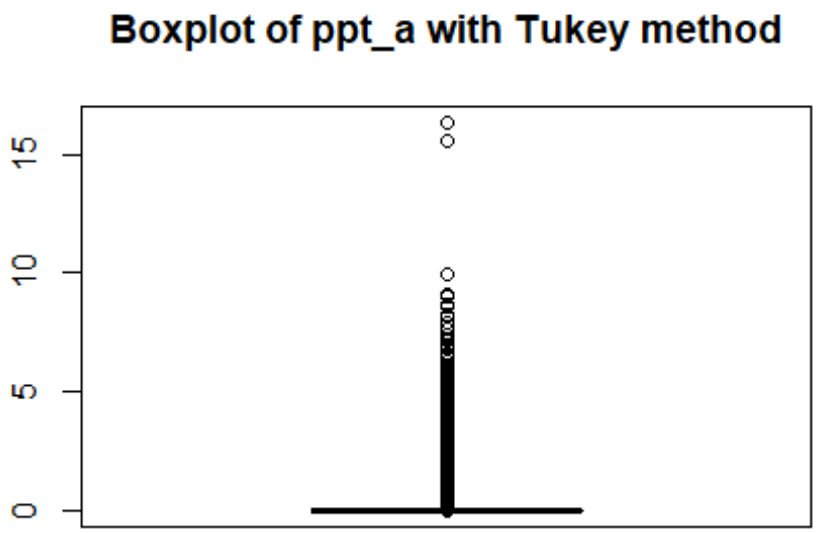

```r
#No Outliers

Q1 <- quantile(weather_Snow_Soil_PPt_merged$T_g_190, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$T_g_190, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$T_g_190 < lower |
weather_Snow_Soil_PPt_merged$T_g_190 > upper)

boxplot(weather_Snow_Soil_PPt_merged$T_g_190, main = "Boxplot of T_g_190 with
Tukey method")
points(outliers, weather_Snow_Soil_PPt_merged$T_g_190[outliers], col = "red",
pch = 19)
```



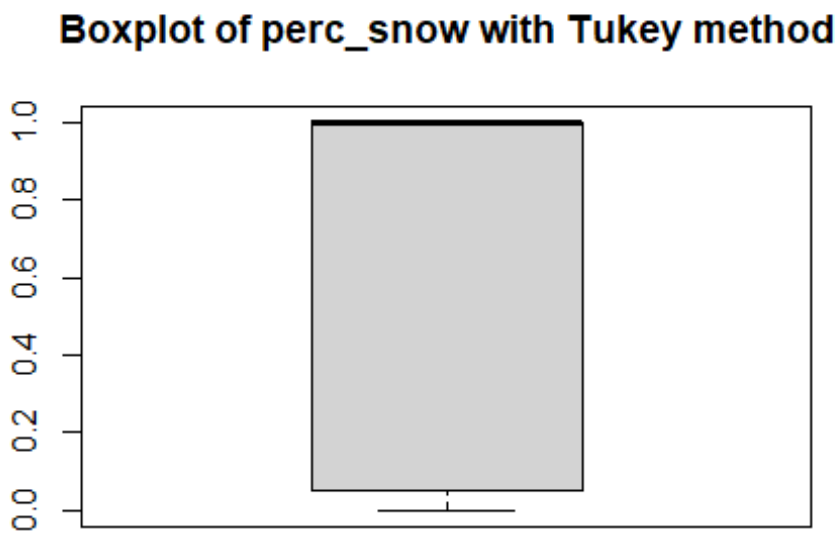

```r
#No Outliers

Q1 <- quantile(weather_Snow_Soil_PPt_merged$s_m_190, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$s_m_190, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$s_m_190 < lower |
weather_Snow_Soil_PPt_merged$s_m_190 > upper)

boxplot(weather_Snow_Soil_PPt_merged$s_m_190, main = "Boxplot of s_m_190 with
Tukey method")
points(outliers, weather_Snow_Soil_PPt_merged$s_m_190[outliers], col = "red",
pch = 19)
```




```r
#No Outliers

Q1 <- quantile(weather_Snow_Soil_PPt_merged$ppt_a, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$ppt_a, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$ppt_a < lower |
weather_Snow_Soil_PPt_merged$ppt_a > upper)

boxplot(weather_Snow_Soil_PPt_merged$ppt_a, main = "Boxplot of ppt_a with
Tukey method")
points(outliers, weather_Snow_Soil_PPt_merged$ppt_a[outliers], col = "red",
pch = 19)
```




```r
#No Outliers

Q1 <- quantile(weather_Snow_Soil_PPt_merged$perc_snow, 0.25)
Q3 <- quantile(weather_Snow_Soil_PPt_merged$perc_snow, 0.75)
IQR <- Q3 - Q1
lower <- Q1 - 1.5 * IQR
upper <- Q3 + 1.5 * IQR

outliers <- which(weather_Snow_Soil_PPt_merged$perc_snow < lower |
weather_Snow_Soil_PPt_merged$perc_snow > upper)

boxplot(weather_Snow_Soil_PPt_merged$perc_snow, main = "Boxplot of perc_snow
with Tukey method")
points(outliers, weather_Snow_Soil_PPt_merged$perc_snow[outliers], col =
"red", pch = 19)
```




```
#No Outliers
```

#After handling outliers

#########Multiple Linear Regression on the combined with correlation dataset without outliers#####################################

```
lm.fits2 <- lm(z_s ~. , data = weather_Snow_Soil_PPt_merged)
summary(lm.fits2)

##
## Call:
## lm(formula = z_s ~ ., data = weather_Snow_Soil_PPt_merged)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.9299  -3.2151  -0.6447   2.1766  29.5461
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.934e+03  3.168e+01 -61.035  < 2e-16 ***
## WY           3.453e+00  9.485e-02  36.407  < 2e-16 ***
## Year        -2.493e+00  9.490e-02 -26.274  < 2e-16 ***
## Month       -1.072e+00  1.153e-02 -93.005  < 2e-16 ***
## Day          2.869e-03  2.056e-03   1.395  0.16297
## Hour         1.681e-02  2.713e-03   6.195 5.87e-10 ***
## T_a         -7.582e-02  1.093e-02  -6.939 3.96e-12 ***
```

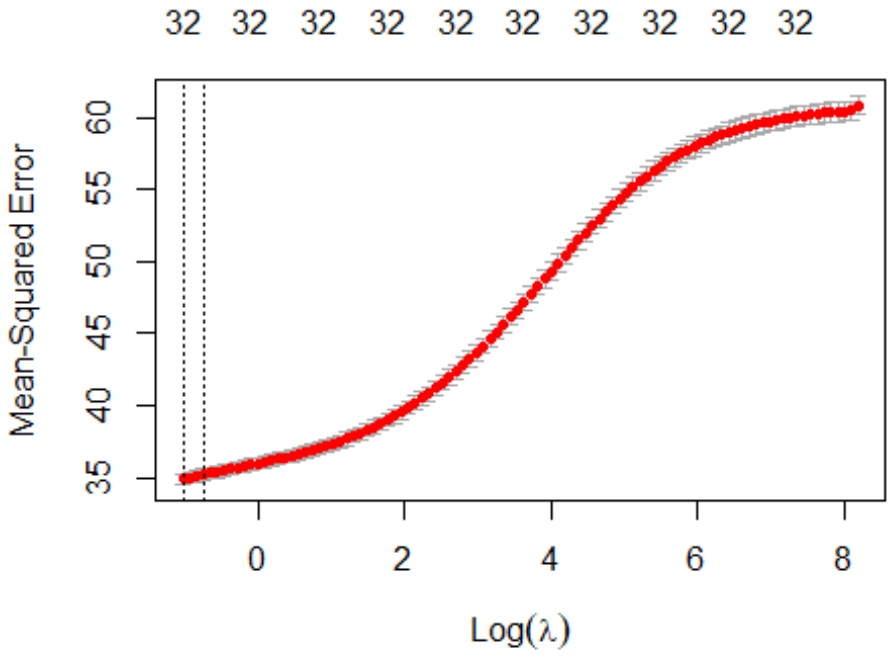
```
## RH           6.856e+00  3.080e-01  22.260  < 2e-16 ***
## e_a          2.005e-02  7.871e-04  25.469  < 2e-16 ***
## T_d         -9.861e-01  3.196e-02 -30.856  < 2e-16 ***
## S_i          7.275e-03  3.813e-04  19.077  < 2e-16 ***
## w_s         -4.249e-01  5.662e-02  -7.504 6.24e-14 ***
## w_d         -7.201e-03  5.335e-04 -13.497  < 2e-16 ***
## T_g_5        3.852e-01  1.843e-02  20.901  < 2e-16 ***
## s_m_5       -2.745e+00  1.795e+00  -1.529  0.12629
## T_g_20      -4.648e-02  4.602e-02  -1.010  0.31253
## s_m_20      -2.200e+01  2.545e+00  -8.646  < 2e-16 ***
## T_g_35       2.907e-01  4.718e-02   6.161 7.25e-10 ***
## s_m_35       4.509e+01  2.601e+00  17.337  < 2e-16 ***
## T_g_50      -1.195e+00  1.206e-01  -9.907  < 2e-16 ***
## s_m_50       7.883e+01  3.584e+00  21.994  < 2e-16 ***
## T_g_75       4.650e+00  1.075e-01  43.246  < 2e-16 ***
## s_m_75      -1.618e+02  3.536e+00 -45.765  < 2e-16 ***
## T_g_90      -5.303e+00  1.074e-01 -49.386  < 2e-16 ***
## s_m_90      -5.434e+01  1.374e+00 -39.554  < 2e-16 ***
## T_g_100     -2.048e-01  7.836e-02  -2.614  0.00895 **
## s_m_100      1.185e+02  3.180e+00  37.273  < 2e-16 ***
## T_g_130     -2.976e-01  1.319e-01  -2.256  0.02408 *
## s_m_130     -2.883e+01  1.597e+00 -18.049  < 2e-16 ***
## T_g_190      1.605e+00  9.116e-02  17.607  < 2e-16 ***
## s_m_190      4.968e+01  1.793e+00  27.712  < 2e-16 ***
## ppt_a        2.551e-01  5.193e-02   4.913 8.99e-07 ***
## perc_snow    2.589e+00  9.069e-02  28.543  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.601 on 96400 degrees of freedom
## Multiple R-squared:  0.4847, Adjusted R-squared:  0.4845
## F-statistic:  2834 on 32 and 96400 DF,  p-value: < 2.2e-16

#Confidence Interval
confint(lm.fits2)

##                     2.5 %        97.5 %
## (Intercept) -1.995824e+03 -1.871630e+03
## WY           3.267297e+00  3.639108e+00
## Year        -2.679268e+00 -2.307274e+00
## Month       -1.094733e+00 -1.049545e+00
## Day         -1.161459e-03  6.899281e-03
## Hour         1.148810e-02  2.212250e-02
## T_a         -9.723592e-02 -5.440567e-02
## RH           6.252160e+00  7.459471e+00
## e_a          1.850406e-02  2.158944e-02
## T_d         -1.048746e+00 -9.234703e-01
## S_i          6.527425e-03  8.022247e-03
## w_s         -5.358782e-01 -3.139195e-01
## w_d         -8.246077e-03 -6.154888e-03
```

```
## T_g_5         3.491129e-01  4.213640e-01
## s_m_5        -6.263287e+00  7.739126e-01
## T_g_20       -1.366720e-01  4.372058e-02
## s_m_20       -2.699159e+01 -1.701562e+01
## T_g_35        1.981995e-01  3.831301e-01
## s_m_35        3.999021e+01  5.018440e+01
## T_g_50       -1.430869e+00 -9.582331e-01
## s_m_50        7.180393e+01  8.585347e+01
## T_g_75        4.439716e+00  4.861254e+00
## s_m_75       -1.687651e+02 -1.549034e+02
## T_g_90       -5.513378e+00 -5.092465e+00
## s_m_90       -5.703195e+01 -5.164671e+01
## T_g_100      -3.584398e-01 -5.125271e-02
## s_m_100       1.122865e+02  1.247511e+02
## T_g_130      -5.560702e-01 -3.903558e-02
## s_m_130      -3.196171e+01 -2.569995e+01
## T_g_190       1.426366e+00  1.783699e+00
## s_m_190       4.616543e+01  5.319272e+01
## ppt_a         1.533527e-01  3.569265e-01
## perc_snow     2.410844e+00  2.766350e+00

#Creating our own function for MSE and RMSE Calculations
MSE2 <- mean(lm.fits2$residuals^2)
RMSE2 <- sqrt(MSE2)

cat("Mean Square Error: ", MSE2)

## Mean Square Error:  31.35586

cat(", Root Mean Square Error: ", RMSE2)

## , Root Mean Square Error:  5.59963

#Compute Error Rate using RSE - Error Rate is RSE divided by mean of response
variable
error2 <- sigma(lm.fits2)/mean(weather_Snow_Soil_PPt_merged$z_s)
cat("\nError rate: ", error2)

##
## Error rate:  1.383973
```

#########Ridge Regression on the combined with correlation dataset without outliers#####################################

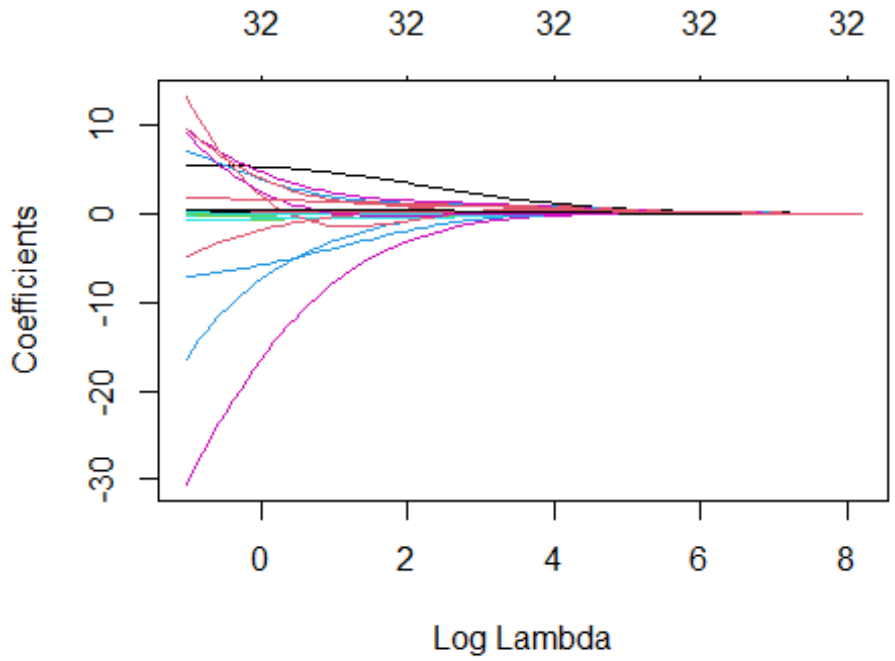
```
library(glmnet)
library(mgcv)
library(visreg)

#Ridge
#pass x matrix and y vector:
x <- model.matrix(z_s ~ ., data=weather_Snow_Soil_PPt_merged )[, -1]
```

```
y <- weather_Snow_Soil_PPt_merged$z_s

model <- glmnet(x, y, alpha = 0)

#find optimal lambda value
ridge.mod <- cv.glmnet(x, y, alpha = 0)
plot(ridge.mod)
```




```
min_lambda_ridge <- ridge.mod$lambda.min
cat("Minimum value of Lambda for ridge: ", min_lambda_ridge,"\n")

## Minimum value of Lambda for ridge:  0.360662

ridge.mod2 <- glmnet(x, y, alpha = 0, lambda = min_lambda_ridge)

coef(ridge.mod2)

## 33 x 1 sparse Matrix of class "dgCMatrix"
##                          s0
## (Intercept) -1.115722e+03
## WY           3.706981e-01
## Year         1.876046e-01
## Month       -7.265427e-01
## Day         -7.244474e-04
## Hour         3.449229e-02
## T_a         -1.757518e-01
## RH           5.379409e+00
```

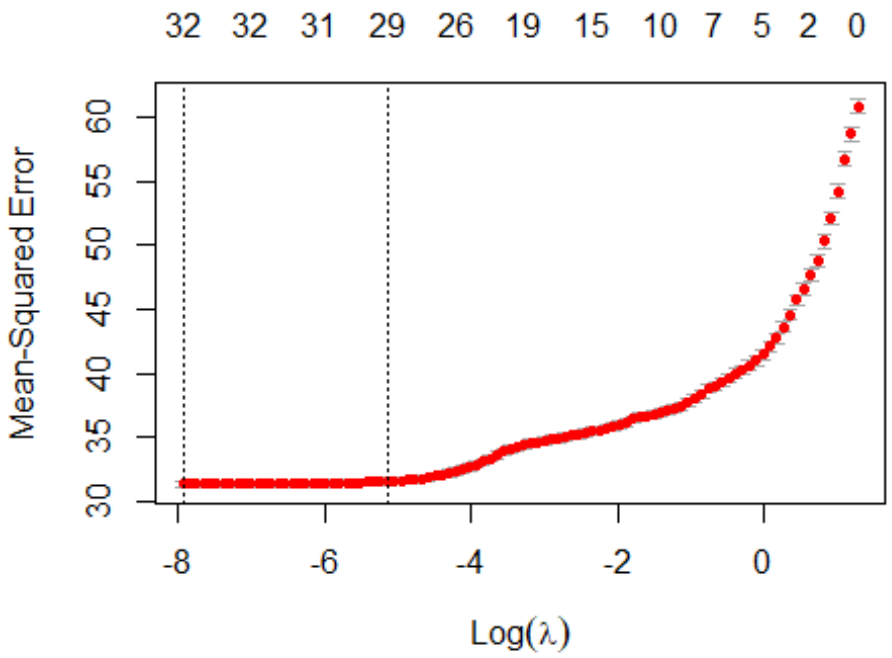
```
## e_a           1.966999e-03
## T_d          -2.561164e-01
## S_i           6.296856e-03
## w_s          -6.271977e-01
## w_d          -1.007415e-02
## T_g_5         1.048187e-01
## s_m_5        -4.603972e+00
## T_g_20       -2.461533e-02
## s_m_20        7.398238e+00
## T_g_35       -8.275310e-02
## s_m_35        9.907609e+00
## T_g_50       -5.880593e-02
## s_m_50        1.002165e+01
## T_g_75        8.815525e-02
## s_m_75       -1.626114e+01
## T_g_90       -1.930241e-01
## s_m_90       -3.125455e+01
## T_g_100       1.250781e-01
## s_m_100       1.189364e+01
## T_g_130      -6.807481e-02
## s_m_130      -7.226266e+00
## T_g_190       6.716946e-02
## s_m_190       9.289761e+00
## ppt_a         3.896236e-01
## perc_snow     1.839239e+00

#produce Ridge trace plot
plot(model, xvar = "lambda")
```




```
#use fitted best model to make predictions on train data

y_pred_ridge <- predict(ridge.mod2, s = min_lambda_ridge, newx=x)

mse_ridge <- mean((y - y_pred_ridge)^2)
rmse_ridge <- sqrt(mse_ridge)
RSS_ridge <- sum((y - y_pred_ridge)^2)
TSS_ridge <- (sum((y - mean(y))^2))
rsquared_ridge <- 1-(RSS_ridge/TSS_ridge)

cat("Mean Square Error Ridge: ", mse_ridge)

## Mean Square Error Ridge:  34.87724

cat("\nRoot Mean Square Error Ridge: ", rmse_ridge)

##
## Root Mean Square Error Ridge:  5.905695

cat("\nR^2 Ridge: ", rsquared_ridge)

##
## R^2 Ridge:  0.4268345
```

#########Lasso Regression on the combined with correlation dataset without outliers#####################################

```r
#Lasso

#pass x matrix and y vector:
x <- model.matrix(z_s ~ ., data=weather_Snow_Soil_PPt_merged )[, -1]
y <- weather_Snow_Soil_PPt_merged$z_s

lasso.mod <- cv.glmnet(x, y, alpha = 1)
#lasso.mod <- cv.glmnet(x, y, alpha = 1)

min_lambda_lasso <- lasso.mod$lambda.min
cat("Minimum value of Lambda: ", min_lambda_lasso,"\n")

## Minimum value of Lambda:  0.000360662

#produce plot of test MSE by lambda value
plot(lasso.mod)
```




```r
lasso.mod2 <- glmnet(x, y, alpha = 1, lambda = min_lambda_lasso)

coef(lasso.mod2)

## 33 x 1 sparse Matrix of class "dgCMatrix"
##                          s0
## (Intercept) -1.918165e+03
## WY           3.419375e+00
## Year        -2.467003e+00
## Month       -1.069781e+00
```

```
## Day           2.986246e-03
## Hour          1.693989e-02
## T_a          -7.828058e-02
## RH            6.750395e+00
## e_a           1.956282e-02
## T_d          -9.677326e-01
## S_i           7.230407e-03
## w_s          -4.269312e-01
## w_d          -7.247369e-03
## T_g_5         3.807209e-01
## s_m_5        -3.516970e+00
## T_g_20       -8.137978e-02
## s_m_20       -1.926109e+01
## T_g_35        1.064487e-01
## s_m_35        3.834742e+01
## T_g_50       -8.135308e-01
## s_m_50        8.320649e+01
## T_g_75        4.731202e+00
## s_m_75       -1.563074e+02
## T_g_90       -5.396720e+00
## s_m_90       -5.471223e+01
## T_g_100      -3.738888e-01
## s_m_100       1.104291e+02
## T_g_130      -3.668086e-01
## s_m_130      -2.854180e+01
## T_g_190       1.697923e+00
## s_m_190       5.040468e+01
## ppt_a         2.627587e-01
## perc_snow     2.561495e+00
```

```
#use fitted best model to make predictions on train data

y_pred_lasso <- predict(lasso.mod2, s = min_lambda_lasso, newx=x)

mse_lasso <- mean((y - y_pred_lasso)^2)
rmse_lasso <- sqrt(mse_lasso)
RSS_lasso <- sum((y - y_pred_lasso)^2)
TSS_lasso <- (sum((y - mean(y))^2))
rsquared_lasso <- 1-(RSS_lasso/TSS_lasso)

cat("Mean Square Error Lasso: ", mse_lasso)

## Mean Square Error Lasso:  31.36329

cat("\n Root Mean Square Error Lasso: ", rmse_lasso)

##
##  Root Mean Square Error Lasso:  5.600294

cat("\n R^2 Lasso: ", rsquared_lasso)
```

```
##
##  R^2 Lasso:  0.484582
```

############################PCA on the combined with correlation dataset without outliers for dimensionality reduction and reducing correlation###################

```
library("FactoMineR")
```

```
## Warning: package 'FactoMineR' was built under R version 4.2.3
```

```
library("factoextra")
```

```
## Warning: package 'factoextra' was built under R version 4.2.3
```

```
## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa
```

```
PCA_DATA = (weather_Snow_Soil_PPt_merged)
#PCA_DATA = subset(PCA_DATA, select = -c(WY, Year, Month, Day, Hour, Minute))
summary(PCA_DATA)
```

```
##        WY            Year          Month             Day              Hour
##  Min.   :2004   Min.   :2003   Min.   : 1.000   Min.   : 1.00   Min.   :
0.0
##  1st Qu.:2006   1st Qu.:2006   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:
5.0
##  Median :2009   Median :2009   Median : 7.000   Median :16.00   Median
:11.0
##  Mean   :2009   Mean   :2009   Mean   : 6.523   Mean   :15.73   Mean
:11.5
##  3rd Qu.:2012   3rd Qu.:2011   3rd Qu.:10.000   3rd Qu.:23.00   3rd
Qu.:17.0
##  Max.   :2015   Max.   :2014   Max.   :12.000   Max.   :31.00   Max.
:23.0
##       T_a               RH              e_a             T_d
##  Min.   :-4.092   Min.   :0.06333   Min.   :277.3   Min.   :-9.8833
##  1st Qu.: 1.725   1st Qu.:0.37500   1st Qu.:410.3   1st Qu.:-4.9687
##  Median : 6.642   Median :0.53333   Median :522.4   Median :-1.9667
##  Mean   : 7.728   Mean   :0.53987   Mean   :543.2   Mean   :-2.0793
##  3rd Qu.:13.633   3rd Qu.:0.69917   3rd Qu.:652.8   3rd Qu.: 0.8167
##  Max.   :21.975   Max.   :1.00000   Max.   :909.9   Max.   : 5.3667
##       S_i              w_s              w_d             z_s
##  Min.   :  0.00   Min.   :0.0000   Min.   :  0.0   Min.   : 0.000
##  1st Qu.:  0.00   1st Qu.:0.0000   1st Qu.:  0.0   1st Qu.: 0.000
##  Median :  0.00   Median :0.0000   Median :  0.0   Median : 0.000
##  Mean   : 19.91   Mean   :0.3839   Mean   : 36.0   Mean   : 4.047
##  3rd Qu.:  0.00   3rd Qu.:0.0000   3rd Qu.:  0.0   3rd Qu.: 4.364
##  Max.   :286.00   Max.   :2.9000   Max.   :253.7   Max.   :42.091
##      T_g_5            s_m_5            T_g_20           s_m_20
##  Min.   : 0.000   Min.   :0.00000   Min.   : 0.000   Min.   :0.00000
```

```
##  1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.: 0.000   1st Qu.:0.00000
##  Median : 0.000   Median :0.00000   Median : 0.000   Median :0.00000
##  Mean   : 3.205   Mean   :0.04797   Mean   : 3.302   Mean   :0.05712
##  3rd Qu.: 2.339   3rd Qu.:0.08813   3rd Qu.: 2.800   3rd Qu.:0.11233
##  Max.   :20.133   Max.   :0.19896   Max.   :19.567   Max.   :0.22778
##      T_g_35           s_m_35           T_g_50           s_m_50
##  Min.   : 0.000   Min.   :0.00000   Min.   : 0.000   Min.   :0.00000
##  1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.: 0.000   1st Qu.:0.00000
##  Median : 0.000   Median :0.00000   Median : 0.000   Median :0.00000
##  Mean   : 4.038   Mean   :0.06436   Mean   : 3.191   Mean   :0.05568
##  3rd Qu.: 5.500   3rd Qu.:0.12950   3rd Qu.: 3.549   3rd Qu.:0.11300
##  Max.   :21.500   Max.   :0.28700   Max.   :17.440   Max.   :0.26011
##      T_g_75           s_m_75           T_g_90           s_m_90
##  Min.   : 0.000   Min.   :0.0000   Min.   : 0.000   Min.   :0.00000
##  1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.: 0.000   1st Qu.:0.00000
##  Median : 0.000   Median :0.0000   Median : 0.000   Median :0.00000
##  Mean   : 2.834   Mean   :0.0575   Mean   : 2.815   Mean   :0.06718
##  3rd Qu.: 3.588   3rd Qu.:0.1182   3rd Qu.: 4.386   3rd Qu.:0.13350
##  Max.   :14.630   Max.   :0.2464   Max.   :13.565   Max.   :0.32100
##     T_g_100          s_m_100          T_g_130          s_m_130
##  Min.   : 0.000   Min.   :0.00000   Min.   : 0.000   Min.   :0.00000
##  1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.: 0.000   1st Qu.:0.00000
##  Median : 0.000   Median :0.00000   Median : 0.000   Median :0.00000
##  Mean   : 2.492   Mean   :0.02824   Mean   : 2.976   Mean   :0.04812
##  3rd Qu.: 3.200   3rd Qu.:0.04616   3rd Qu.: 5.200   3rd Qu.:0.11000
##  Max.   :13.100   Max.   :0.13467   Max.   :13.500   Max.   :0.24200
##     T_g_190          s_m_190           ppt_a            perc_snow
##  Min.   : 0.000   Min.   :0.00000   Min.   : 0.00000   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.: 0.00000   1st Qu.:0.0500
##  Median : 0.000   Median :0.00000   Median : 0.00000   Median :1.0000
##  Mean   : 3.068   Mean   :0.05527   Mean   : 0.06945   Mean   :0.6661
##  3rd Qu.: 6.300   3rd Qu.:0.13800   3rd Qu.: 0.00000   3rd Qu.:1.0000
##  Max.   :13.500   Max.   :0.24600   Max.   :16.33333   Max.   :1.0000
```

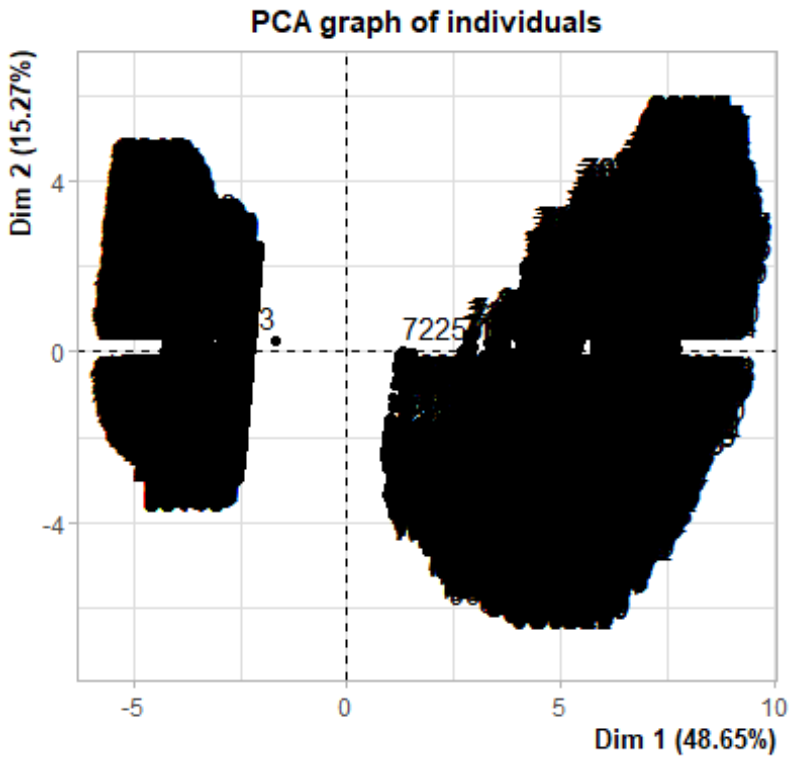
PCA: Help to explain the variability of the dataset using fewer variables.

The goals of PCA are to:

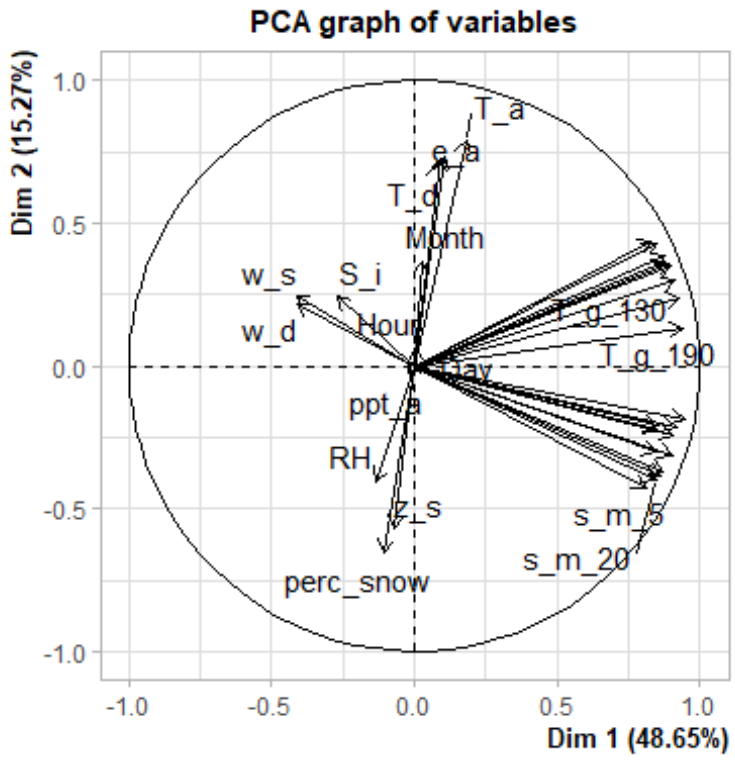
1. Gain an overall structure of the large dimension data,
2. determine key numerical variables based on their contribution to maximum variances in the dataset,
3. compress the size of the data set by keeping only the key variables and removing redundant variables, and
4. find out the correlation among key variables and construct new components for further analysis.

Note that, the PCA method is particularly useful when the variables within the data set are highly correlated and redundant.

```r
#Perform PCA and do biplot visualization
PCA_DATA_1 = PCA(PCA_DATA, scale.unit = TRUE, graph = TRUE)
```




```
## Warning: ggrepel: 16 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



1. Positively correlated variables are grouped together.
2. Negatively correlated variables are located on opposite sides of the plot origin
3. The distance between variables and the origin measures the quality of the variables on the factor map. Variables that are away from the origin are well represented on the factor map.

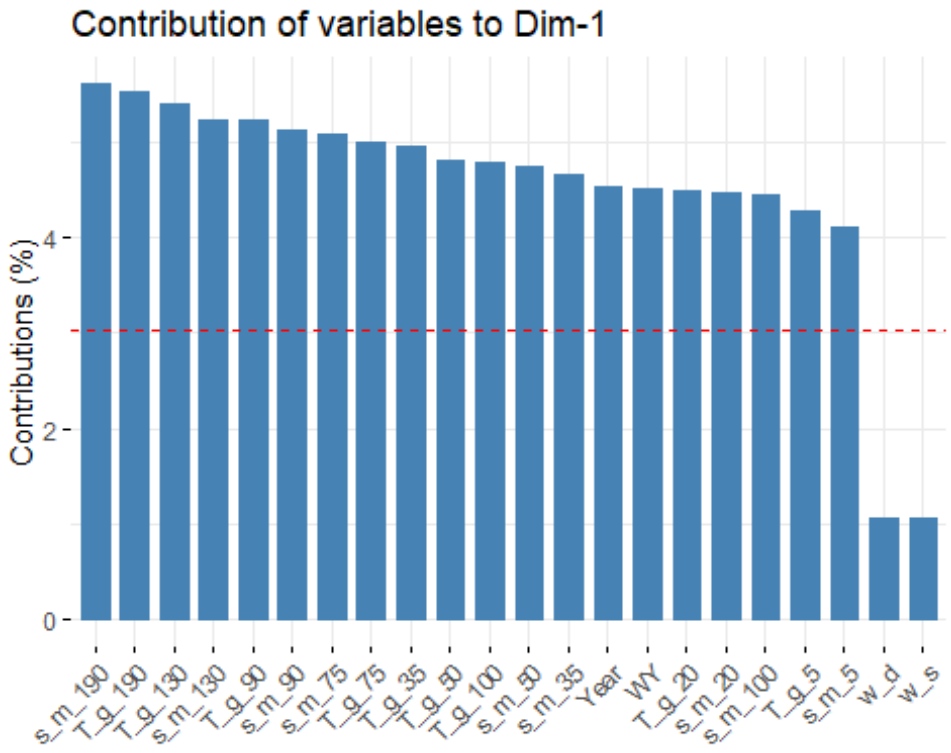
Variables that are closed to circumference (like T_g_20, s_m_5) manifest the maximum representation of the principal components. However, variables like RH, ppt_a show week representation of the principal components.

```
#Get PCA variables
var_PCA = get_pca_var(PCA_DATA_1)
var_PCA

## Principal Component Analysis Results for variables
##  ===================================================
##   Name       Description
## 1 "$coord"   "Coordinates for the variables"
## 2 "$cor"     "Correlations between variables and dimensions"
## 3 "$cos2"    "Cos2 for the variables"
## 4 "$contrib" "contributions of the variables"

#Looking at the variation contribution of each feature
PCA1_Contributions = (var_PCA$contrib)
PCA1_Contributions = subset(PCA1_Contributions, select = c(Dim.1)) #Keeping
only Dim.1
```

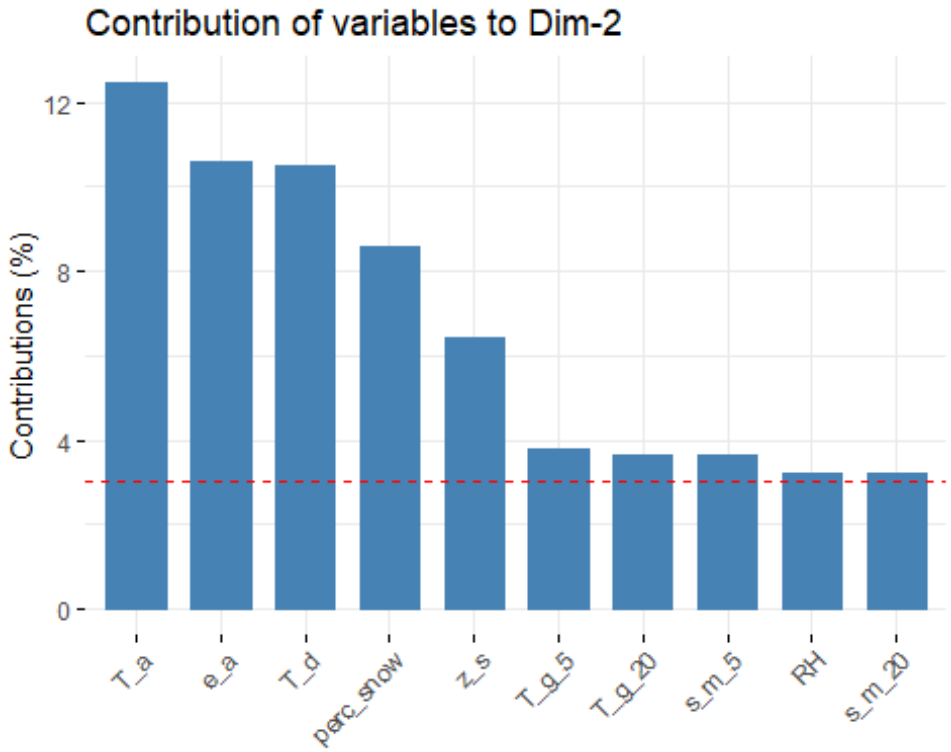
```
BB = ordered(PCA1_Contributions)
PCA1_Contributions
```

```
##                  Dim.1
## WY        4.503361e+00
## Year      4.525336e+00
## Month     3.686727e-03
## Day       3.246947e-06
## Hour      1.044972e-04
## T_a       2.024532e-01
## RH        1.137724e-01
## e_a       6.493501e-02
## T_d       5.259825e-02
## S_i       4.563616e-01
## w_s       1.066420e+00
## w_d       1.067385e+00
## z_s       3.109685e-02
## T_g_5     4.271940e+00
## s_m_5     4.100422e+00
## T_g_20    4.480561e+00
## s_m_20    4.474998e+00
## T_g_35    4.953516e+00
## s_m_35    4.653442e+00
## T_g_50    4.796857e+00
## s_m_50    4.731069e+00
## T_g_75    4.988126e+00
## s_m_75    5.067006e+00
## T_g_90    5.219346e+00
## s_m_90    5.124052e+00
## T_g_100   4.789854e+00
## s_m_100   4.449666e+00
## T_g_130   5.384774e+00
## s_m_130   5.232131e+00
## T_g_190   5.521818e+00
## s_m_190   5.597144e+00
## ppt_a     1.325161e-03
## perc_snow 7.443724e-02
```

```
#Visualizing the variation contributions
fviz_contrib(PCA_DATA_1, choice = "var", axes = 1, top = 22) #Contributions
of variables to PC1 : top = top 22 contributors
```




```
fviz_contrib(PCA_DATA_1, choice = "var", axes = 2, top = 10) #Contributions of variables to PC2 : top = top 22 contributors
```



```r
pca <- prcomp(PCA_DATA)

# Get the loadings (correlations between variables and PCs)
loadings <- pca$rotation

# Get the top 10 contributors for each PC
top_20 <- apply(loadings, 2, function(x) names(sort(abs(x), decreasing =
TRUE)[1:20]))

# Print the names of the top 10 contributors for each PC
for (i in 1:1) {
  #cat(sprintf("Top 10 contributors for PC%d:\n", i))
  AA = (top_20[, i])
}

LL = noquote(AA)
LL
```

```
##  [1] e_a       S_i       w_d       T_a       T_d       z_s       T_g_5
##  [8] T_g_20    T_g_35    T_g_50    T_g_75    T_g_100   T_g_90    T_g_130
## [15] T_g_190   Month     perc_snow Hour      WY        Year
```

```r
#PP = toString(ZZ)
PCA_weather_Snow_Soil_PPt_merged = subset(PCA_DATA, select = c(e_a, S_i, w_d,
T_a, T_d, z_s, T_g_5, T_g_20, T_g_35, T_g_50, T_g_75, T_g_100, T_g_90,
T_g_130, T_g_190, Month, perc_snow, Hour, WY, Year))

#Creating a sub-dataframe to perform MLR after using PCA components

summary(PCA_weather_Snow_Soil_PPt_merged)
```

```
##       e_a             S_i              w_d             T_a
##  Min.   :277.3   Min.   :  0.00   Min.   :  0.0   Min.   :-4.092
##  1st Qu.:410.3   1st Qu.:  0.00   1st Qu.:  0.0   1st Qu.: 1.725
##  Median :522.4   Median :  0.00   Median :  0.0   Median : 6.642
##  Mean   :543.2   Mean   : 19.91   Mean   : 36.0   Mean   : 7.728
##  3rd Qu.:652.8   3rd Qu.:  0.00   3rd Qu.:  0.0   3rd Qu.:13.633
##  Max.   :909.9   Max.   :286.00   Max.   :253.7   Max.   :21.975
##       T_d               z_s             T_g_5            T_g_20
##  Min.   :-9.8833   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:-4.9687   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
##  Median :-1.9667   Median : 0.000   Median : 0.000   Median : 0.000
##  Mean   :-2.0793   Mean   : 4.047   Mean   : 3.205   Mean   : 3.302
##  3rd Qu.: 0.8167   3rd Qu.: 4.364   3rd Qu.: 2.339   3rd Qu.: 2.800
##  Max.   : 5.3667   Max.   :42.091   Max.   :20.133   Max.   :19.567
##      T_g_35           T_g_50           T_g_75           T_g_100
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
##  Median : 0.000   Median : 0.000   Median : 0.000   Median : 0.000
##  Mean   : 4.038   Mean   : 3.191   Mean   : 2.834   Mean   : 2.492
```

```
##  3rd Qu.: 5.500   3rd Qu.: 3.549   3rd Qu.: 3.588   3rd Qu.: 3.200
##  Max.   :21.500   Max.   :17.440   Max.   :14.630   Max.   :13.100
##      T_g_90          T_g_130          T_g_190           Month
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 1.000
##  1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 4.000
##  Median : 0.000   Median : 0.000   Median : 0.000   Median : 7.000
##  Mean   : 2.815   Mean   : 2.976   Mean   : 3.068   Mean   : 6.523
##  3rd Qu.: 4.386   3rd Qu.: 5.200   3rd Qu.: 6.300   3rd Qu.:10.000
##  Max.   :13.565   Max.   :13.500   Max.   :13.500   Max.   :12.000
##    perc_snow           Hour            WY            Year
##  Min.   :0.0000   Min.   : 0.0   Min.   :2004   Min.   :2003
##  1st Qu.:0.0500   1st Qu.: 5.0   1st Qu.:2006   1st Qu.:2006
##  Median :1.0000   Median :11.0   Median :2009   Median :2009
##  Mean   :0.6661   Mean   :11.5   Mean   :2009   Mean   :2009
##  3rd Qu.:1.0000   3rd Qu.:17.0   3rd Qu.:2012   3rd Qu.:2011
##  Max.   :1.0000   Max.   :23.0   Max.   :2015   Max.   :2014
```

Performing MLR on PCA data

```
PCA_MLR_ALL_Merged_data = PCA_weather_Snow_Soil_PPt_merged

#plot(PCA_MLR_ALL_Merged_data) # the plot() gives a visual representation of
the relation between the various columns in the dataset

#Observe new correlation
cor(PCA_MLR_ALL_Merged_data)
```

```
##                   e_a          S_i          w_d         T_a         T_d
## e_a        1.00000000  0.120661837  0.067831787  0.61742557  0.98611795
## S_i        0.12066184  1.000000000  0.489065295  0.15250881  0.12282399
## w_d        0.06783179  0.489065295  1.000000000 -0.03451613  0.07517742
## T_a        0.61742557  0.152508811 -0.034516128  1.00000000  0.62669098
## T_d        0.98611795  0.122823989  0.075177420  0.62669098  1.00000000
## z_s       -0.35015686 -0.134638774 -0.197961680 -0.46234840 -0.36011266
## T_g_5      0.29117608 -0.149959152 -0.233428799  0.46262465  0.28399237
## T_g_20     0.27846310 -0.155509614 -0.242068737  0.42616171  0.27052495
## T_g_35     0.23224689 -0.171130410 -0.266384316  0.37063262  0.22406700
## T_g_50     0.24782178 -0.166184065 -0.258684757  0.38387488  0.23907420
## T_g_75     0.22243659 -0.173568709 -0.270179810  0.34875805  0.21363228
## T_g_100    0.21750967 -0.170666503 -0.265662191  0.34428562  0.20836215
## T_g_90     0.19364098 -0.182241049 -0.283679313  0.31276368  0.18359674
## T_g_130    0.16149762 -0.190402471 -0.296383512  0.27039606  0.15048157
## T_g_190    0.10472496 -0.201161535 -0.313131242  0.19370504  0.09230060
## Month      0.11617687 -0.002217975  0.024976387  0.18920978  0.11508724
## perc_snow -0.86729730 -0.080578150 -0.023153870 -0.53167528 -0.82437373
## Hour       0.04196407  0.070792439  0.001871272  0.11248321  0.04378648
## WY        -0.04876461 -0.407451140 -0.634214411  0.01807248 -0.05943743
## Year      -0.01920259 -0.399472733 -0.634870797  0.05894019 -0.03006423
##                   z_s       T_g_5       T_g_20      T_g_35      T_g_50
## e_a       -0.350156862  0.29117608  0.278463103  0.23224689  0.24782178
## S_i       -0.134638774 -0.14995915 -0.155509614 -0.17113041 -0.16618407
```

```
## w_d       -0.197961680 -0.23342880 -0.242068737 -0.26638432 -0.25868476
## T_a       -0.462348395  0.46262465  0.426161705  0.37063262  0.38387488
## T_d       -0.360112655  0.28399237  0.270524952  0.22406700  0.23907420
## z_s        1.000000000 -0.23021764 -0.230220539 -0.21056975 -0.21594116
## T_g_5     -0.230217636  1.00000000  0.983065585  0.96037249  0.96386492
## T_g_20    -0.230220539  0.98306559  1.000000000  0.98590016  0.98948831
## T_g_35    -0.210569752  0.96037249  0.985900158  1.00000000  0.99563219
## T_g_50    -0.215941160  0.96386492  0.989488307  0.99563219  1.00000000
## T_g_75    -0.196559377  0.94221453  0.972189304  0.98940821  0.99451386
## T_g_100   -0.195713955  0.92547743  0.958377544  0.97658868  0.98600761
## T_g_90    -0.183721441  0.92322144  0.954285802  0.98115082  0.98413291
## T_g_130   -0.157439710  0.89247164  0.925659957  0.96309987  0.96344529
## T_g_190   -0.111111895  0.82788971  0.863817842  0.92014612  0.91425693
## Month     -0.400244114  0.10242006  0.123174065  0.14235004  0.15068903
## perc_snow  0.333650821 -0.25607846 -0.248341066 -0.20978338 -0.22249906
## Hour      -0.002158867  0.04850397  0.007012558 -0.00276354 -0.00103857
## WY         0.122058656  0.57460419  0.594888246  0.65597282  0.63724789
## Year       0.132324428  0.59003757  0.607061970  0.66100869  0.64235648
##                  T_g_75       T_g_100        T_g_90       T_g_130
T_g_190
## e_a        0.2224365904  2.175097e-01  1.936410e-01  1.614976e-01
1.047250e-01
## S_i       -0.1735687089 -1.706665e-01 -1.822410e-01 -1.904025e-01 -
2.011615e-01
## w_d       -0.2701798101 -2.656622e-01 -2.836793e-01 -2.963835e-01 -
3.131312e-01
## T_a        0.3487580532  3.442856e-01  3.127637e-01  2.703961e-01
1.937050e-01
## T_d        0.2136322837  2.083622e-01  1.835967e-01  1.504816e-01
9.230060e-02
## z_s       -0.1965593772 -1.957140e-01 -1.837214e-01 -1.574397e-01 -
1.111119e-01
## T_g_5      0.9422145339  9.254774e-01  9.232214e-01  8.924716e-01
8.278897e-01
## T_g_20     0.9721893039  9.583775e-01  9.542858e-01  9.256600e-01
8.638178e-01
## T_g_35     0.9894082097  9.765887e-01  9.811508e-01  9.630999e-01
9.201461e-01
## T_g_50     0.9945138620  9.860076e-01  9.841329e-01  9.634453e-01
9.142569e-01
## T_g_75     1.0000000000  9.953722e-01  9.948690e-01  9.813909e-01
9.430370e-01
## T_g_100    0.9953721545  1.000000e+00  9.926223e-01  9.818103e-01
9.448250e-01
## T_g_90     0.9948689538  9.926223e-01  1.000000e+00  9.938842e-01
9.672065e-01
## T_g_130    0.9813909230  9.818103e-01  9.938842e-01  1.000000e+00
9.874740e-01
## T_g_190    0.9430370381  9.448250e-01  9.672065e-01  9.874740e-01
1.000000e+00
```

```
## Month      0.1652226451  1.894735e-01  1.671436e-01  1.657556e-01
1.560463e-01
## perc_snow -0.2013744309 -1.985079e-01 -1.801720e-01 -1.556072e-01 -
1.095802e-01
## Hour       0.0002114189 -5.801379e-06  5.110767e-05  1.031523e-06
4.869719e-05
## WY         0.6639539506  6.511199e-01  7.015644e-01  7.331454e-01
7.759520e-01
## Year       0.6638035772  6.468344e-01  6.978550e-01  7.253182e-01
7.627321e-01
##                   Month   perc_snow          Hour            WY
Year
## e_a        1.161769e-01 -0.86729730  4.196407e-02 -4.876461e-02 -
1.920259e-02
## S_i       -2.217975e-03 -0.08057815  7.079244e-02 -4.074511e-01 -
3.994727e-01
## w_d        2.497639e-02 -0.02315387  1.871272e-03 -6.342144e-01 -
6.348708e-01
## T_a        1.892098e-01 -0.53167528  1.124832e-01  1.807248e-02
5.894019e-02
## T_d        1.150872e-01 -0.82437373  4.378648e-02 -5.943743e-02 -
3.006423e-02
## z_s       -4.002441e-01  0.33365082 -2.158867e-03  1.220587e-01
1.323244e-01
## T_g_5      1.024201e-01 -0.25607846  4.850397e-02  5.746042e-01
5.900376e-01
## T_g_20     1.231741e-01 -0.24834107  7.012558e-03  5.948882e-01
6.070620e-01
## T_g_35     1.423500e-01 -0.20978338 -2.763540e-03  6.559728e-01
6.610087e-01
## T_g_50     1.506890e-01 -0.22249906 -1.038570e-03  6.372479e-01
6.423565e-01
## T_g_75     1.652226e-01 -0.20137443  2.114189e-04  6.639540e-01
6.638036e-01
## T_g_100    1.894735e-01 -0.19850792 -5.801379e-06  6.511199e-01
6.468344e-01
## T_g_90     1.671436e-01 -0.18017204  5.110767e-05  7.015644e-01
6.978550e-01
## T_g_130    1.657556e-01 -0.15560715  1.031523e-06  7.331454e-01
7.253182e-01
## T_g_190    1.560463e-01 -0.10958021  4.869719e-05  7.759520e-01
7.627321e-01
## Month      1.000000e+00 -0.12116424 -1.737015e-05  3.294394e-04 -
1.021104e-01
## perc_snow -1.211642e-01  1.00000000 -3.017604e-02  1.261852e-02 -
1.258286e-02
## Hour      -1.737015e-05 -0.03017604  1.000000e+00 -3.268834e-05 -
2.834783e-05
## WY         3.294394e-04  0.01261852 -3.268834e-05  1.000000e+00
9.907104e-01
```

```
## Year      -1.021104e-01 -0.01258286 -2.834783e-05  9.907104e-01
1.000000e+00

library("caTools")
```

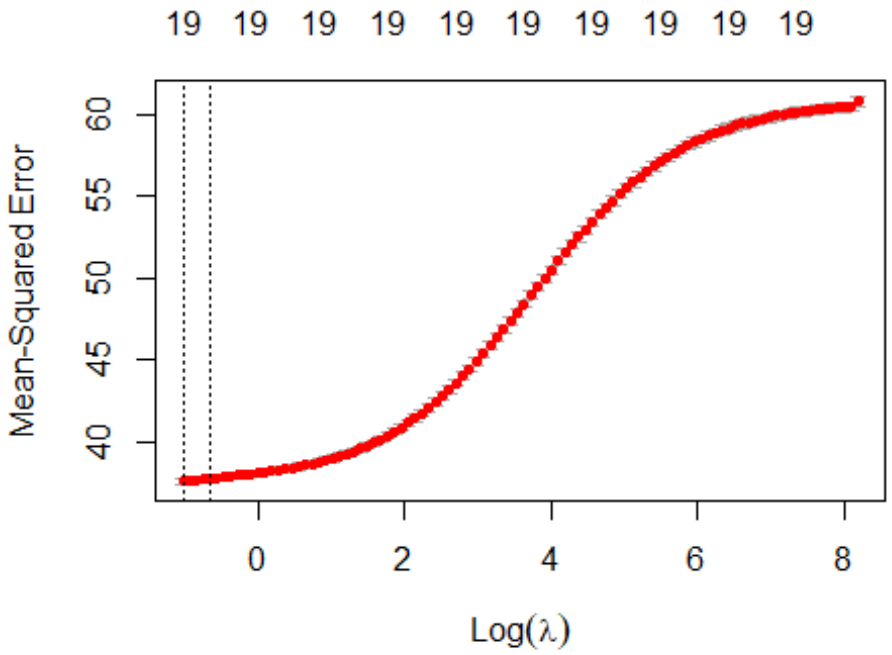
#########Multiple linear regression on the combined with correlation dataset without outliers with PCA Dimensionality Reduction#####################################

```
model_PCA_MLR_ALL_Merged_data <-lm(z_s ~., data = PCA_MLR_ALL_Merged_data)
summary(model_PCA_MLR_ALL_Merged_data)

##
## Call:
## lm(formula = z_s ~ ., data = PCA_MLR_ALL_Merged_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.8060  -3.5483  -0.9602   2.0800  31.0636
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.322e+03  3.085e+01 -42.873  < 2e-16 ***
## e_a          2.576e-02  8.304e-04  31.020  < 2e-16 ***
## S_i          4.322e-03  3.381e-04  12.783  < 2e-16 ***
## w_d         -1.268e-02  3.667e-04 -34.581  < 2e-16 ***
## T_a         -3.054e-01  4.652e-03 -65.639  < 2e-16 ***
## T_d         -9.473e-01  3.099e-02 -30.566  < 2e-16 ***
## T_g_5        3.358e-01  1.944e-02  17.271  < 2e-16 ***
## T_g_20       3.131e-02  4.609e-02   0.679    0.497
## T_g_35      -3.857e-01  4.585e-02  -8.412  < 2e-16 ***
## T_g_50      -6.406e-01  1.202e-01  -5.329 9.92e-08 ***
## T_g_75       2.441e+00  1.023e-01  23.852  < 2e-16 ***
## T_g_100      2.559e+00  7.004e-02  36.545  < 2e-16 ***
## T_g_90      -4.239e+00  1.043e-01 -40.652  < 2e-16 ***
## T_g_130     -1.750e+00  1.239e-01 -14.126  < 2e-16 ***
## T_g_190      1.593e+00  7.751e-02  20.556  < 2e-16 ***
## Month       -8.584e-01  1.194e-02 -71.887  < 2e-16 ***
## perc_snow    3.085e+00  9.583e-02  32.198  < 2e-16 ***
## Hour         1.926e-02  2.875e-03   6.701 2.08e-11 ***
## WY           2.177e+00  9.911e-02  21.965  < 2e-16 ***
## Year        -1.521e+00  9.936e-02 -15.308  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.946 on 96413 degrees of freedom
## Multiple R-squared:  0.4191, Adjusted R-squared:  0.419
## F-statistic:  3662 on 19 and 96413 DF,  p-value: < 2.2e-16

#Creating our own function for MSE and RMSE Calculations
MSE3 <- mean(model_PCA_MLR_ALL_Merged_data$residuals^2)
```

```
RMSE3 <- sqrt(MSE3)

cat("Mean Square Error: ", MSE3)

## Mean Square Error:  35.34552

cat(", Root Mean Square Error: ", RMSE3)

## , Root Mean Square Error:  5.94521

#Compute Error Rate using RSE - Error Rate is RSE divided by mean of response
variable
error3 <-
sigma(model_PCA_MLR_ALL_Merged_data)/mean(PCA_MLR_ALL_Merged_data$z_s)
cat("\nError rate: ", error3)

##
## Error rate:  1.469285
```

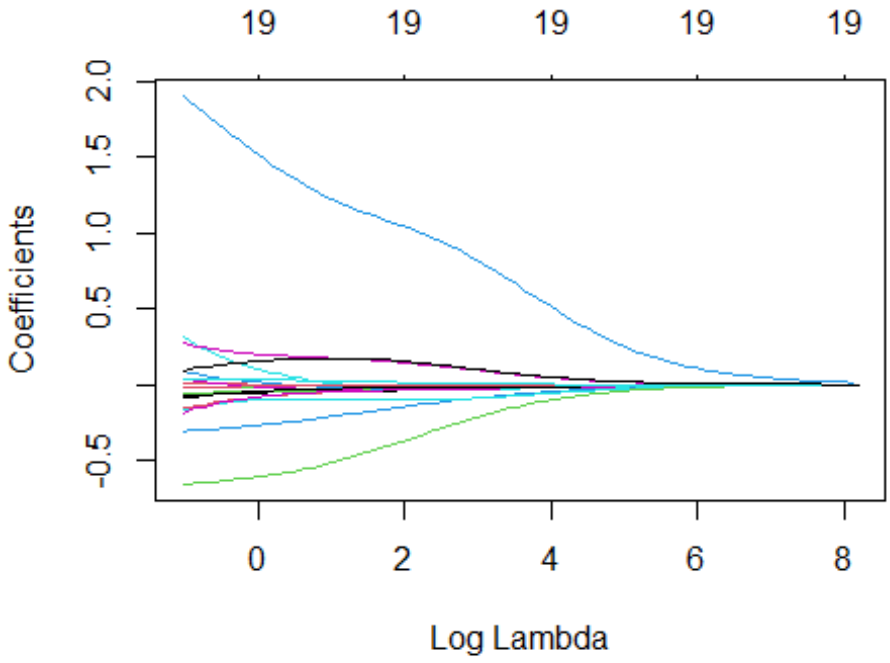
#########Ridge Regression on the combined with correlation dataset without outliers with PCA Dimensionality Reduction#####################################

```
library(glmnet)
library(mgcv)
library(visreg)

#Ridge
#pass x matrix and y vector:
x <- model.matrix(z_s ~ ., data=PCA_MLR_ALL_Merged_data )[, -1]
y <- PCA_MLR_ALL_Merged_data$z_s

model <- glmnet(x, y, alpha = 0)

#find optimal lambda value
ridge.mod <- cv.glmnet(x, y, alpha = 0)
plot(ridge.mod)
```




```
min_lambda_ridge <- ridge.mod$lambda.min
cat("Minimum value of Lambda for ridge: ", min_lambda_ridge,"\n")

## Minimum value of Lambda for ridge:  0.360662

ridge.mod2 <- glmnet(x, y, alpha = 0, lambda = min_lambda_ridge)

coef(ridge.mod2)

## 20 x 1 sparse Matrix of class "dgCMatrix"
##                         s0
## (Intercept) -7.269076e+02
## e_a          4.192297e-03
## S_i          2.364112e-03
## w_d         -1.505801e-02
## T_a         -3.071267e-01
## T_d         -1.724136e-01
## T_g_5        3.960157e-02
## T_g_20      -7.136839e-02
## T_g_35      -1.635921e-01
## T_g_50      -7.828437e-02
## T_g_75       7.204021e-02
## T_g_100      3.163996e-01
## T_g_90      -1.689559e-01
## T_g_130     -8.322181e-02
## T_g_190     -2.129360e-02
## Month       -6.584241e-01
```

```
## perc_snow    1.909090e+00
## Hour         3.431895e-02
## WY           2.711134e-01
## Year         9.456842e-02

#produce Ridge trace plot
plot(model, xvar = "lambda")
```

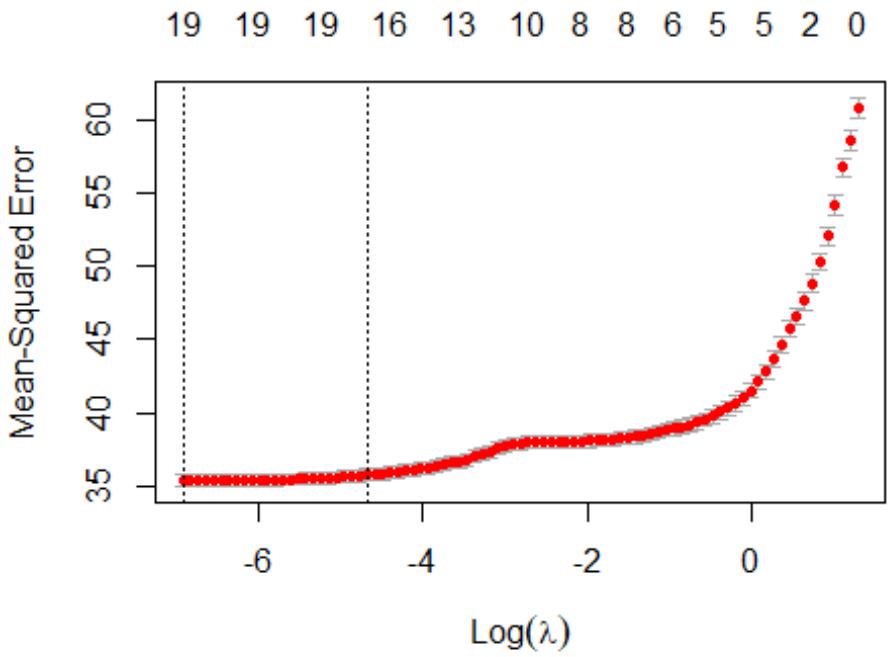



```
#use fitted best model to make predictions on train data

y_pred_ridge <- predict(ridge.mod2, s = min_lambda_ridge, newx=x)

mse_ridge <- mean((y - y_pred_ridge)^2)
rmse_ridge <- sqrt(mse_ridge)
RSS_ridge <- sum((y - y_pred_ridge)^2)
TSS_ridge <- (sum((y - mean(y))^2))
rsquared_ridge <- 1-(RSS_ridge/TSS_ridge)

cat("Mean Square Error Ridge: ", mse_ridge)

## Mean Square Error Ridge:  37.5346

cat("\nRoot Mean Square Error Ridge: ", rmse_ridge)

##
## Root Mean Square Error Ridge:  6.126549

cat("\nR^2 Ridge: ", rsquared_ridge)
```

```
##
## R^2 Ridge:  0.383164
```

#########Lasso Regression on the combined with correlation dataset without outliers with PCA Dimensionality Reduction#####################################

```
#Lasso

#pass x matrix and y vector:
x <- model.matrix(z_s ~ ., data=PCA_MLR_ALL_Merged_data )[, -1]
y <- PCA_MLR_ALL_Merged_data$z_s

lasso.mod <- cv.glmnet(x, y, alpha = 1)
#lasso.mod <- cv.glmnet(x, y, alpha = 1)

min_lambda_lasso <- lasso.mod$lambda.min
cat("Minimum value of Lambda: ", min_lambda_lasso,"\n")

## Minimum value of Lambda:  0.001003563

#produce plot of test MSE by lambda value
plot(lasso.mod)
```




```
lasso.mod2 <- glmnet(x, y, alpha = 1, lambda = min_lambda_lasso)

coef(lasso.mod2)
```

```
## 20 x 1 sparse Matrix of class "dgCMatrix"
##                          s0
## (Intercept) -1.328606e+03
## e_a          2.488327e-02
## S_i          4.348133e-03
## w_d         -1.262963e-02
## T_a         -3.090233e-01
## T_d         -9.161381e-01
## T_g_5        3.351829e-01
## T_g_20      -5.634336e-02
## T_g_35      -4.589518e-01
## T_g_50      -3.155454e-01
## T_g_75       2.410494e+00
## T_g_100      2.455155e+00
## T_g_90      -4.532589e+00
## T_g_130     -1.406062e+00
## T_g_190      1.492215e+00
## Month       -8.408470e-01
## perc_snow    3.043685e+00
## Hour         2.023579e-02
## WY           2.027377e+00
## Year        -1.368216e+00
```

```
#use fitted best model to make predictions on train data

y_pred_lasso <- predict(lasso.mod2, s = min_lambda_lasso, newx=x)

mse_lasso <- mean((y - y_pred_lasso)^2)
rmse_lasso <- sqrt(mse_lasso)
RSS_lasso <- sum((y - y_pred_lasso)^2)
TSS_lasso <- (sum((y - mean(y))^2))
rsquared_lasso <- 1-(RSS_lasso/TSS_lasso)

cat("Mean Square Error Lasso: ", mse_lasso)
```

```
## Mean Square Error Lasso:  35.35376
```

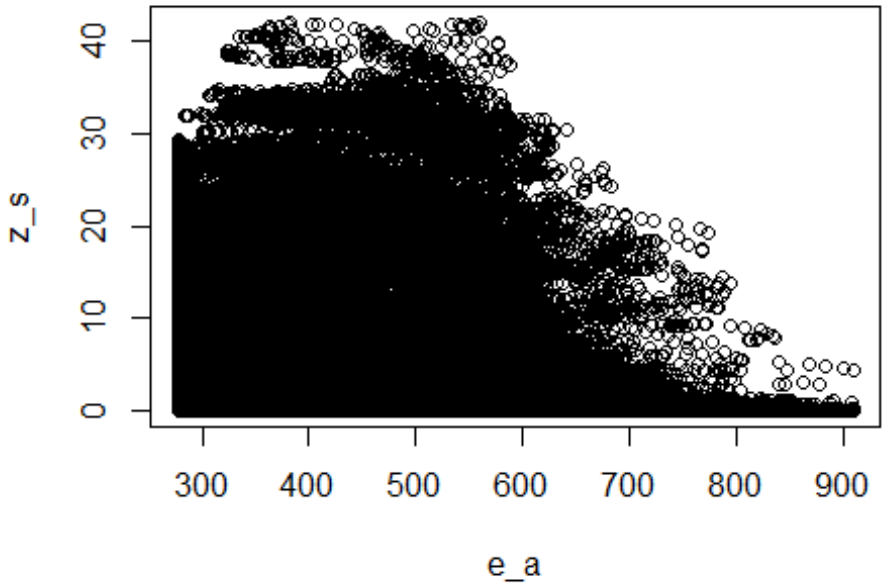
```
cat("\n Root Mean Square Error Lasso: ", rmse_lasso)
```

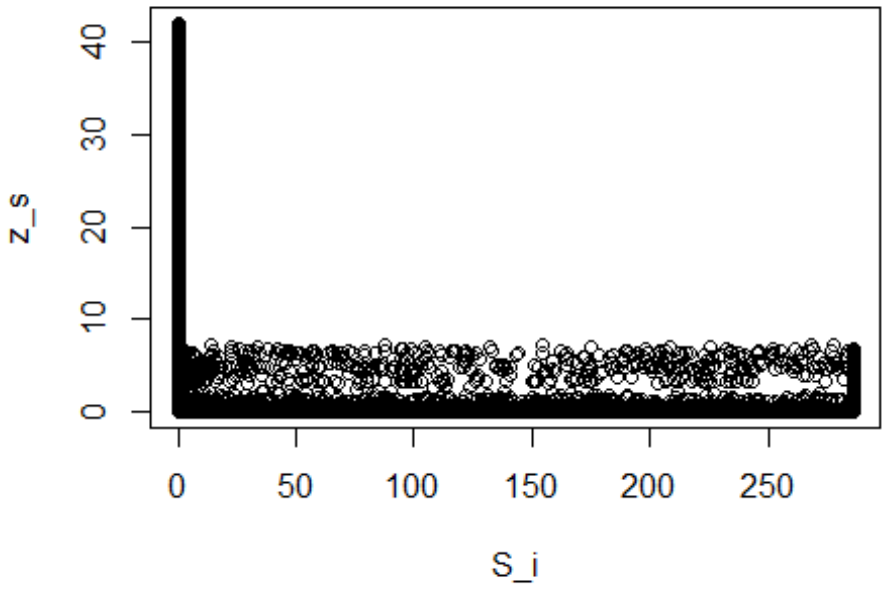
```
##
##  Root Mean Square Error Lasso:  5.945903
```

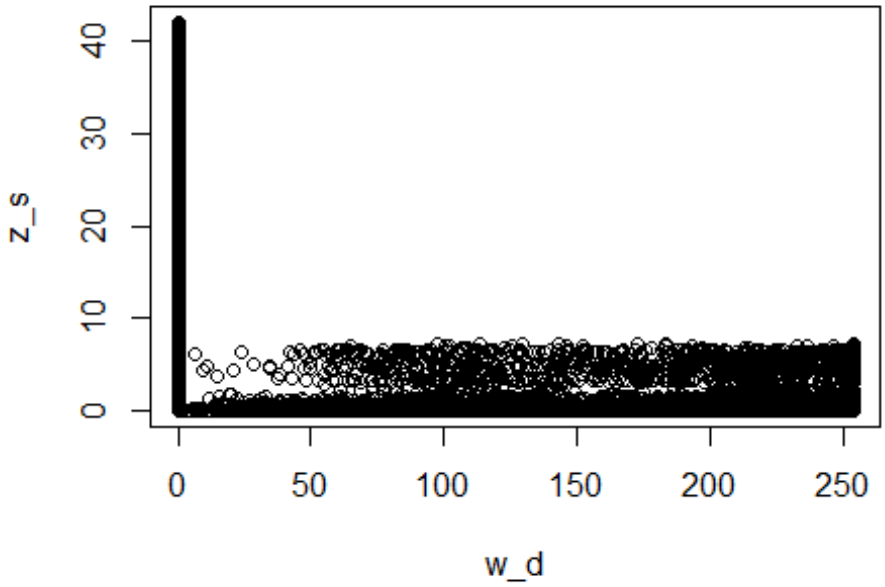
```
cat("\n R^2 Lasso: ", rsquared_lasso)
```

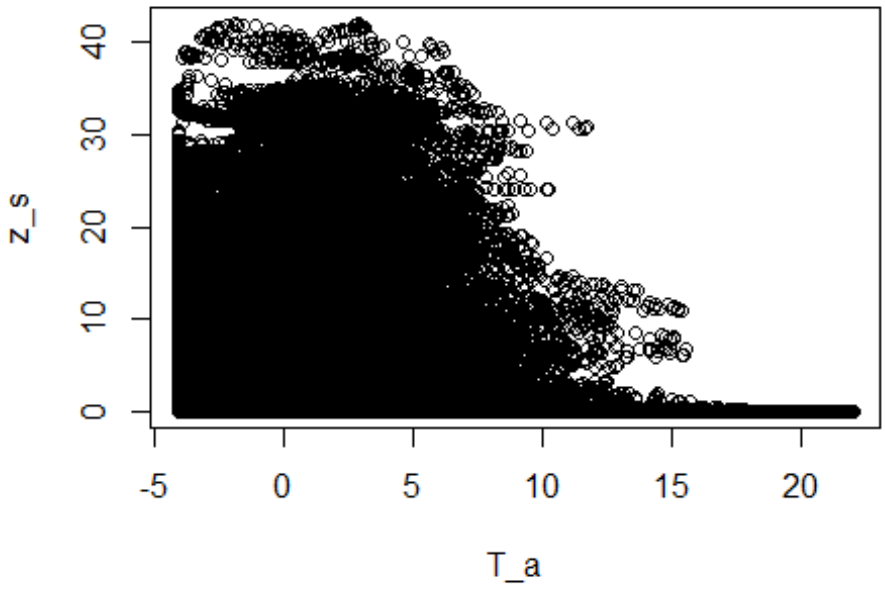
```
##
##  R^2 Lasso:  0.4190035
```

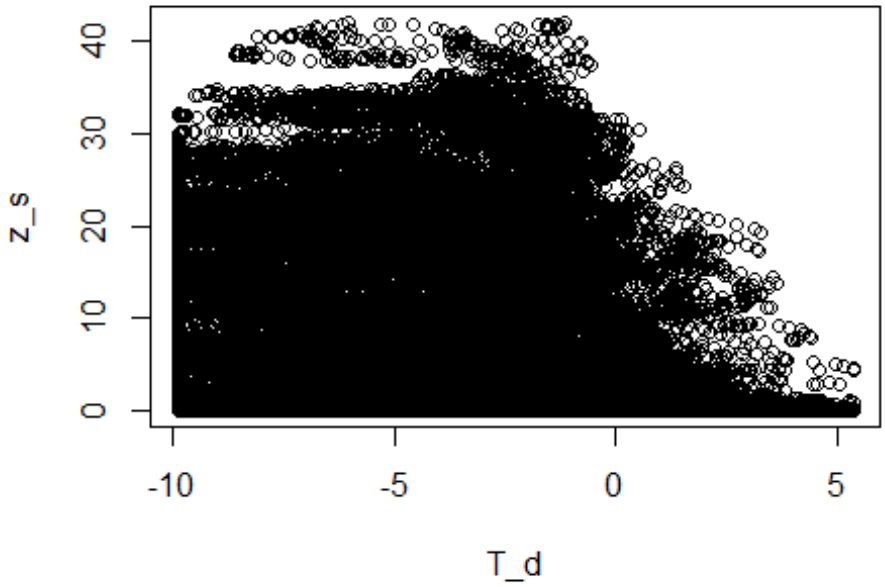
```
#Bivariate visualization - Plotting snow depth as a function of each feature
plot(z_s ~., data = PCA_MLR_ALL_Merged_data)
```

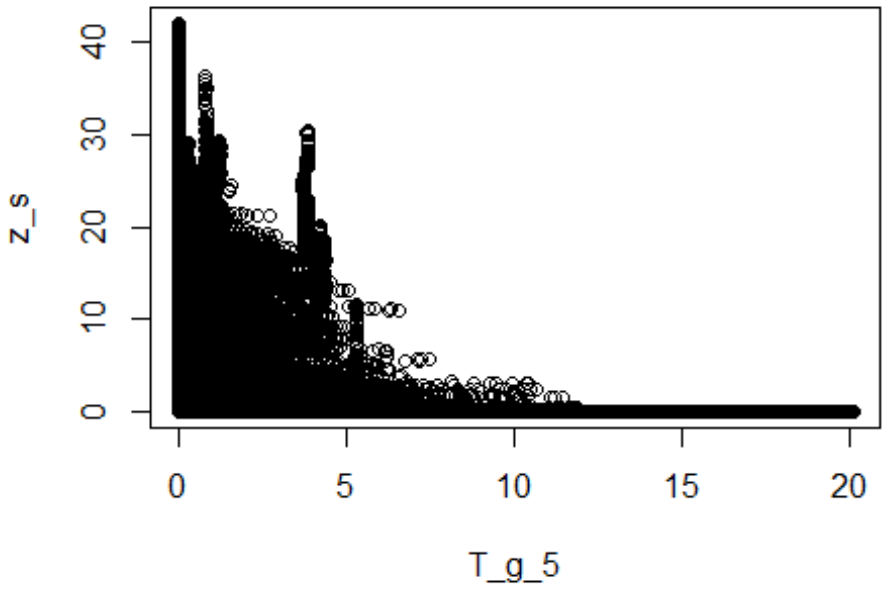

z_s
0
10
20
30
40
300
400
500
600
700
800
900
e_a

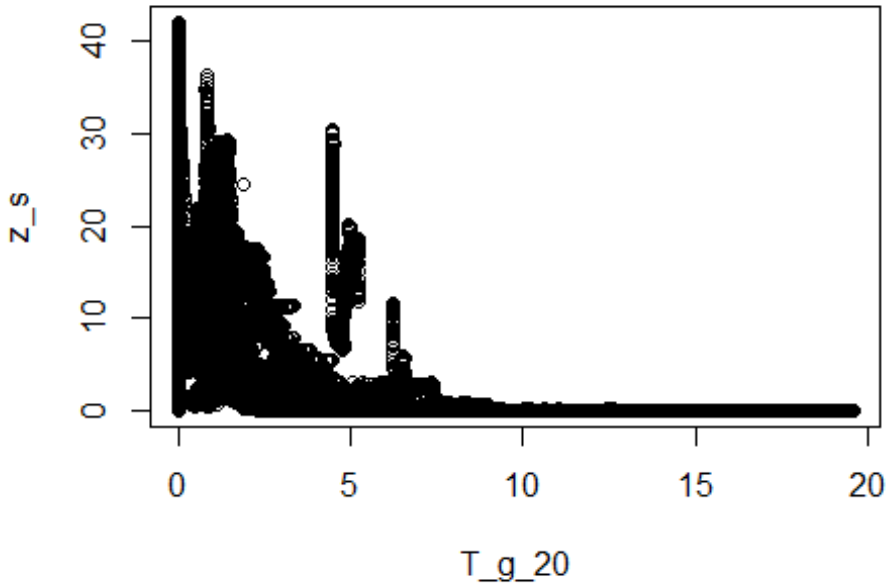

z_s
0
10
20
30
40
0
50
100
150
200
250
S_i

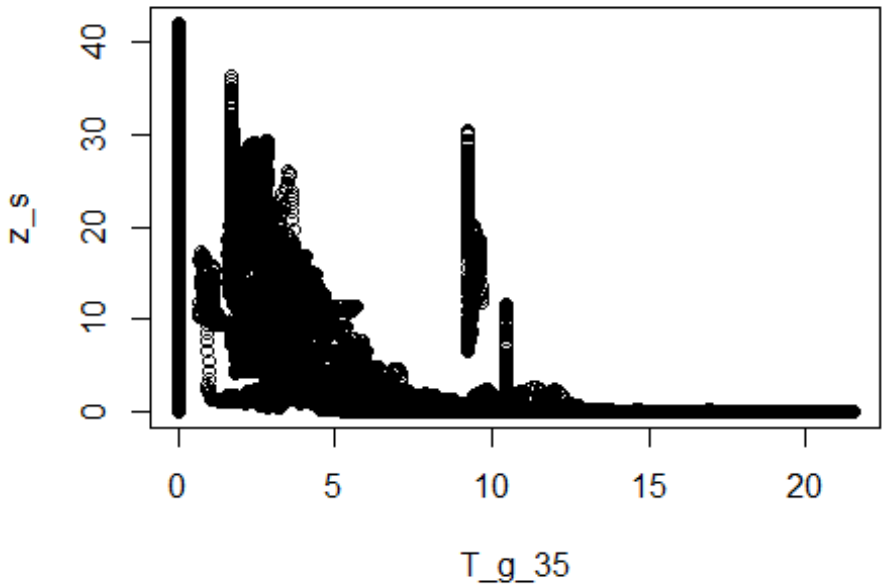

z_s
0
10
20
30
40
0
50
100
150
200
250
w_d

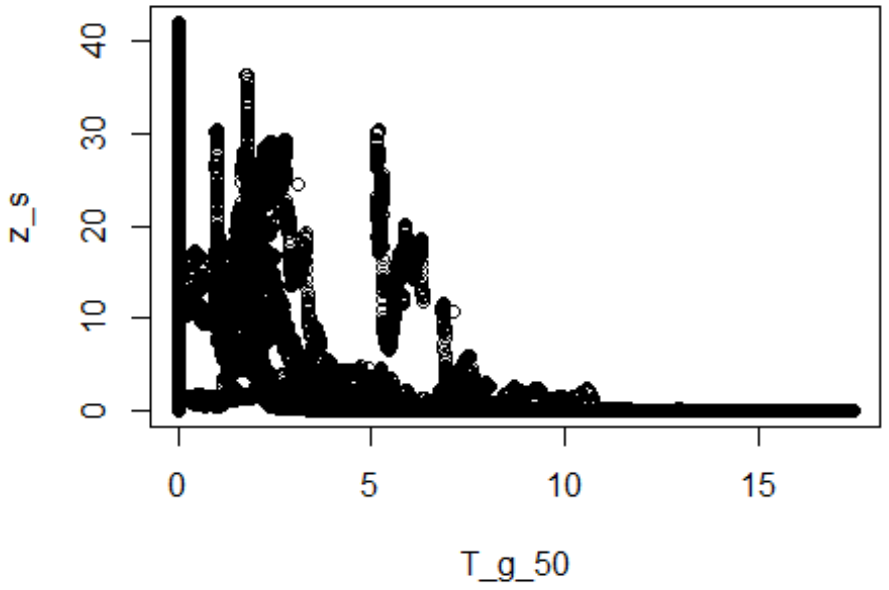

z_s
0
10
20
30
40
-5
0
5
10
15
20
T_a

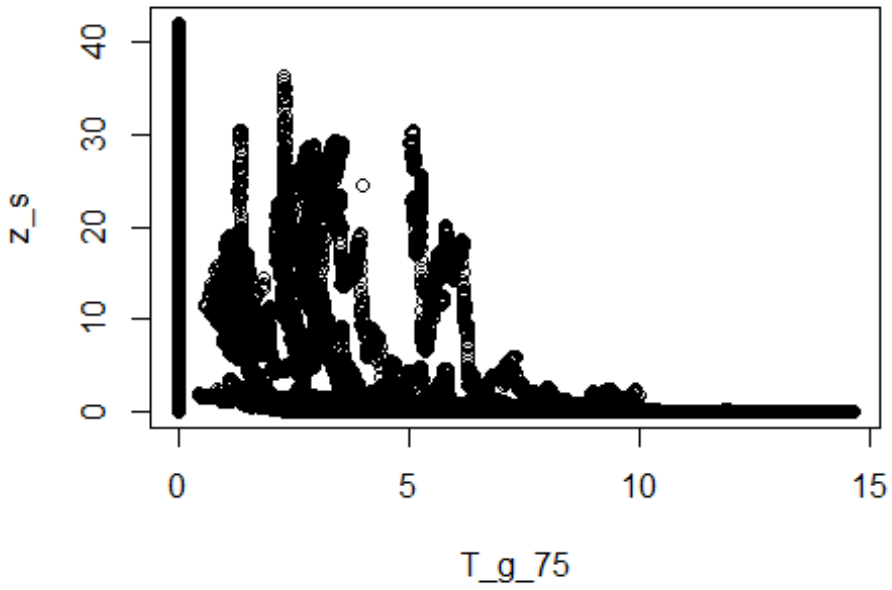

z_s
0
10
20
30
40
-10
-5
0
5
T_d

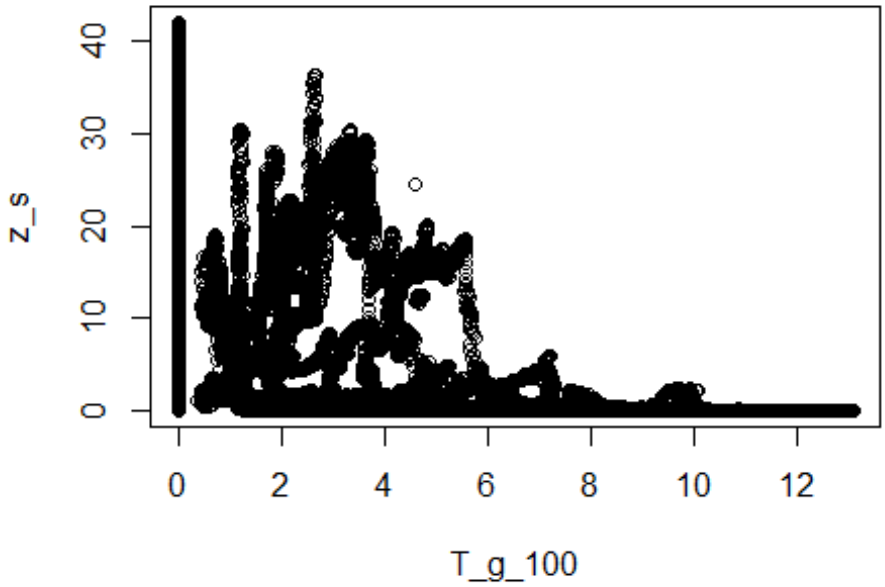

z_s
0
10
20
30
40
0
5
10
15
20
T_g_5

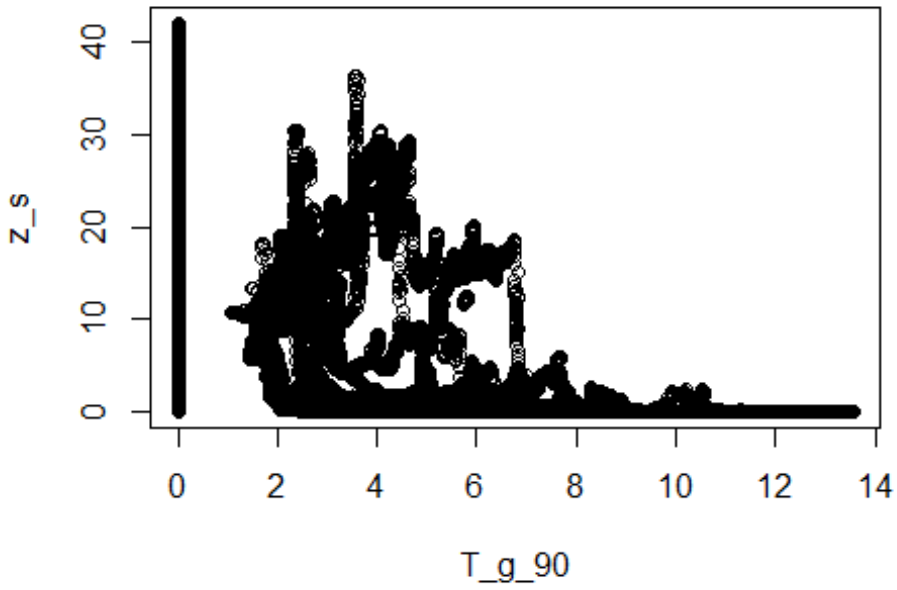

z_s
0
10
20
30
40
0
5
10
15
20
T_g_20

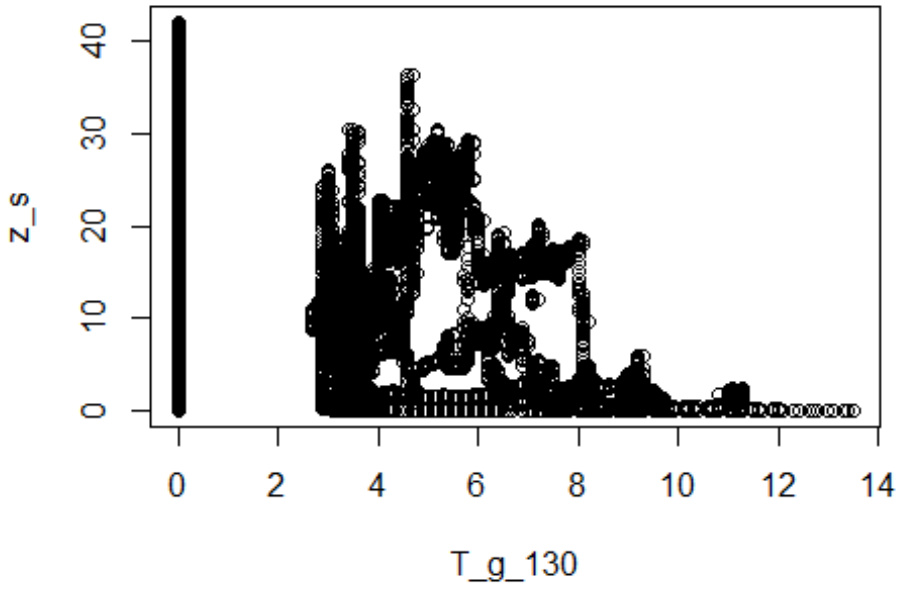

z_s
0
10
20
30
40
0
5
10
15
20
T_g_35

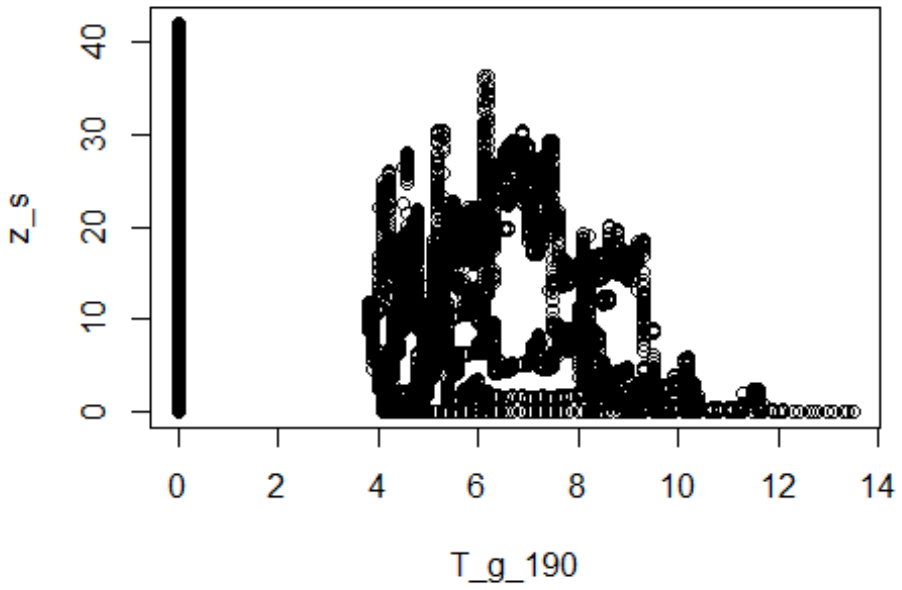

z_s
0
10
20
30
40
0
5
10
15
T_g_50

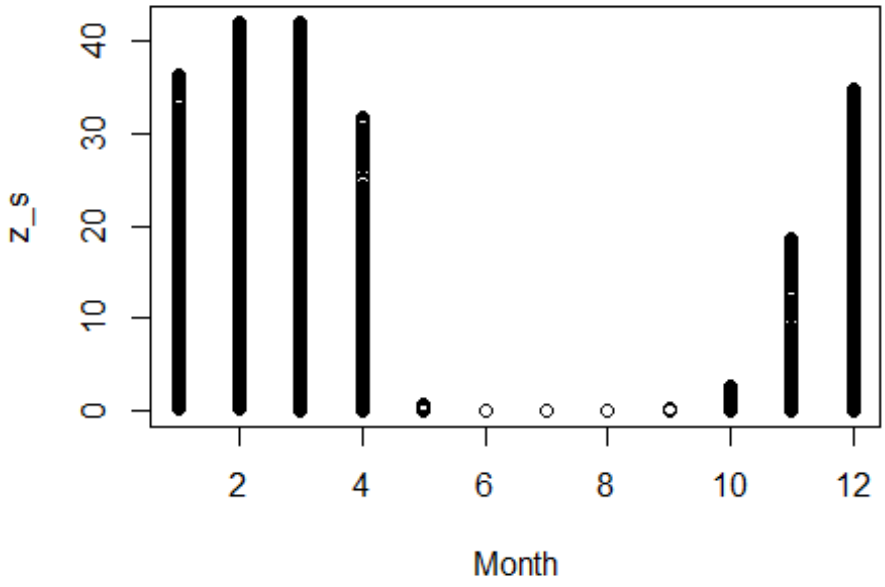

z_s
0
10
20
30
40
0
5
10
15
T_g_75

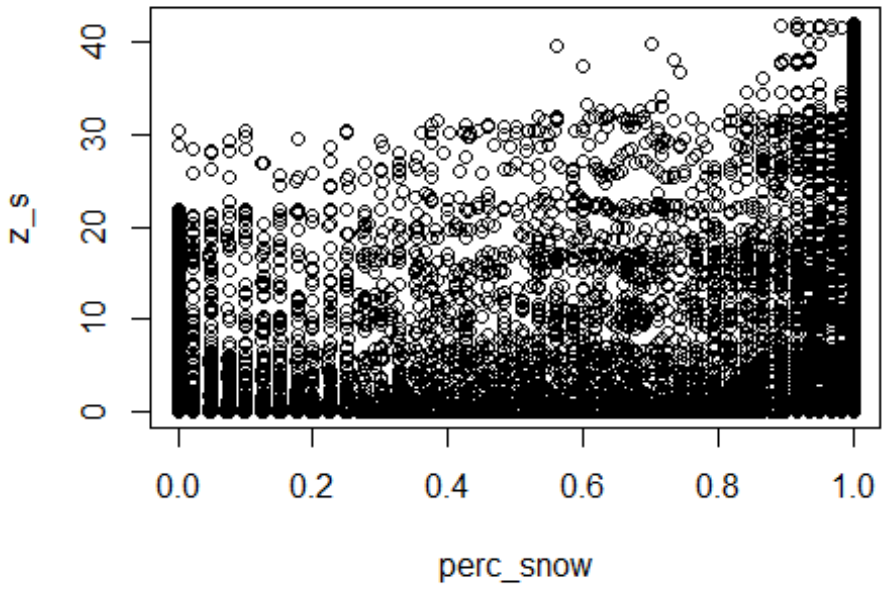

z_s
0
10
20
30
40
0
2
4
6
8
10
12
T_g_100

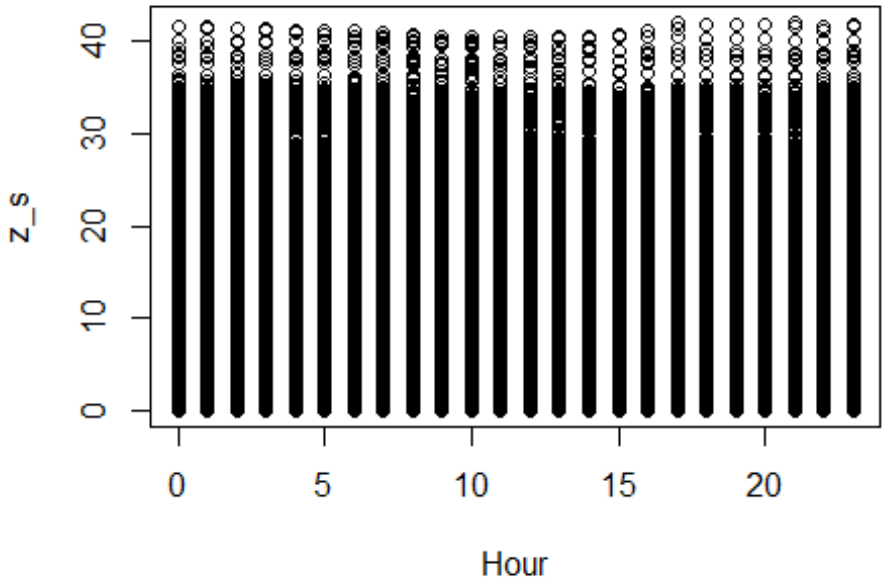

z_s
0
10
20
30
40
0
2
4
6
8
10
12
14
T_g_90

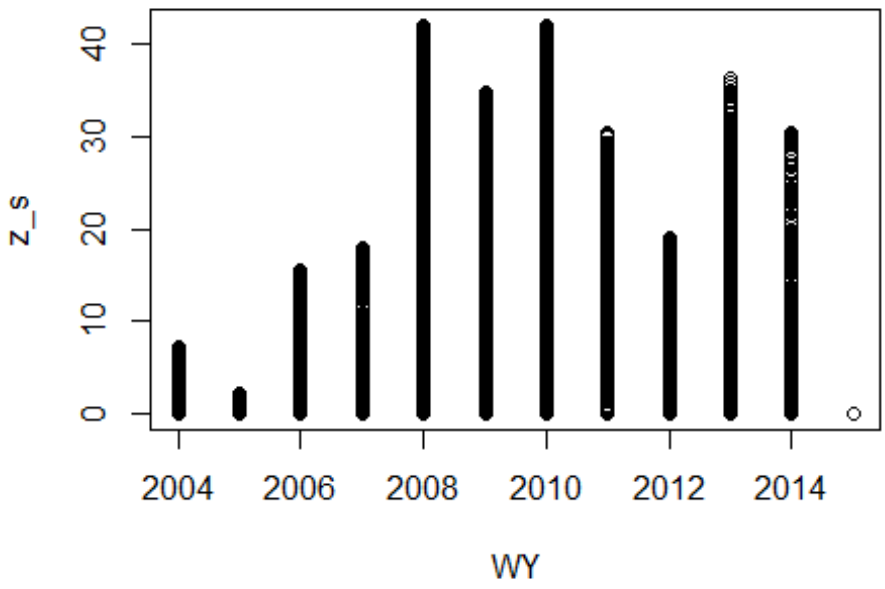

z_s
0
10
20
30
40
0
2
4
6
8
10
12
14
T_g_130


z_s
0
10
20
30
40
0
2
4
6
8
10
12
14
T_g_190


z_s
0
10
20
30
40
2
4
6
8
10
12
Month

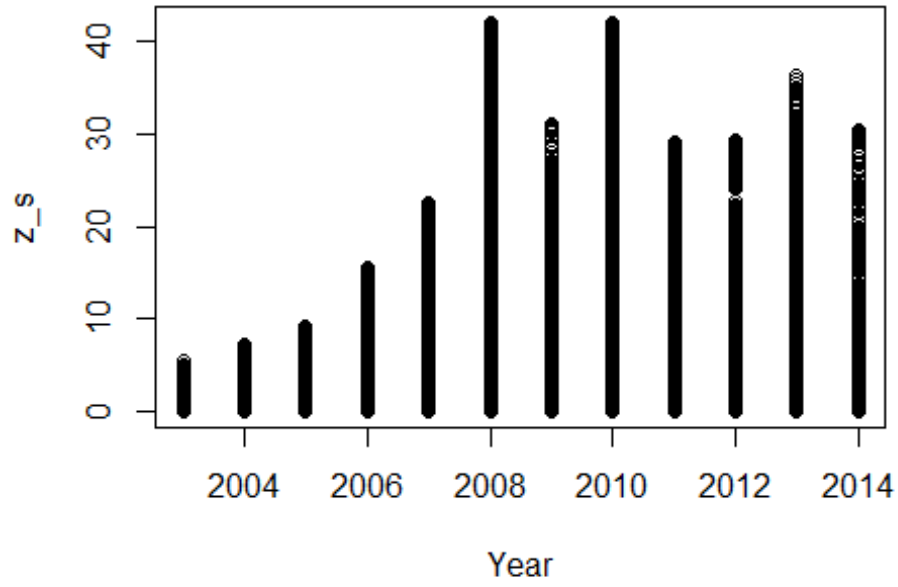

z_s
0
10
20
30
40
0.0
0.2
0.4
0.6
0.8
1.0
perc_snow



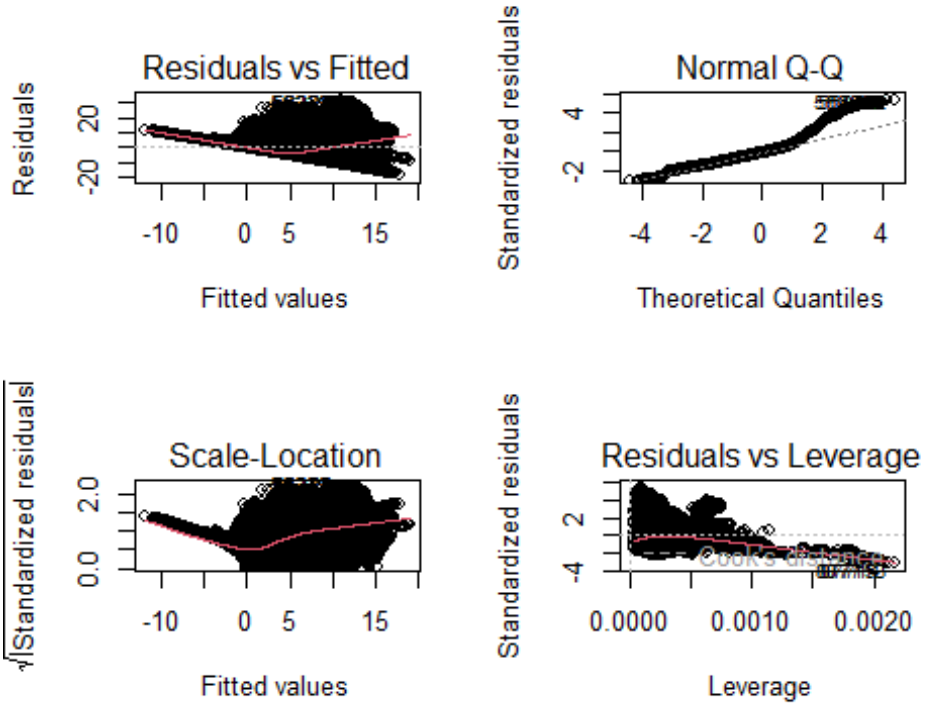



```
#plot(perc_snow ~ e_a, data = data)
abline(model_PCA_MLR_ALL_Merged_data)
```

```
## Warning in abline(model_PCA_MLR_ALL_Merged_data): only using the first two
of 20
## regression coefficients
```

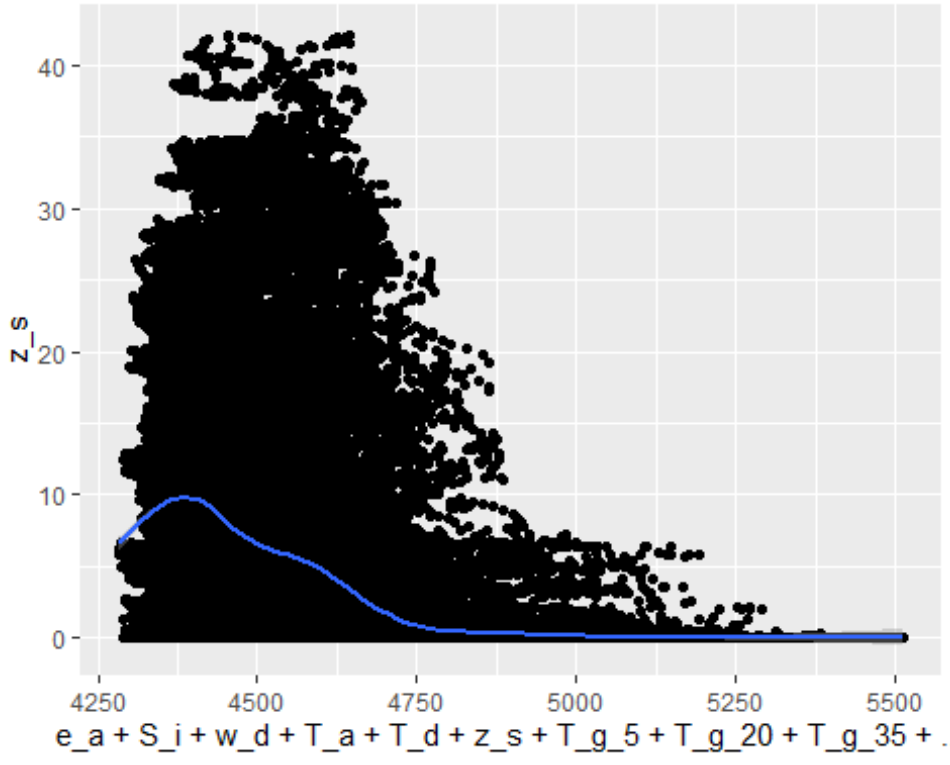



```
#Plotting residuals vs fitted to view the non linear relationship
par(mfrow = c(2, 2))
plot(model_PCA_MLR_ALL_Merged_data)
```



It can be observed from the above graphs that the relationship between selected features are non-linear.

```
#stat_smooth plot using ggplot to observe the relationships
#We can also observe the relation with single / multiple features
#ggplot(data, aes(x = e_a, y = z_s)) + geom_point() +
#stat_smooth()
ggplot(PCA_MLR_ALL_Merged_data, aes(x = e_a + S_i + w_d + T_a + T_d + z_s +
T_g_5 + T_g_20 + T_g_35 + T_g_50 + T_g_75 + T_g_100 + T_g_90 + T_g_130 +
T_g_190 + Month + Hour + WY + Year, y = z_s)) + geom_point() +
stat_smooth()

## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



In above graph it can be observed that plot fitt is doing fine be keeping the residuals low however it still need further optimization.