Dengue.R

dubey

Sat May 06 19:41:45 2017

```
#Dengue Analysis
#Created by Ratnam Dubey
#https://www.drivendata.org/competitions/44/dengai-predicting-disease-
spread/submissions/
# load libraries

pkgs <- c('tidyverse','scales' ,'corrplot', 'magrittr','corrplot' ,'zoo',
'RColorBrewer', 'gridExtra','MASS','plyr' , 'dplyr' ,'plotly' )
invisible(lapply(pkgs, require, character.only = T))
```

```
## Loading required package: tidyverse

## Warning in library(package, lib.loc = lib.loc, character.only = TRUE,
## logical.return = TRUE, : there is no package called 'tidyverse'

## Loading required package: scales

## Warning: package 'scales' was built under R version 3.3.3

## Loading required package: corrplot

## Warning: package 'corrplot' was built under R version 3.3.3

## Loading required package: magrittr

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.3.3

##
## Attaching package: 'zoo'

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

## Loading required package: RColorBrewer

## Loading required package: gridExtra

## Loading required package: MASS

## Loading required package: plyr

## Loading required package: dplyr
```

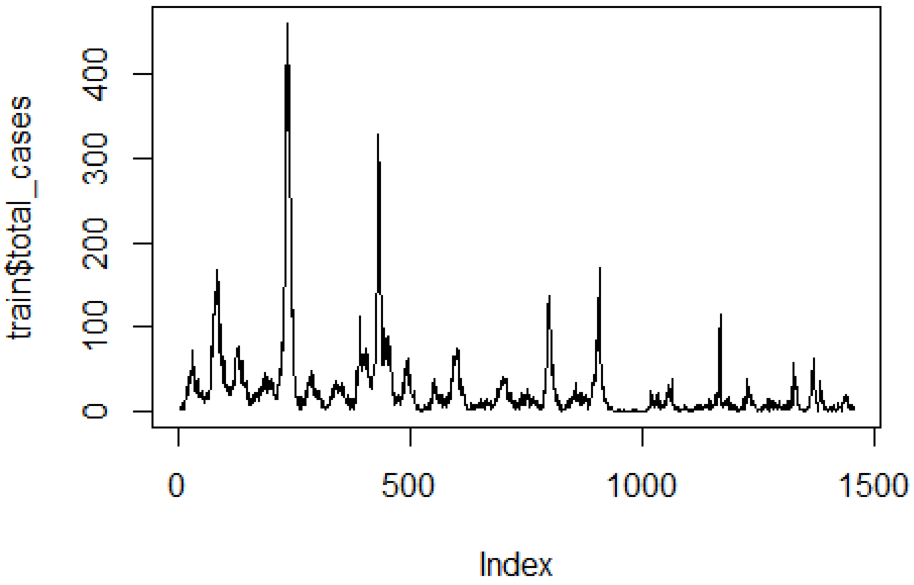
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following object is masked from 'package:MASS':
##
##     select

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following objects are masked from 'package:stats':
##
##     filter, lag

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

## Loading required package: plotly

## Warning: package 'plotly' was built under R version 3.3.3

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.3.3

##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following objects are masked from 'package:plyr':
##
##     arrange, mutate, rename, summarise

## The following object is masked from 'package:MASS':
##
##     select

## The following object is masked from 'package:stats':
##
##     filter
## The following object is masked from 'package:graphics':
##
##     layout
```

```
# Importing the Data
train <- read.table("D:\\Kaggle Projects\\Dengue\\dengue_features_train.csv",
header=TRUE, sep=",")
test <- read.table("D:\\Kaggle Projects\\Dengue\\dengue_features_test.csv",
header=TRUE, sep=",")

attach(train)
attach(test)
```

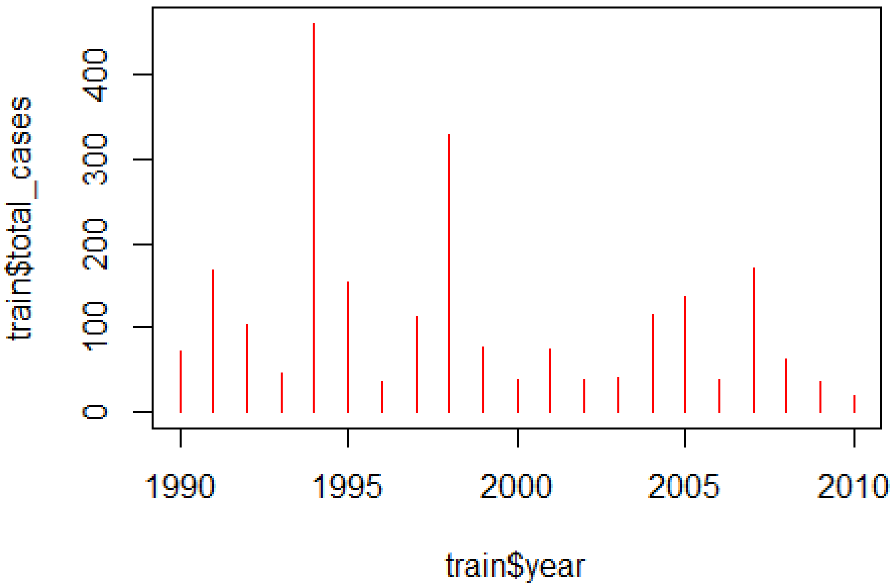
```
## The following objects are masked from train:
##
##     city, ndvi_ne, ndvi_nw, ndvi_se, ndvi_sw,
##     precipitation_amt_mm, reanalysis_air_temp_k,
##     reanalysis_avg_temp_k, reanalysis_dew_point_temp_k,
##     reanalysis_max_air_temp_k, reanalysis_min_air_temp_k,
##     reanalysis_precip_amt_kg_per_m2,
##     reanalysis_relative_humidity_percent,
##     reanalysis_sat_precip_amt_mm,
##     reanalysis_specific_humidity_g_per_kg, reanalysis_tdtr_k,
##     station_avg_temp_c, station_diur_temp_rng_c,
##     station_max_temp_c, station_min_temp_c, station_precip_mm,
##     week_start_date, weekofyear, year
```

```
#Exploring the Data
head(train,5)
```

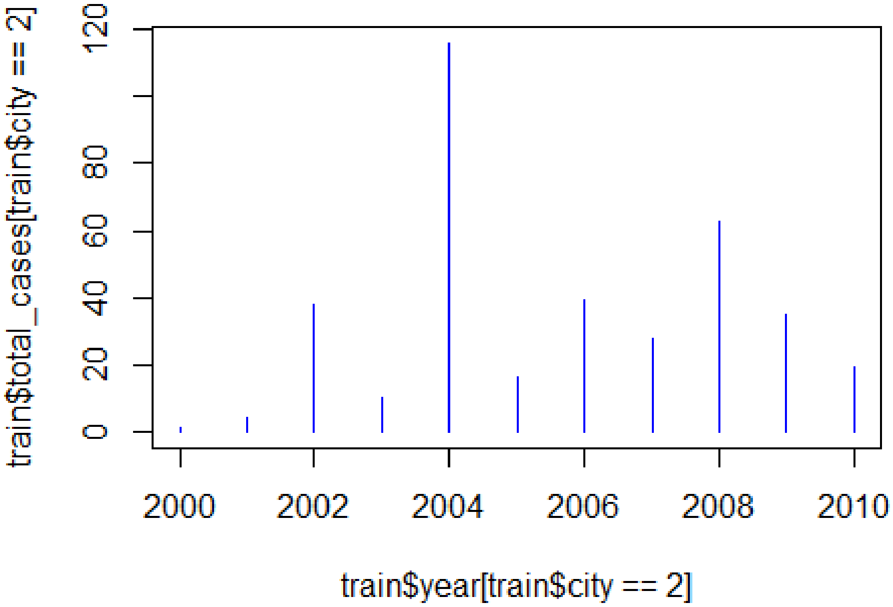
```
##   city year weekofyear week_start_date   ndvi_ne   ndvi_nw   ndvi_se
## 1    1 1990         18       4/30/1990 0.1226000 0.1037250 0.1984833
## 2    1 1990         19        5/7/1990 0.1699000 0.1421750 0.1623571
## 3    1 1990         20       5/14/1990 0.0322500 0.1729667 0.1572000
## 4    1 1990         21       5/21/1990 0.1286333 0.2450667 0.2275571
## 5    1 1990         22       5/28/1990 0.1962000 0.2622000 0.2512000
##     ndvi_sw precipitation_amt_mm reanalysis_air_temp_k
## 1 0.1776167                12.42              297.5729
## 2 0.1554857                22.82              298.2114
## 3 0.1708429                34.54              298.7814
## 4 0.2358857                15.36              298.9871
## 5 0.2473400                 7.52              299.5186
##   reanalysis_avg_temp_k reanalysis_dew_point_temp_k
## 1              297.7429                    292.4143
## 2              298.4429                    293.9514
## 3              298.8786                    295.4343
## 4              299.2286                    295.3100
## 5              299.6643                    295.8214
##   reanalysis_max_air_temp_k reanalysis_min_air_temp_k
## 1                     299.8                     295.9
## 2                     300.9                     296.4
## 3                     300.5                     297.3
## 4                     301.4                     297.0
## 5                     301.9                     297.5
##   reanalysis_precip_amt_kg_per_m2 reanalysis_relative_humidity_percent
## 1                           32.00                             73.36571
## 2                           17.94                             77.36857
## 3                           26.10                             82.05286
## 4                           13.90                             80.33714
## 5                           12.20                             80.46000
##   reanalysis_sat_precip_amt_mm reanalysis_specific_humidity_g_per_kg
## 1                        12.42                              14.01286
## 2                        22.82                              15.37286
## 3                        34.54                              16.84857
## 4                        15.36                              16.67286
## 5                         7.52                              17.21000
##   reanalysis_tdtr_k station_avg_temp_c station_diur_temp_rng_c
## 1          2.628571           25.44286                6.900000
## 2          2.371429           26.71429                6.371429
## 3          2.300000           26.71429                6.485714
## 4          2.428571           27.47143                6.771429
## 5          3.014286           28.94286                9.371429
##   station_max_temp_c station_min_temp_c station_precip_mm total_cases
## 1               29.4               20.0              16.0           4
## 2               31.7               22.2               8.6           5
## 3               32.2               22.8              41.4           4
## 4               33.3               23.3               4.0           3
## 5               35.0               23.9               5.8           6

#getting the Data over the Time
aggdata <-aggregate(train$city, by=list(train$year,train$city),FUN=mean,
na.rm=TRUE)

#Conclusion here is that the Sj has the Data from 1990 to 2008
#where as the iq has the Data from 2000 to 2010
#plotting the Data based on the Number of Cases


plot(train$total_cases,type="l")
```



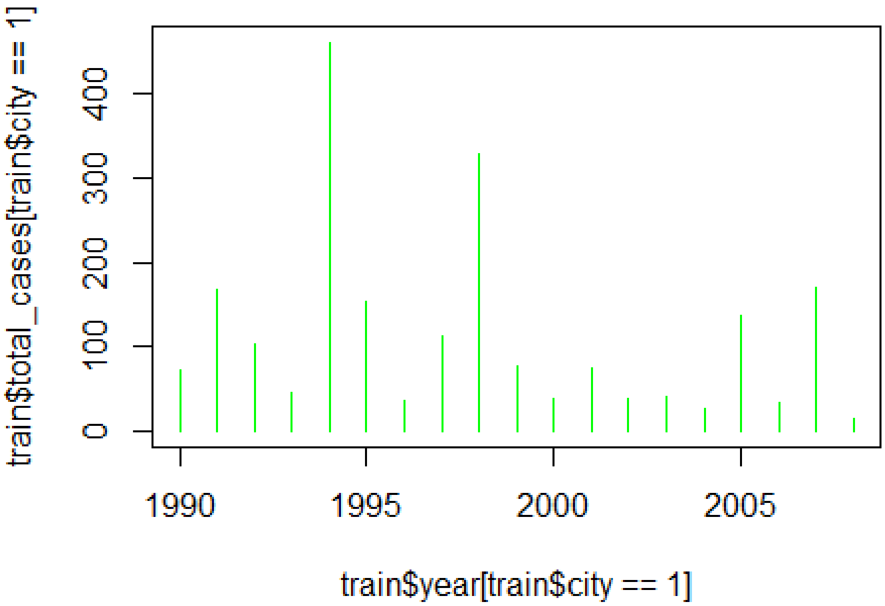
```
plot(train$year,train$total_cases,type = "h" , col="red")
```



```r
# As we can see the Cases are distributed over the year
# But we dont know for which city it is for as we have two city

unique(train$city)

## [1] 1 2

#two cities are Sj = 1 and iq = 2

plot(train$year[train$city==2],train$total_cases[train$city==2],type="h" ,
col="blue")
```

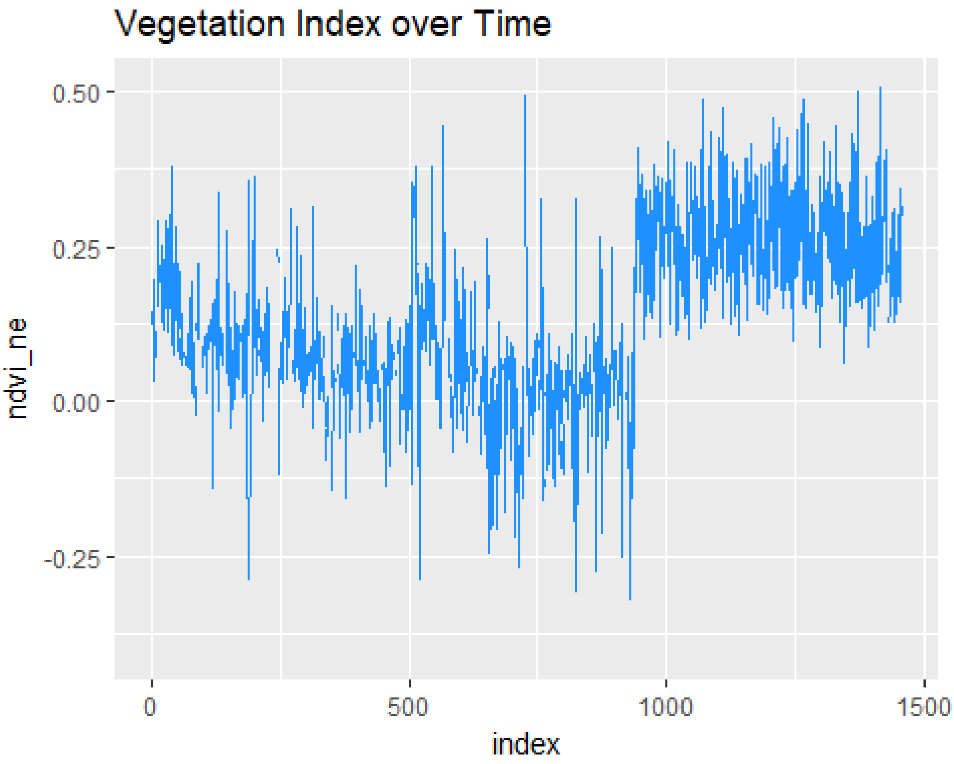


```r
plot(train$year[train$city==1],train$total_cases[train$city==1],type="h" ,
col="green")
```

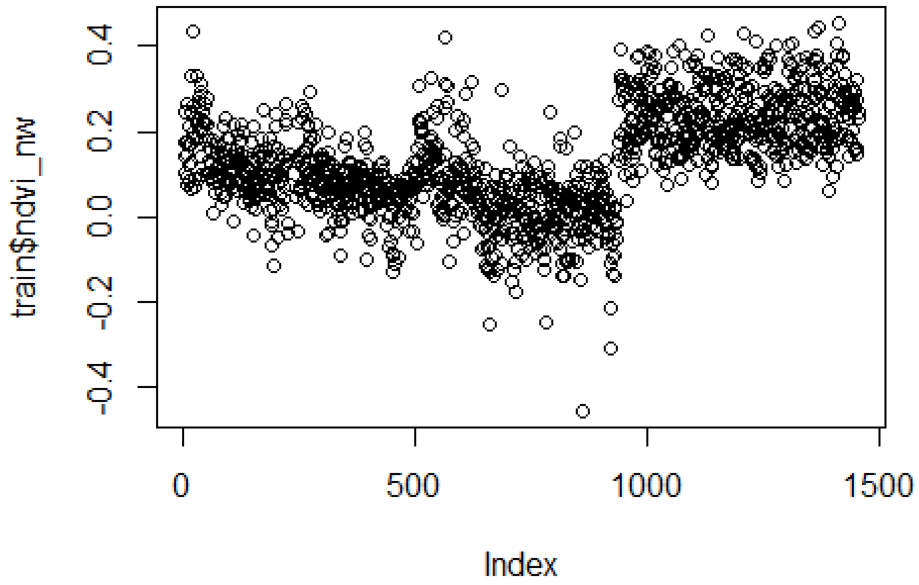


```
# Conclusion
# For City Sj the Maximum number of cases are in 1994
# For City iq the Maximum number of cases are in  2004

# count missing values (as percent)
apply(train, 2, function(x)
  round(100 * (length(which(is.na(x))))/length(x) , digits = 1)) %>%
  as.data.frame() %>%
  `names<-`('Percent of Missing Values')

##                                       Percent of Missing Values
## city                                                        0.0
## year                                                        0.0
## weekofyear                                                  0.0
## week_start_date                                             0.0
## ndvi_ne                                                    13.3
## ndvi_nw                                                     3.6
## ndvi_se                                                     1.5
## ndvi_sw                                                     1.5
## precipitation_amt_mm                                        0.9
## reanalysis_air_temp_k                                       0.7
## reanalysis_avg_temp_k                                       0.7
## reanalysis_dew_point_temp_k                                 0.7
## reanalysis_max_air_temp_k                                   0.7
## reanalysis_min_air_temp_k                                   0.7
## reanalysis_precip_amt_kg_per_m2                             0.7
## reanalysis_relative_humidity_percent                        0.7
## reanalysis_sat_precip_amt_mm                                0.9
## reanalysis_specific_humidity_g_per_kg                       0.7
## reanalysis_tdtr_k                                           0.7
## station_avg_temp_c                                          3.0
## station_diur_temp_rng_c                                     3.0
## station_max_temp_c                                          1.4
## station_min_temp_c                                          1.0
## station_precip_mm                                           1.5
## total_cases                                                 0.0
```

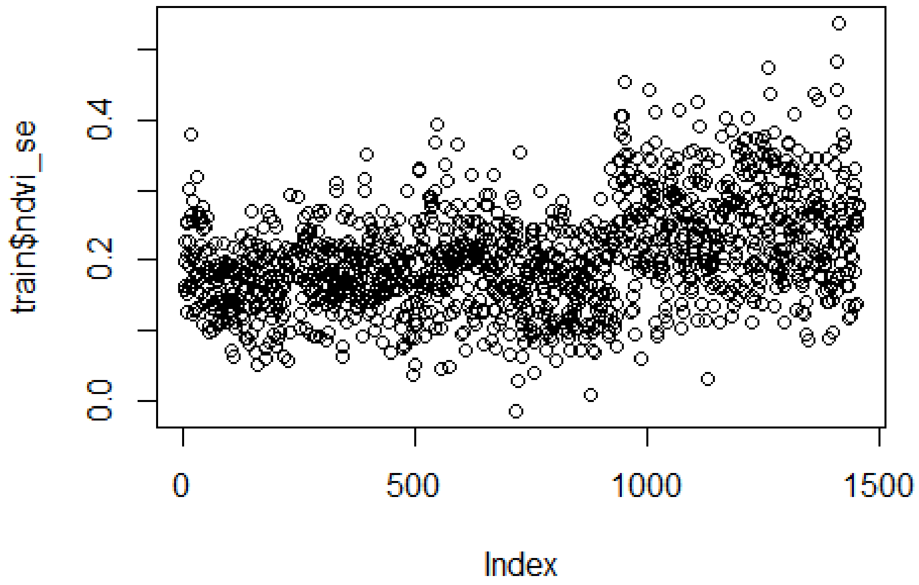
```
# Plotting the Data
train %>%
  mutate(index = as.numeric(row.names(.))) %>%
  ggplot(aes(index, ndvi_ne)) +
  geom_line(colour = 'dodgerblue') +
  ggtitle("Vegetation Index over Time")
```



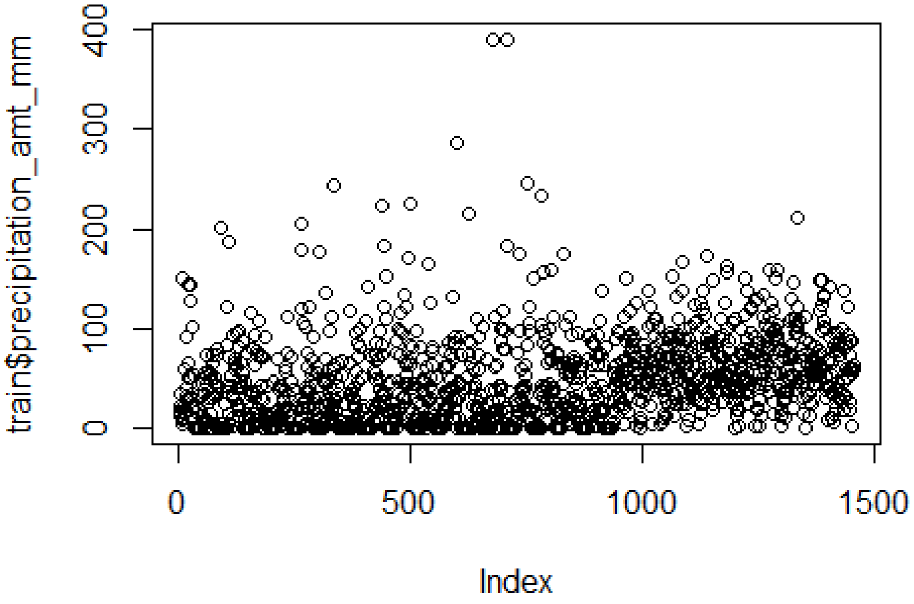
```
# Droping the Coloum with 13% of the missing Data

train$ndvi_ne <- NULL
test$ndvi_ne <- NULL

# Replacing the Values with the Mean
plot(train$ndvi_nw)
```



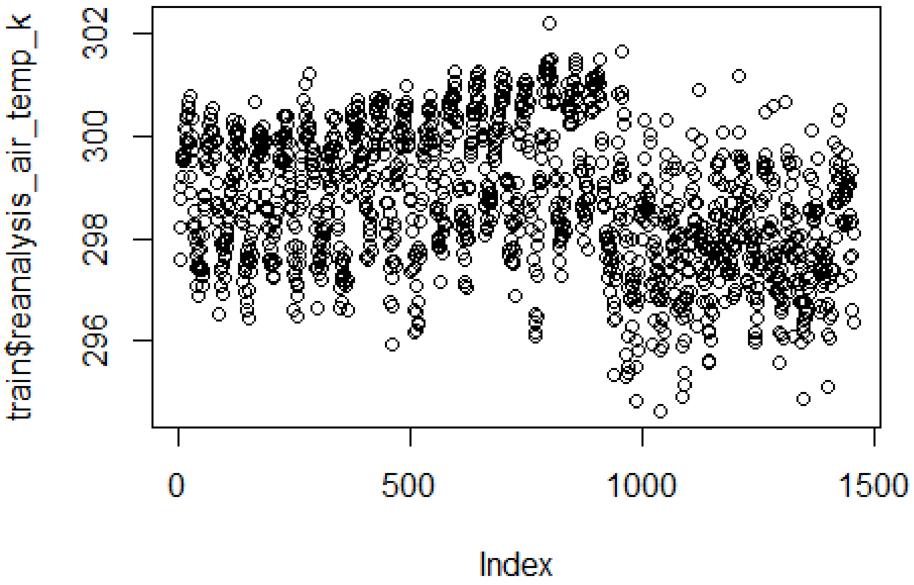
```
mean(train$ndvi_nw,na.rm = TRUE)

## [1] 0.1305526

train$ndvi_nw[is.na(train$ndvi_nw)] <- mean(train$ndvi_nw,na.rm = TRUE)

plot(train$ndvi_se)
```



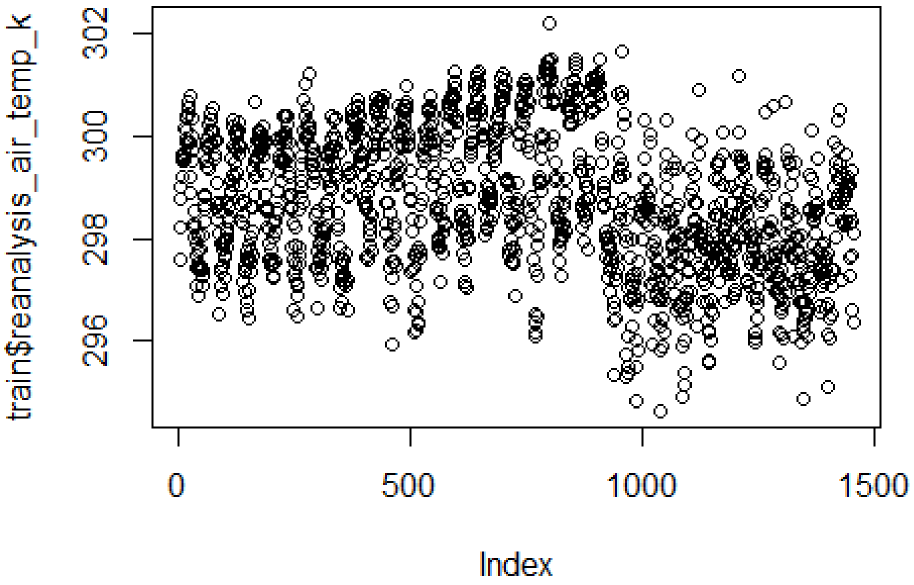
```
mean(train$ndvi_se,na.rm = TRUE)

## [1] 0.2037832

train$ndvi_se[is.na(train$ndvi_se)] <- mean(train$ndvi_se,na.rm = TRUE)


plot(train$precipitation_amt_mm)
```



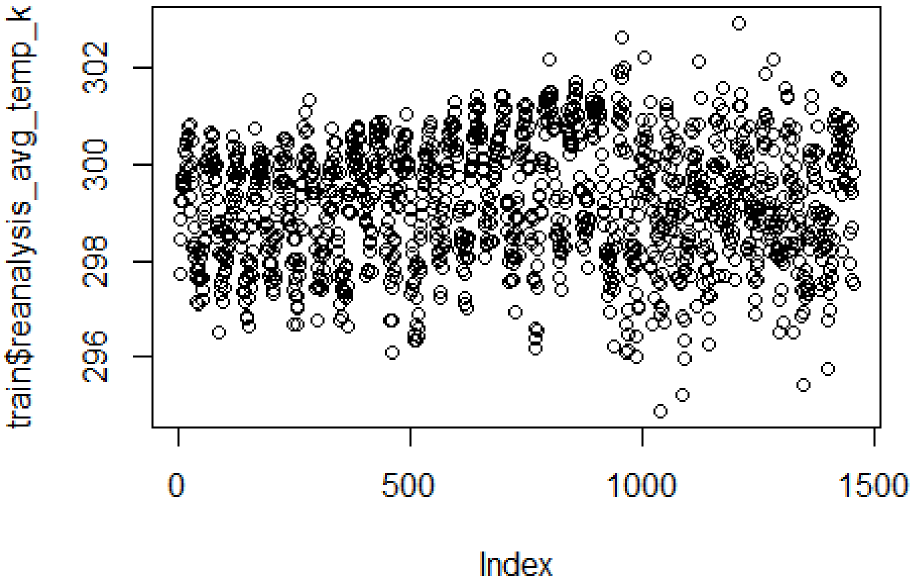
```
mean(train$precipitation_amt_mm,na.rm = TRUE)

## [1] 45.76039

train$precipitation_amt_mm[is.na(train$precipitation_amt_mm)] <-
mean(train$precipitation_amt_mm,na.rm = TRUE)

plot(train$reanalysis_air_temp_k)
```



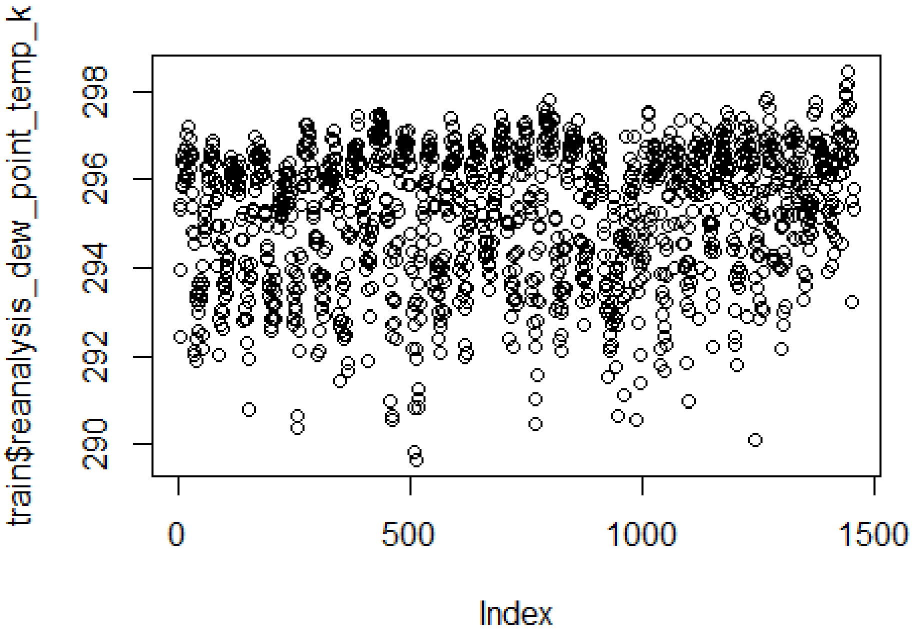
```
mean(train$reanalysis_air_temp_k,na.rm = TRUE)

## [1] 298.7019

train$reanalysis_air_temp_k[is.na(train$reanalysis_air_temp_k)] <-
mean(train$reanalysis_air_temp_k,na.rm = TRUE)

plot(train$reanalysis_air_temp_k)
```



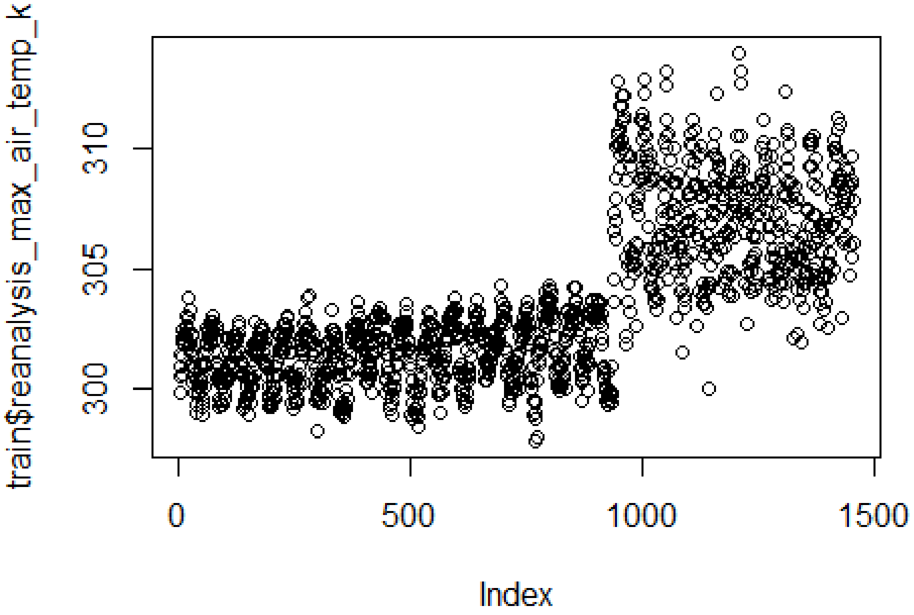
```
mean(train$reanalysis_air_temp_k,na.rm = TRUE)

## [1] 298.7019

train$reanalysis_air_temp_k[is.na(train$reanalysis_air_temp_k)] <-
mean(train$reanalysis_air_temp_k,na.rm = TRUE)

plot(train$reanalysis_avg_temp_k)
```



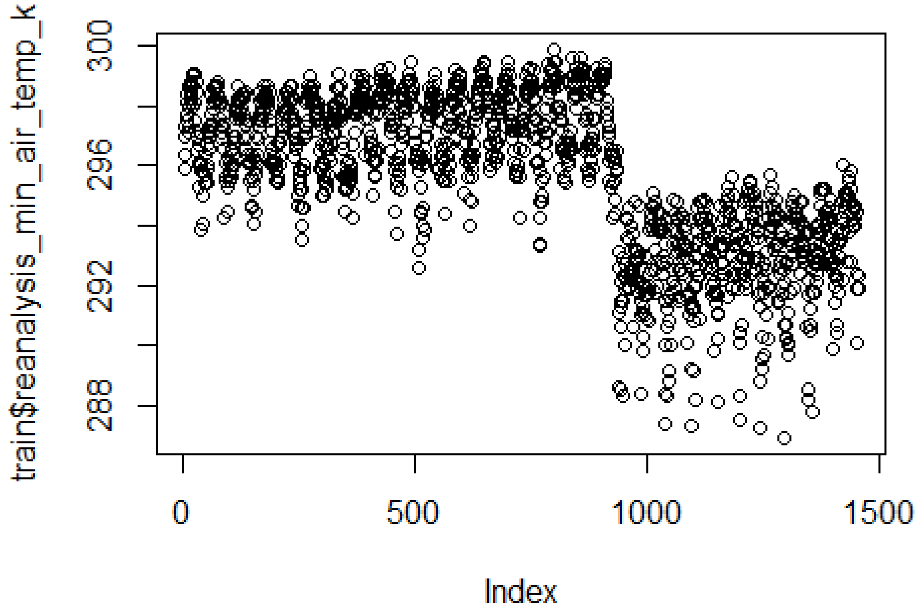
```
mean(train$reanalysis_avg_temp_k,na.rm = TRUE)

## [1] 299.2256

train$reanalysis_avg_temp_k[is.na(train$reanalysis_avg_temp_k)] <-
mean(train$reanalysis_avg_temp_k,na.rm = TRUE)

plot(train$reanalysis_dew_point_temp_k)
```



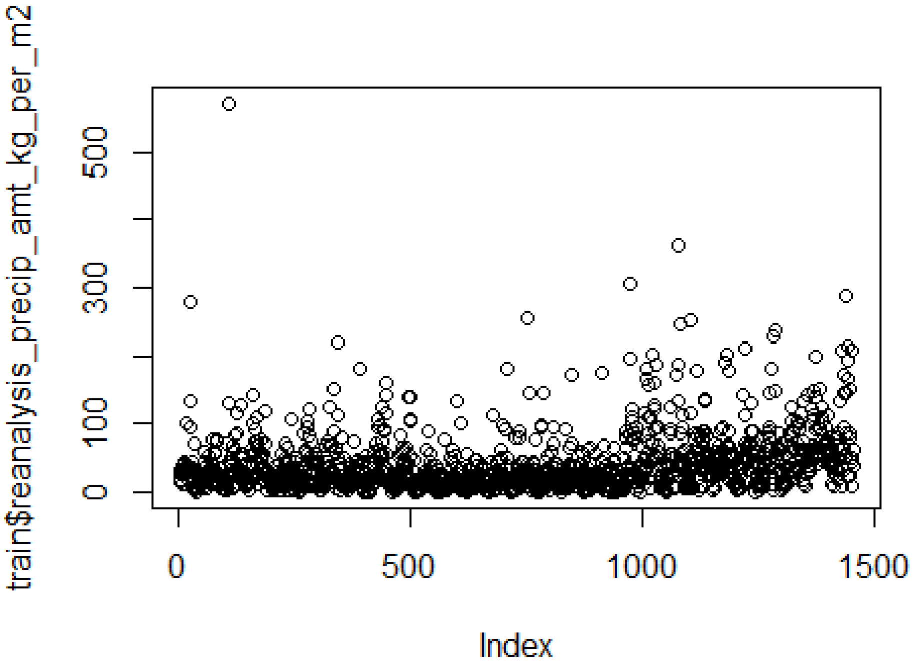
```
mean(train$reanalysis_dew_point_temp_k,na.rm = TRUE)

## [1] 295.2464

train$reanalysis_dew_point_temp_k[is.na(train$reanalysis_dew_point_temp_k)]
<- mean(train$reanalysis_dew_point_temp_k,na.rm = TRUE)

plot(train$reanalysis_max_air_temp_k)
```



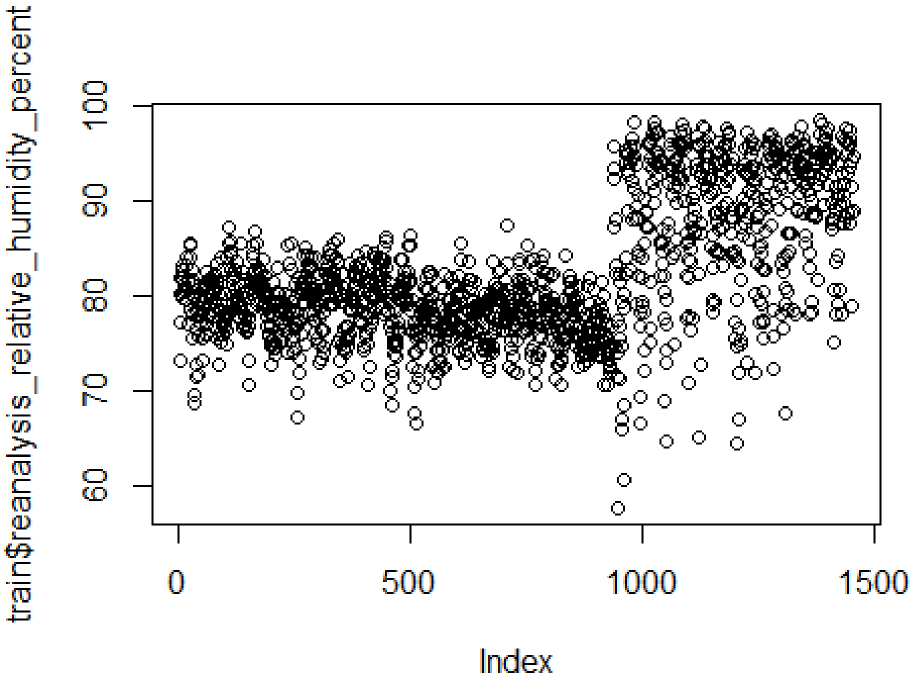
```
mean(train$reanalysis_max_air_temp_k,na.rm = TRUE)

## [1] 303.4271

train$reanalysis_max_air_temp_k[is.na(train$reanalysis_max_air_temp_k)] <-
mean(train$reanalysis_max_air_temp_k,na.rm = TRUE)

plot(train$reanalysis_min_air_temp_k)
```



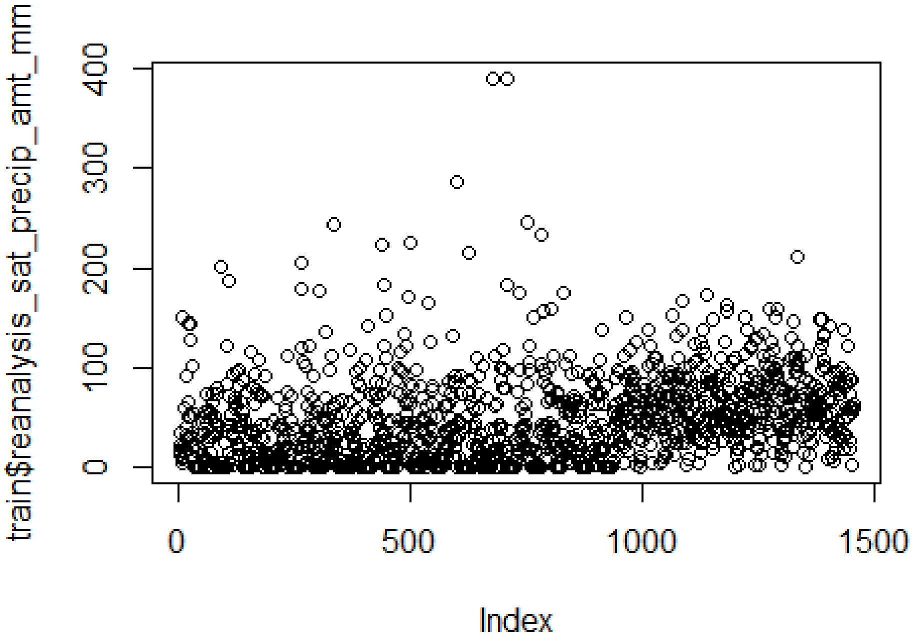
```
mean(train$reanalysis_min_air_temp_k,na.rm = TRUE)

## [1] 295.7192

train$reanalysis_min_air_temp_k[is.na(train$reanalysis_min_air_temp_k)] <-
mean(train$reanalysis_min_air_temp_k,na.rm = TRUE)

plot(train$reanalysis_precip_amt_kg_per_m2)
```



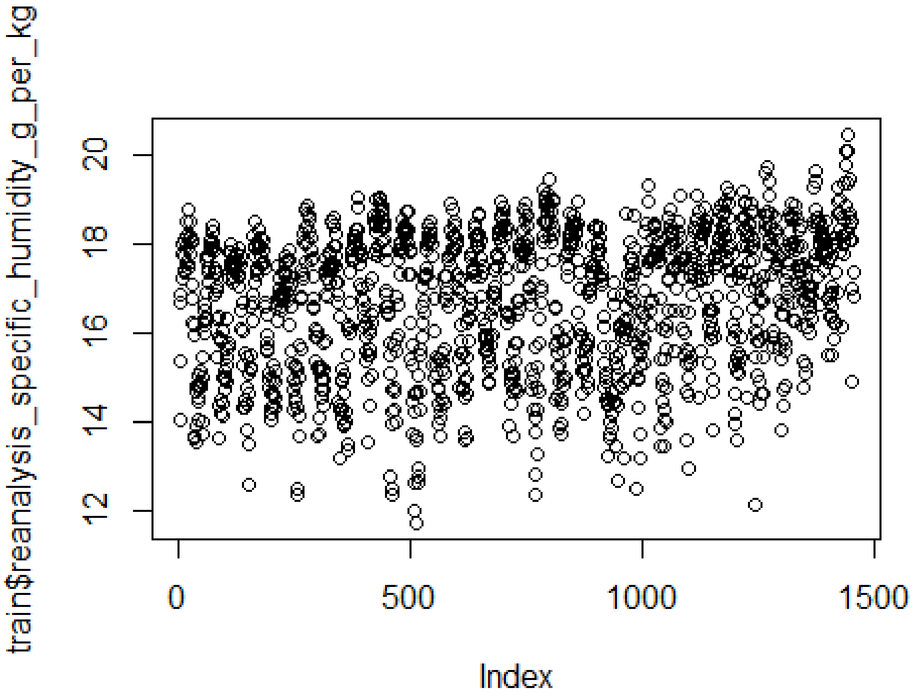
```
mean(train$reanalysis_precip_amt_kg_per_m2,na.rm = TRUE)

## [1] 40.15182

train$reanalysis_precip_amt_kg_per_m2[is.na(train$reanalysis_precip_amt_kg_pe
r_m2)] <- mean(train$reanalysis_precip_amt_kg_per_m2,na.rm = TRUE)

plot(train$reanalysis_relative_humidity_percent)
```



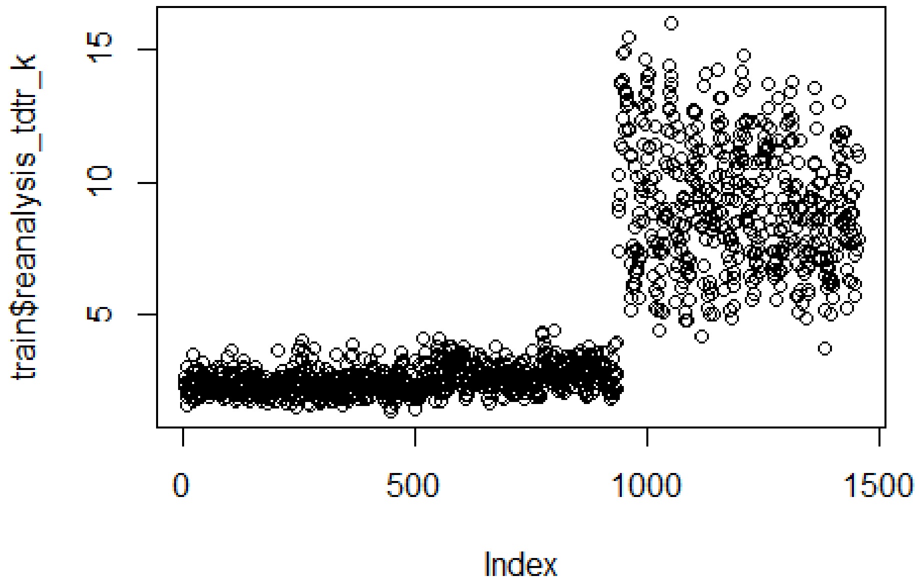
```
mean(train$reanalysis_relative_humidity_percent,na.rm = TRUE)

## [1] 82.16196

train$reanalysis_relative_humidity_percent[is.na(train$reanalysis_relative_humidity_percent)] <- mean(train$reanalysis_relative_humidity_percent,na.rm = TRUE)

plot(train$reanalysis_sat_precip_amt_mm)
```



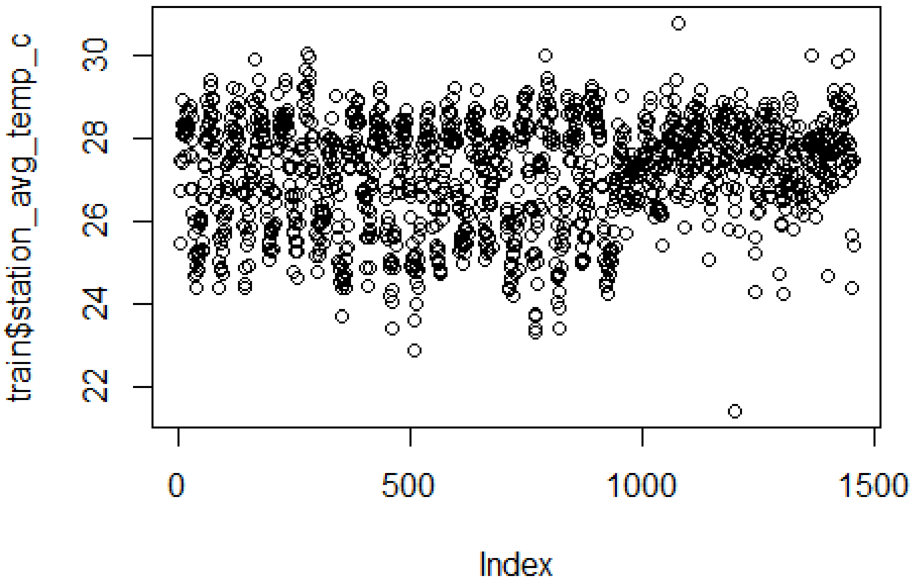
```
mean(train$reanalysis_sat_precip_amt_mm,na.rm = TRUE)

## [1] 45.76039

train$reanalysis_sat_precip_amt_mm[is.na(train$reanalysis_sat_precip_amt_mm)]
<- mean(train$reanalysis_sat_precip_amt_mm,na.rm = TRUE)

plot(train$reanalysis_specific_humidity_g_per_kg)
```



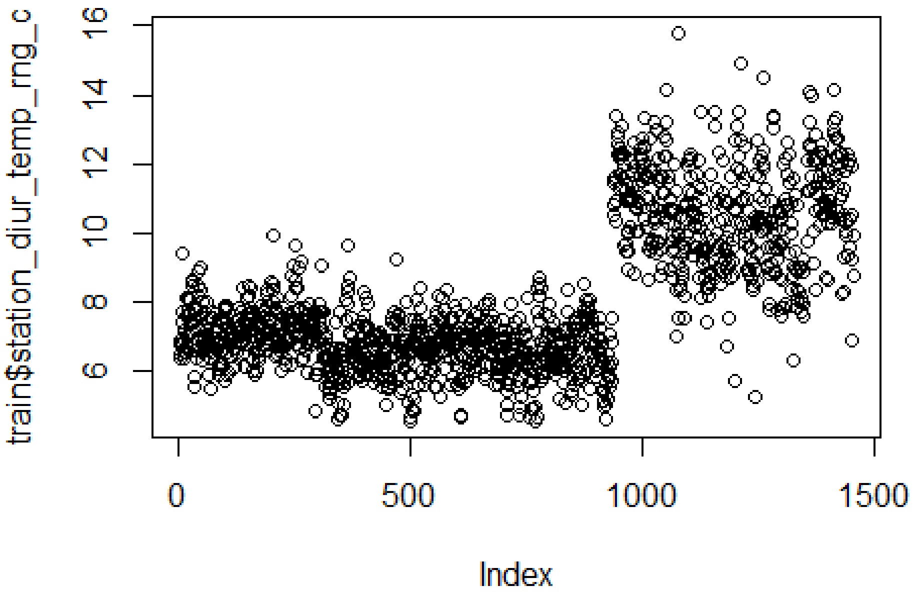
```
mean(train$reanalysis_specific_humidity_g_per_kg,na.rm = TRUE)

## [1] 16.74643

train$reanalysis_specific_humidity_g_per_kg[is.na(train$reanalysis_specific_h
umidity_g_per_kg)] <- mean(train$reanalysis_specific_humidity_g_per_kg,na.rm
= TRUE)

plot(train$reanalysis_tdtr_k)
```



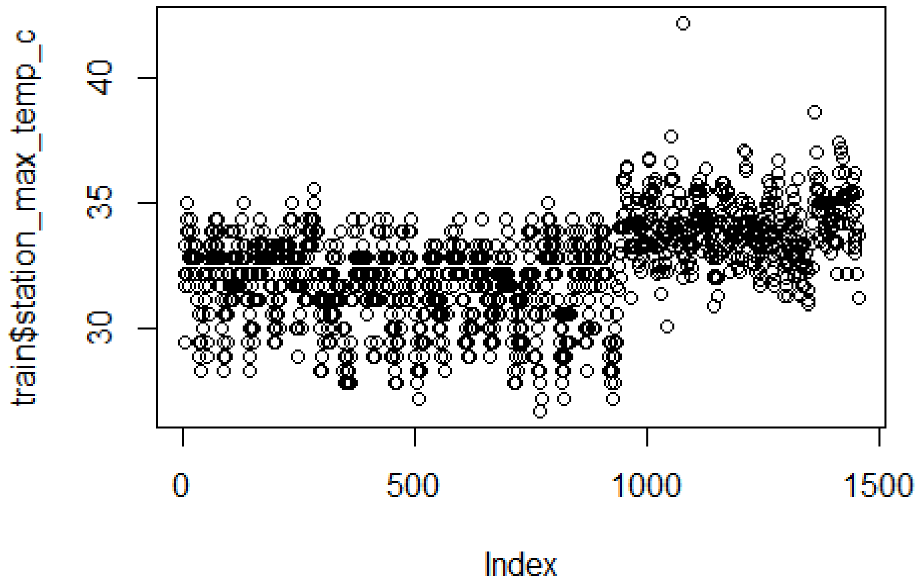
```
mean(train$reanalysis_tdtr_k,na.rm = TRUE)

## [1] 4.903754

train$reanalysis_tdtr_k[is.na(train$reanalysis_tdtr_k)] <-
mean(train$reanalysis_tdtr_k,na.rm = TRUE)

plot(train$station_avg_temp_c)
```



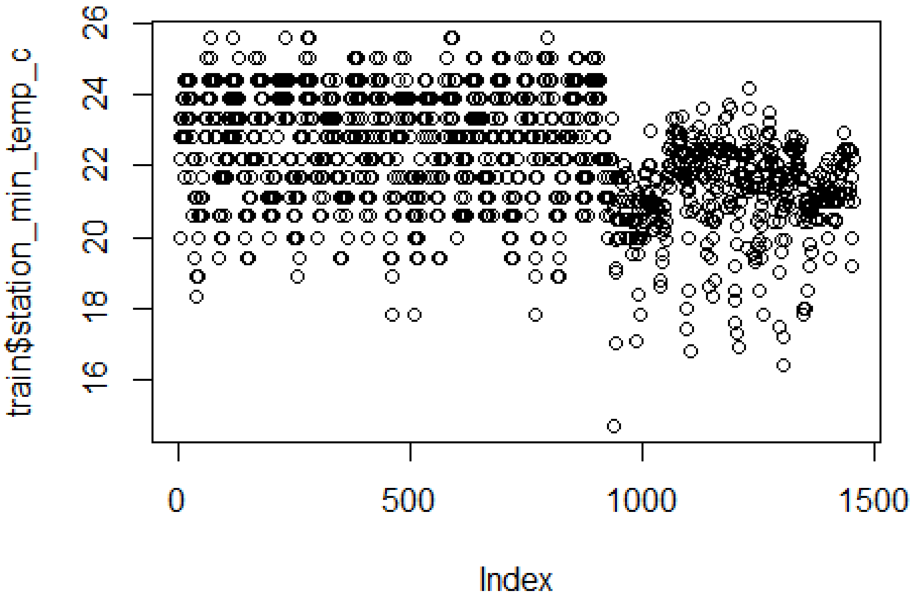
```
mean(train$station_avg_temp_c,na.rm = TRUE)

## [1] 27.18578

train$station_avg_temp_c[is.na(train$station_avg_temp_c)] <-
mean(train$station_avg_temp_c,na.rm = TRUE)

plot(train$station_diur_temp_rng_c)
```



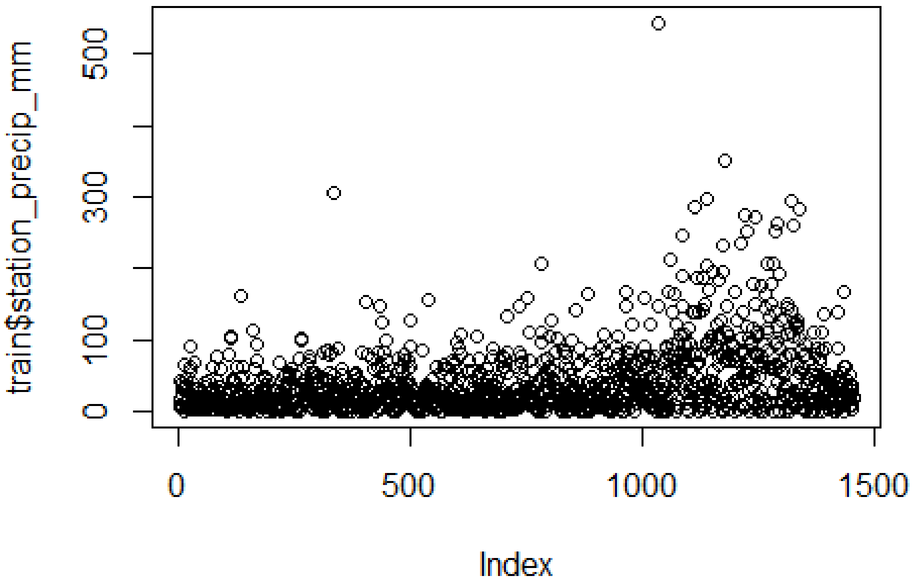
```
mean(train$station_diur_temp_rng_c,na.rm = TRUE)

## [1] 8.059328

train$station_diur_temp_rng_c[is.na(train$station_diur_temp_rng_c)] <-
mean(train$station_diur_temp_rng_c,na.rm = TRUE)

plot(train$station_max_temp_c)
```



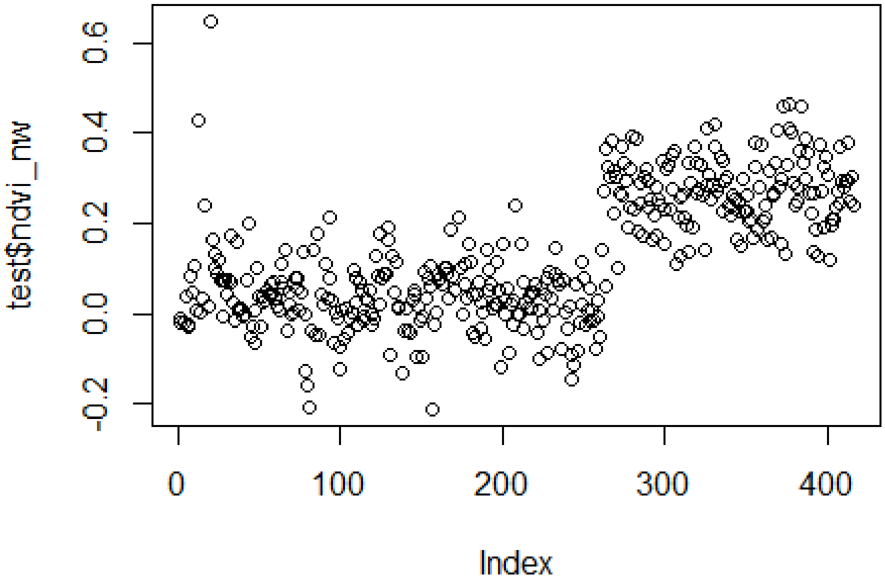
```
mean(train$station_max_temp_c,na.rm = TRUE)

## [1] 32.45244

train$station_max_temp_c[is.na(train$station_max_temp_c)] <-
mean(train$station_max_temp_c,na.rm = TRUE)

plot(train$station_min_temp_c)
```



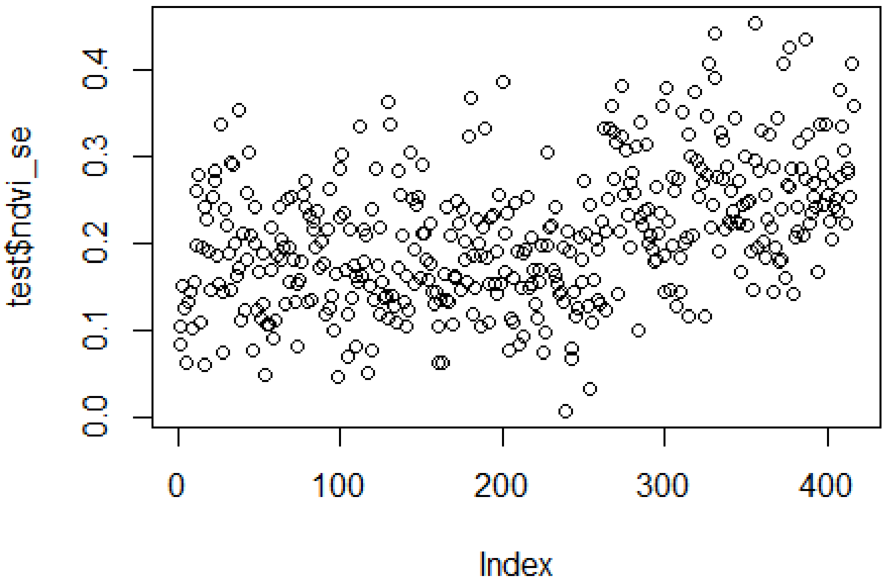
```
mean(train$station_min_temp_c,na.rm = TRUE)

## [1] 22.10215

train$station_min_temp_c[is.na(train$station_min_temp_c)] <-
mean(train$station_min_temp_c,na.rm = TRUE)

plot(train$station_precip_mm)
```



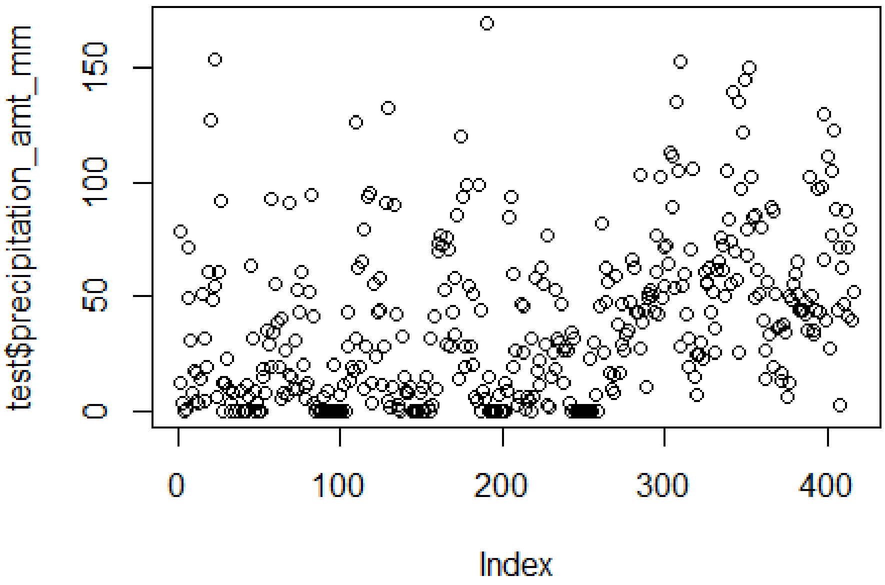
```
mean(train$station_precip_mm,na.rm = TRUE)

## [1] 39.32636

train$station_precip_mm[is.na(train$station_precip_mm)] <-
mean(train$station_precip_mm,na.rm = TRUE)


# count missing values (as percent)
apply(train, 2, function(x)
  round(100 * (length(which(is.na(x))))/length(x) , digits = 1)) %>%
  as.data.frame() %>%
  `names<-`('Percent of Missing Values')

##                                       Percent of Missing Values
## city                                                        0.0
## year                                                        0.0
## weekofyear                                                  0.0
## week_start_date                                             0.0
## ndvi_nw                                                     0.0
## ndvi_se                                                     0.0
## ndvi_sw                                                     1.5
## precipitation_amt_mm                                        0.0
## reanalysis_air_temp_k                                       0.0
## reanalysis_avg_temp_k                                       0.0
## reanalysis_dew_point_temp_k                                 0.0
## reanalysis_max_air_temp_k                                   0.0
## reanalysis_min_air_temp_k                                   0.0
## reanalysis_precip_amt_kg_per_m2                             0.0
## reanalysis_relative_humidity_percent                        0.0
## reanalysis_sat_precip_amt_mm                                0.0
## reanalysis_specific_humidity_g_per_kg                       0.0
## reanalysis_tdtr_k                                           0.0
## station_avg_temp_c                                          0.0
## station_diur_temp_rng_c                                     0.0
## station_max_temp_c                                          0.0
## station_min_temp_c                                          0.0
## station_precip_mm                                           0.0
## total_cases                                                 0.0

# Replacing the Values with the Mean of Test

plot(test$ndvi_nw)
```



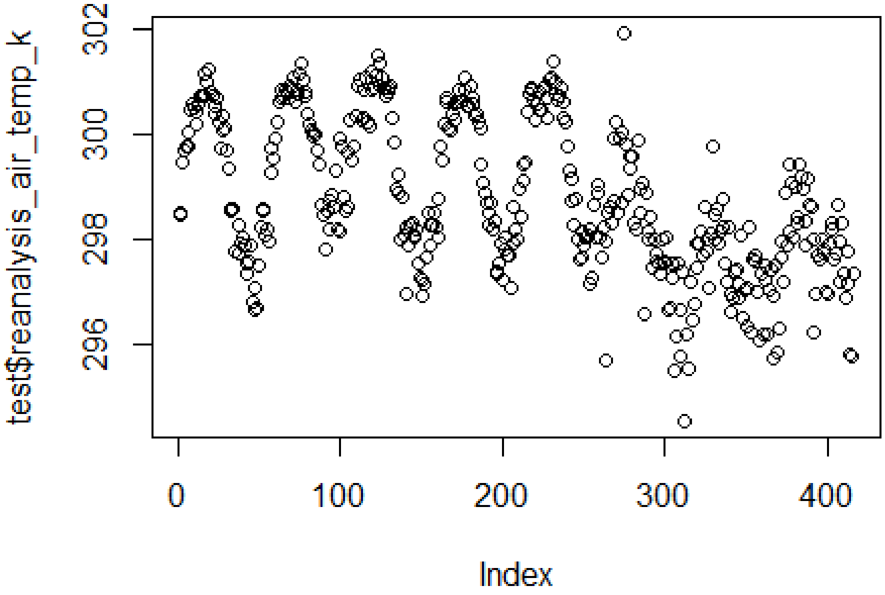
```
mean(test$ndvi_nw,na.rm = TRUE)

## [1] 0.126803

test$ndvi_nw[is.na(test$ndvi_nw)] <- mean(test$ndvi_nw,na.rm = TRUE)

plot(test$ndvi_se)
```



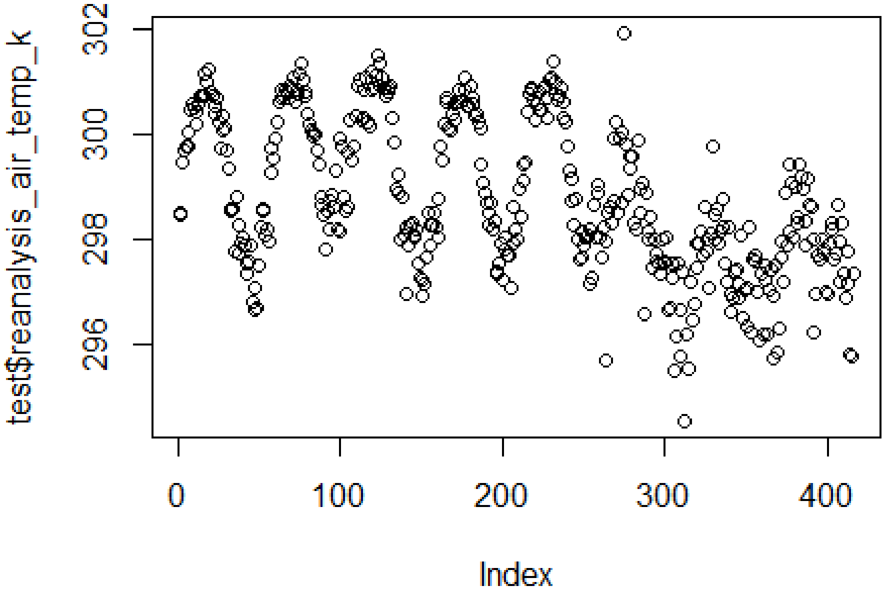
```
mean(test$ndvi_se,na.rm = TRUE)

## [1] 0.2077017

test$ndvi_se[is.na(test$ndvi_se)] <- mean(test$ndvi_se,na.rm = TRUE)


plot(test$precipitation_amt_mm)
```



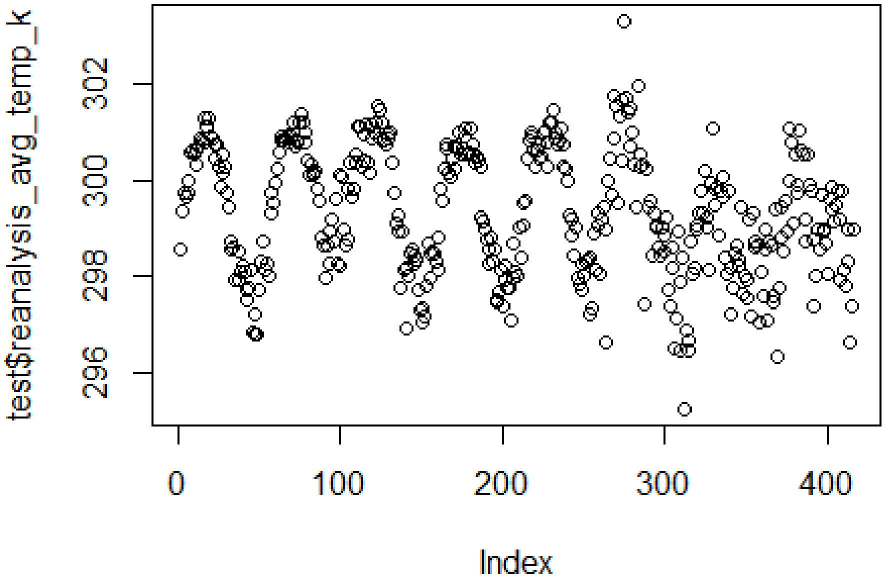
```
mean(test$precipitation_amt_mm,na.rm = TRUE)

## [1] 38.35432

test$precipitation_amt_mm[is.na(test$precipitation_amt_mm)] <-
mean(test$precipitation_amt_mm,na.rm = TRUE)

plot(test$reanalysis_air_temp_k)
```



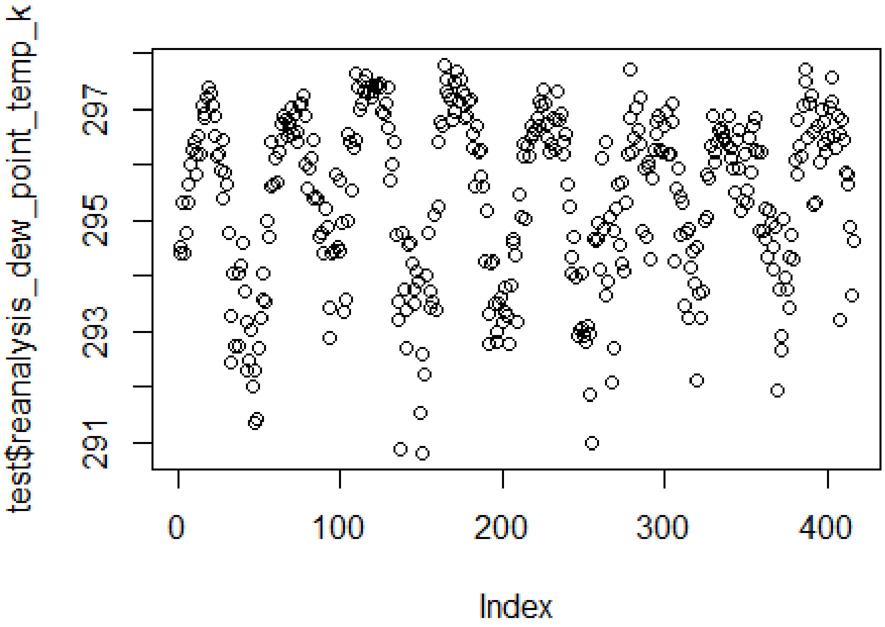
```
mean(test$reanalysis_air_temp_k,na.rm = TRUE)

## [1] 298.8183

test$reanalysis_air_temp_k[is.na(test$reanalysis_air_temp_k)] <-
mean(test$reanalysis_air_temp_k,na.rm = TRUE)

plot(test$reanalysis_air_temp_k)
```



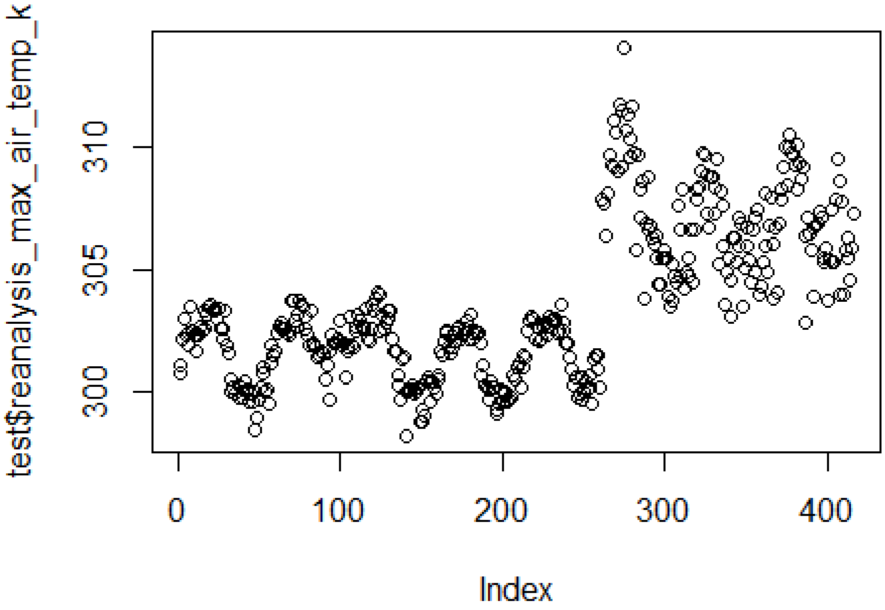
```
mean(test$reanalysis_air_temp_k,na.rm = TRUE)

## [1] 298.8183

test$reanalysis_air_temp_k[is.na(test$reanalysis_air_temp_k)] <-
mean(test$reanalysis_air_temp_k,na.rm = TRUE)

plot(test$reanalysis_avg_temp_k)
```



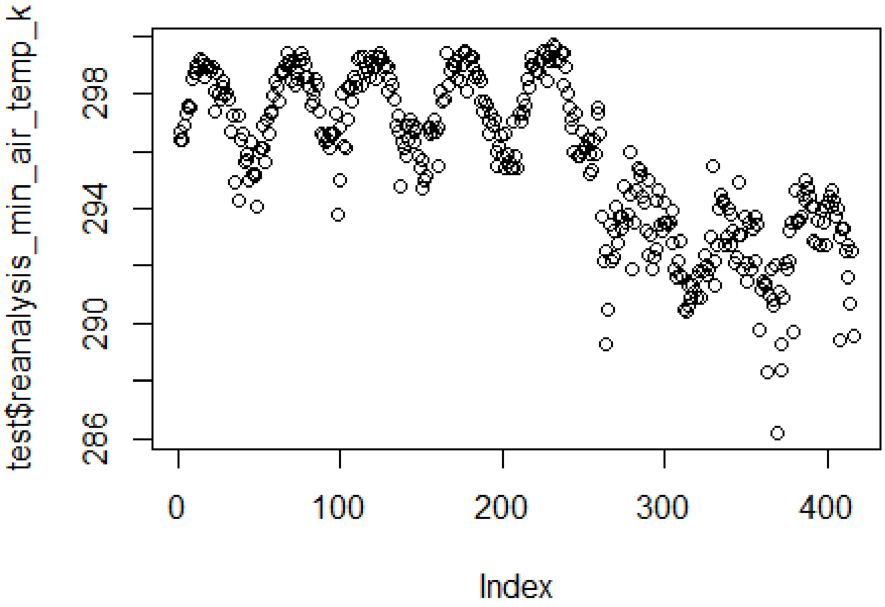
```
mean(test$reanalysis_avg_temp_k,na.rm = TRUE)

## [1] 299.3531

test$reanalysis_avg_temp_k[is.na(test$reanalysis_avg_temp_k)] <-
mean(test$reanalysis_avg_temp_k,na.rm = TRUE)

plot(test$reanalysis_dew_point_temp_k)
```



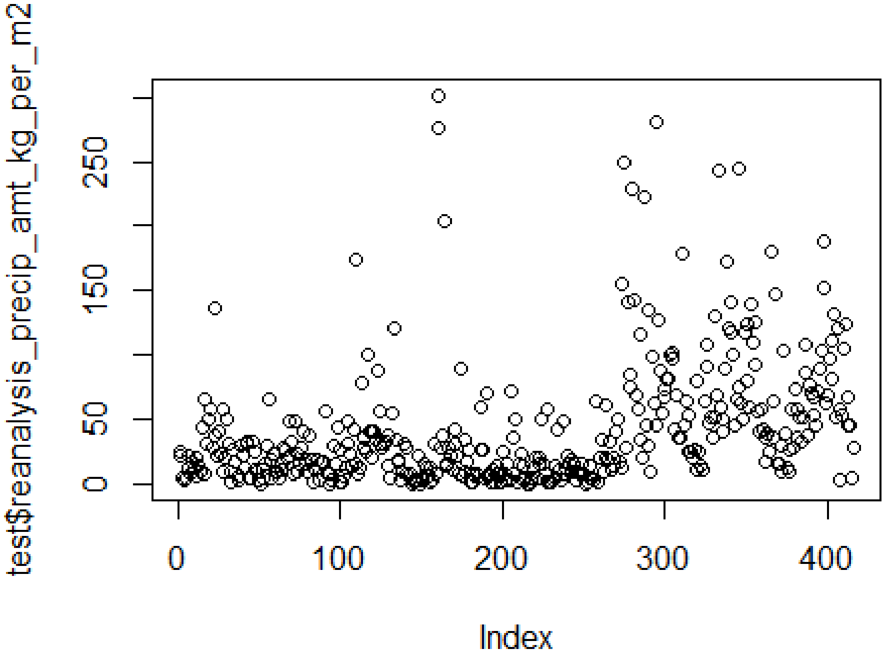
```
mean(test$reanalysis_dew_point_temp_k,na.rm = TRUE)

## [1] 295.4192

test$reanalysis_dew_point_temp_k[is.na(test$reanalysis_dew_point_temp_k)] <-
mean(test$reanalysis_dew_point_temp_k,na.rm = TRUE)

plot(test$reanalysis_max_air_temp_k)
```



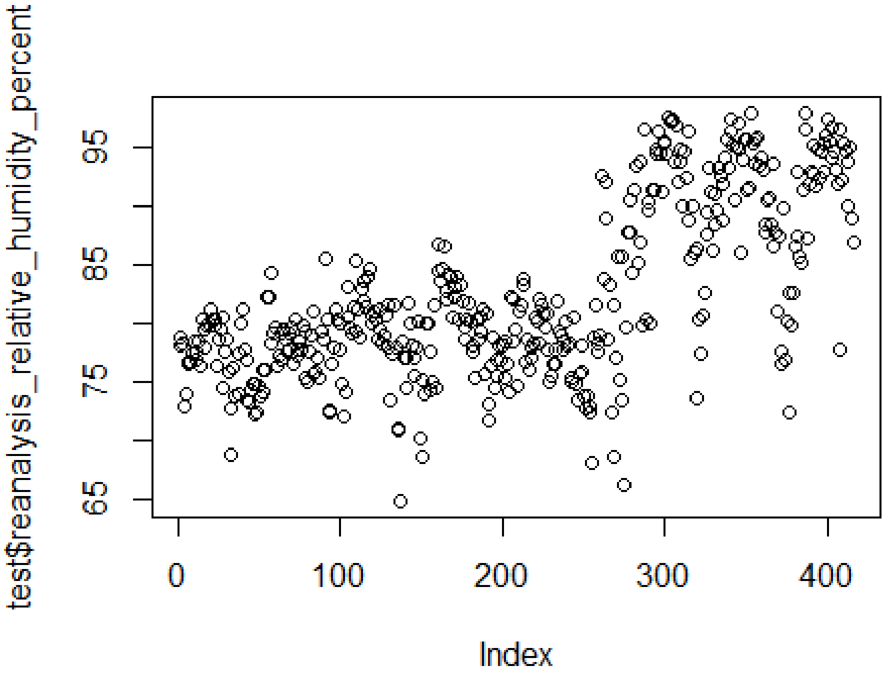
```
mean(test$reanalysis_max_air_temp_k,na.rm = TRUE)

## [1] 303.6234

test$reanalysis_max_air_temp_k[is.na(test$reanalysis_max_air_temp_k)] <-
mean(test$reanalysis_max_air_temp_k,na.rm = TRUE)

plot(test$reanalysis_min_air_temp_k)
```



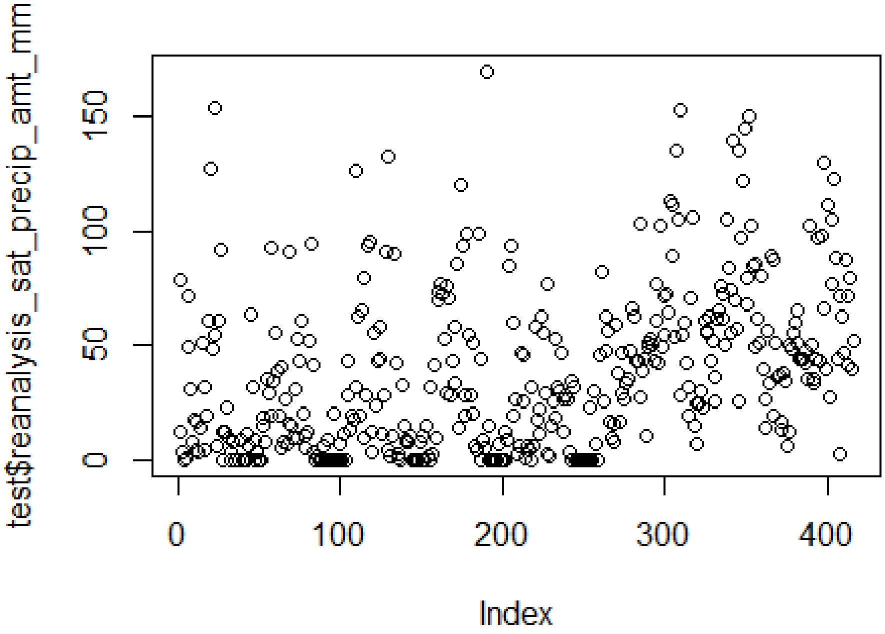
```
mean(test$reanalysis_min_air_temp_k,na.rm = TRUE)

## [1] 295.7435

test$reanalysis_min_air_temp_k[is.na(test$reanalysis_min_air_temp_k)] <-
mean(test$reanalysis_min_air_temp_k,na.rm = TRUE)

plot(test$reanalysis_precip_amt_kg_per_m2)
```



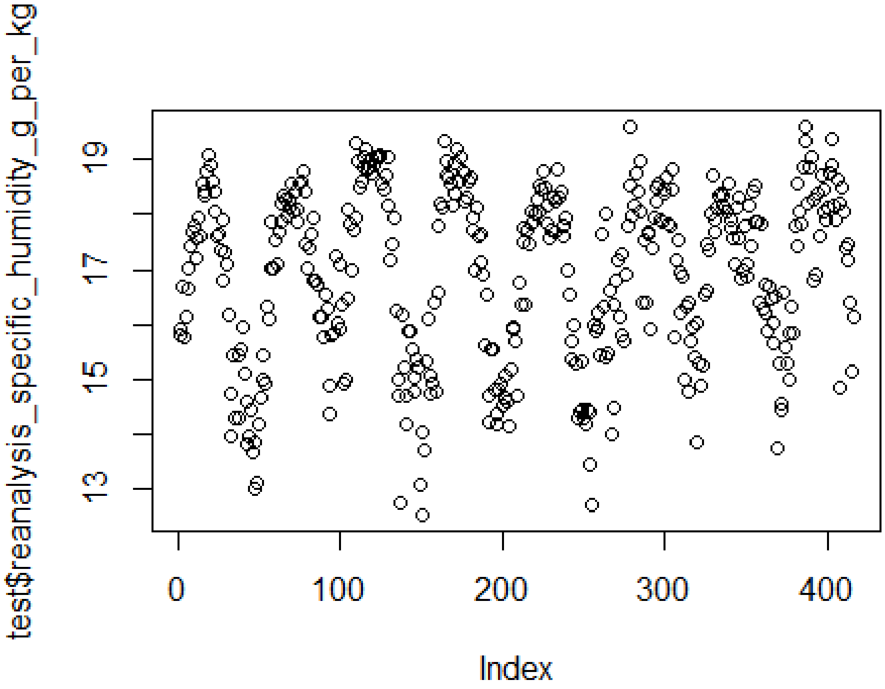
```
mean(test$reanalysis_precip_amt_kg_per_m2,na.rm = TRUE)

## [1] 42.17114

test$reanalysis_precip_amt_kg_per_m2[is.na(test$reanalysis_precip_amt_kg_per_
m2)] <- mean(test$reanalysis_precip_amt_kg_per_m2,na.rm = TRUE)

plot(test$reanalysis_relative_humidity_percent)
```



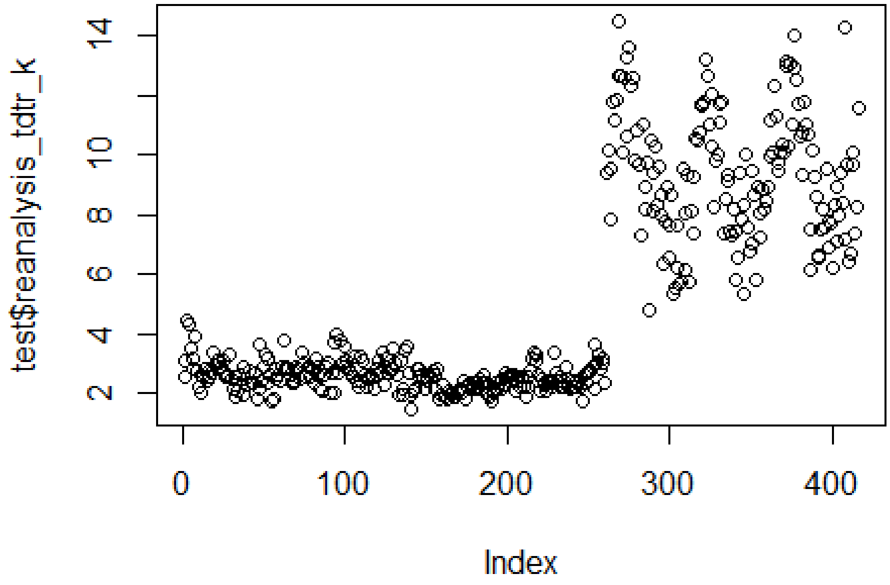
```
mean(test$reanalysis_relative_humidity_percent,na.rm = TRUE)

## [1] 82.49981

test$reanalysis_relative_humidity_percent[is.na(test$reanalysis_relative_humi
dity_percent)] <- mean(test$reanalysis_relative_humidity_percent,na.rm =
TRUE)

plot(test$reanalysis_sat_precip_amt_mm)
```



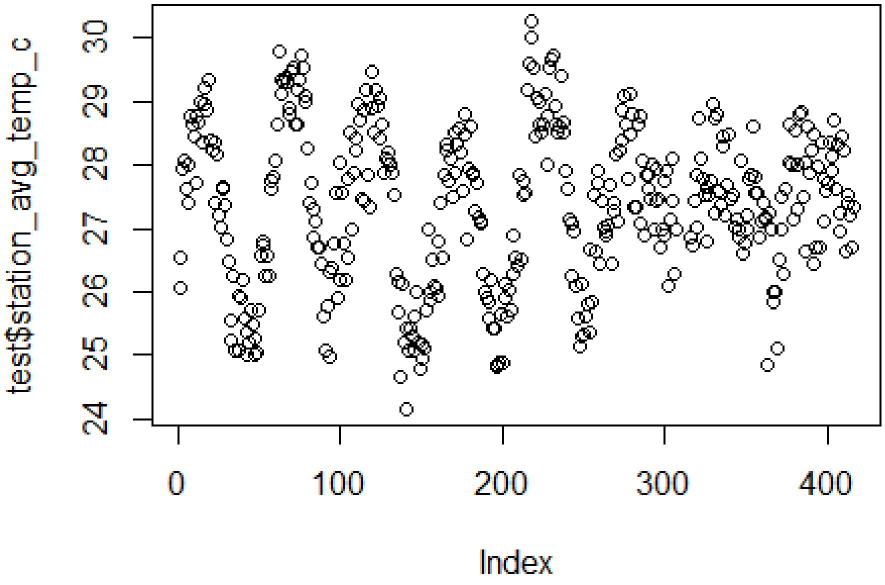
```
mean(test$reanalysis_sat_precip_amt_mm,na.rm = TRUE)

## [1] 38.35432

test$reanalysis_sat_precip_amt_mm[is.na(test$reanalysis_sat_precip_amt_mm)]
<- mean(test$reanalysis_sat_precip_amt_mm,na.rm = TRUE)

plot(test$reanalysis_specific_humidity_g_per_kg)
```



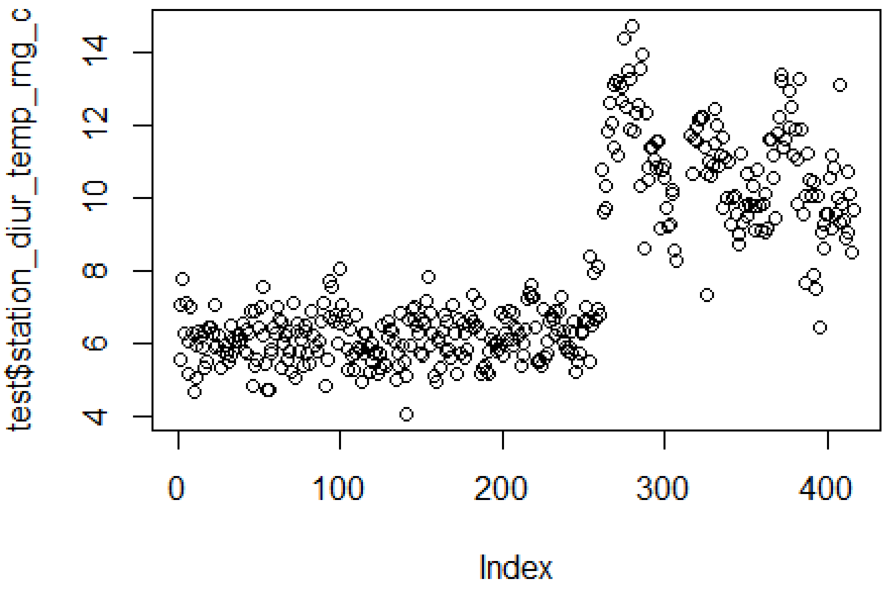
```
mean(test$reanalysis_specific_humidity_g_per_kg,na.rm = TRUE)

## [1] 16.92709

test$reanalysis_specific_humidity_g_per_kg[is.na(test$reanalysis_specific_hum
idity_g_per_kg)] <- mean(test$reanalysis_specific_humidity_g_per_kg,na.rm =
TRUE)

plot(test$reanalysis_tdtr_k)
```



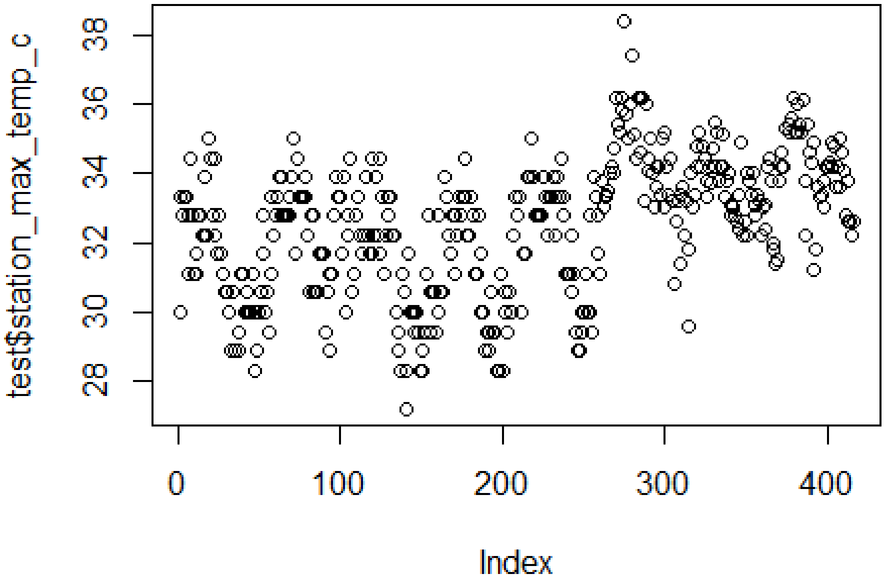
```
mean(test$reanalysis_tdtr_k,na.rm = TRUE)

## [1] 5.124569

test$reanalysis_tdtr_k[is.na(test$reanalysis_tdtr_k)] <-
mean(test$reanalysis_tdtr_k,na.rm = TRUE)

plot(test$station_avg_temp_c)
```



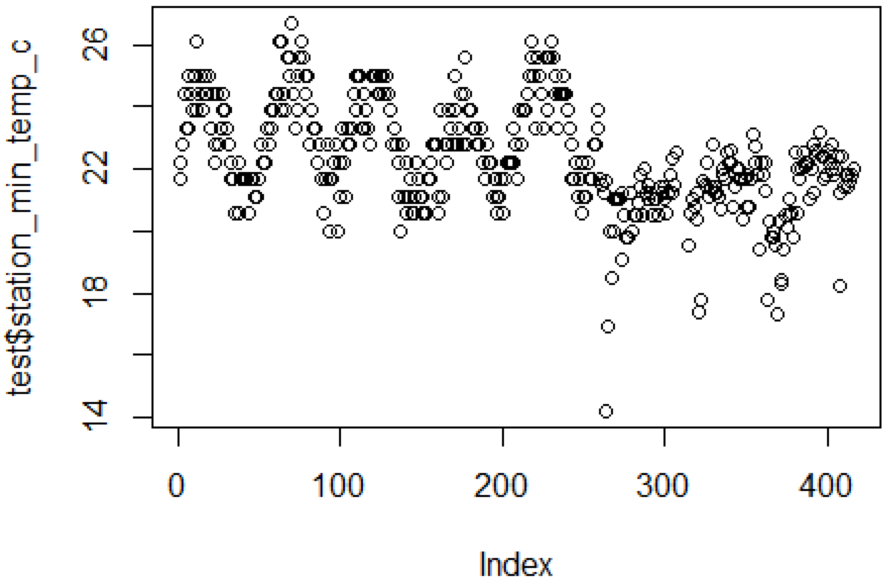
```
mean(test$station_avg_temp_c,na.rm = TRUE)

## [1] 27.36959

test$station_avg_temp_c[is.na(test$station_avg_temp_c)] <-
mean(test$station_avg_temp_c,na.rm = TRUE)

plot(test$station_diur_temp_rng_c)
```



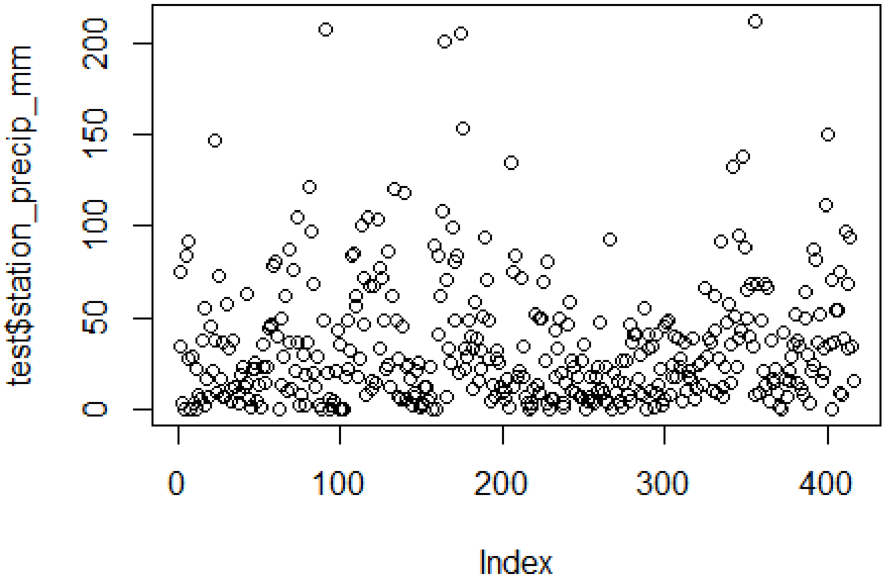
```
mean(test$station_diur_temp_rng_c,na.rm = TRUE)

## [1] 7.810991

test$station_diur_temp_rng_c[is.na(test$station_diur_temp_rng_c)] <-
mean(test$station_diur_temp_rng_c,na.rm = TRUE)

plot(test$station_max_temp_c)
```



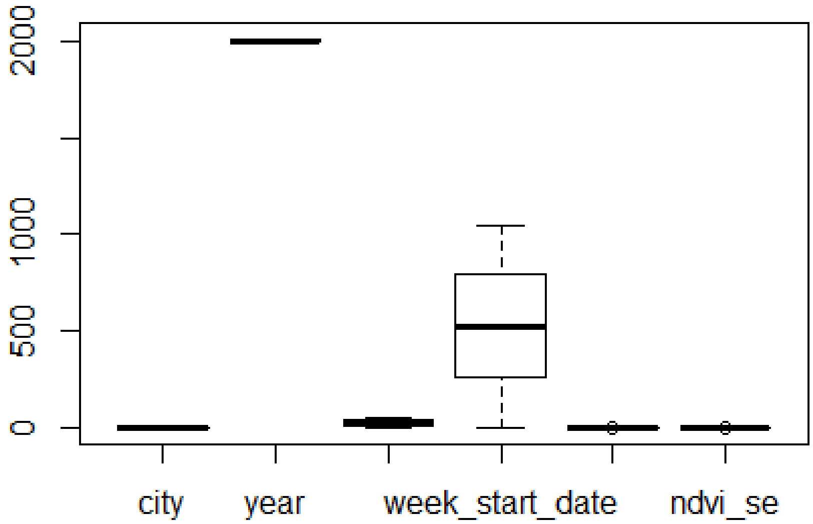
```
mean(test$station_max_temp_c,na.rm = TRUE)

## [1] 32.53462

test$station_max_temp_c[is.na(test$station_max_temp_c)] <-
mean(test$station_max_temp_c,na.rm = TRUE)

plot(test$station_min_temp_c)
```



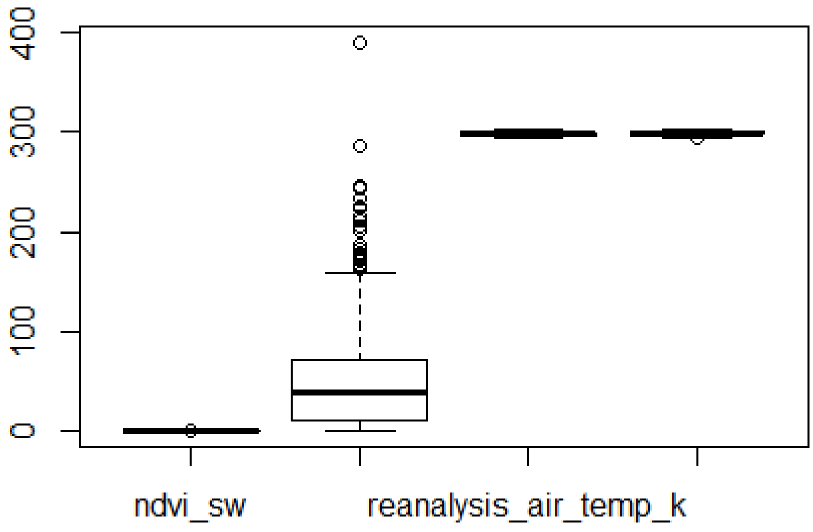
```
mean(test$station_min_temp_c,na.rm = TRUE)

## [1] 22.36855

test$station_min_temp_c[is.na(test$station_min_temp_c)] <-
mean(test$station_min_temp_c,na.rm = TRUE)

plot(test$station_precip_mm)
```



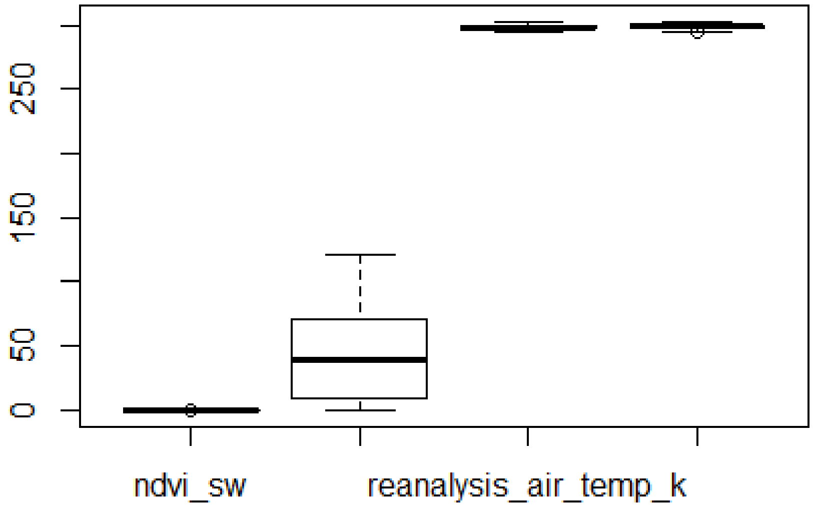
```
mean(test$station_precip_mm,na.rm = TRUE)

## [1] 34.27859

test$station_precip_mm[is.na(test$station_precip_mm)] <-
mean(test$station_precip_mm,na.rm = TRUE)


# count missing values (as percent)
apply(test, 2, function(x)
  round(100 * (length(which(is.na(x))))/length(x) , digits = 1)) %>%
  as.data.frame() %>%
  `names<-`('Percent of Missing Values')

##                                       Percent of Missing Values
## city                                                        0.0
## year                                                        0.0
## weekofyear                                                  0.0
## week_start_date                                             0.0
## ndvi_nw                                                     0.0
## ndvi_se                                                     0.0
## ndvi_sw                                                     0.2
## precipitation_amt_mm                                        0.0
## reanalysis_air_temp_k                                       0.0
## reanalysis_avg_temp_k                                       0.0
## reanalysis_dew_point_temp_k                                 0.0
## reanalysis_max_air_temp_k                                   0.0
## reanalysis_min_air_temp_k                                   0.0
## reanalysis_precip_amt_kg_per_m2                             0.0
## reanalysis_relative_humidity_percent                        0.0
## reanalysis_sat_precip_amt_mm                                0.0
## reanalysis_specific_humidity_g_per_kg                       0.0
## reanalysis_tdtr_k                                           0.0
## station_avg_temp_c                                          0.0
## station_diur_temp_rng_c                                     0.0
## station_max_temp_c                                          0.0
## station_min_temp_c                                          0.0
## station_precip_mm                                           0.0

##Checking the Outliers for the Values
boxplot(train[1:6])
```



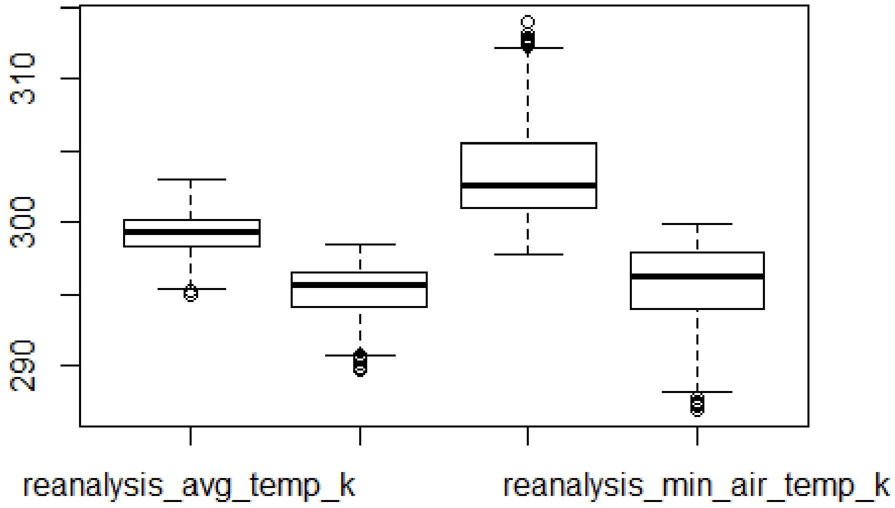
```
boxplot(train[7:10])
```



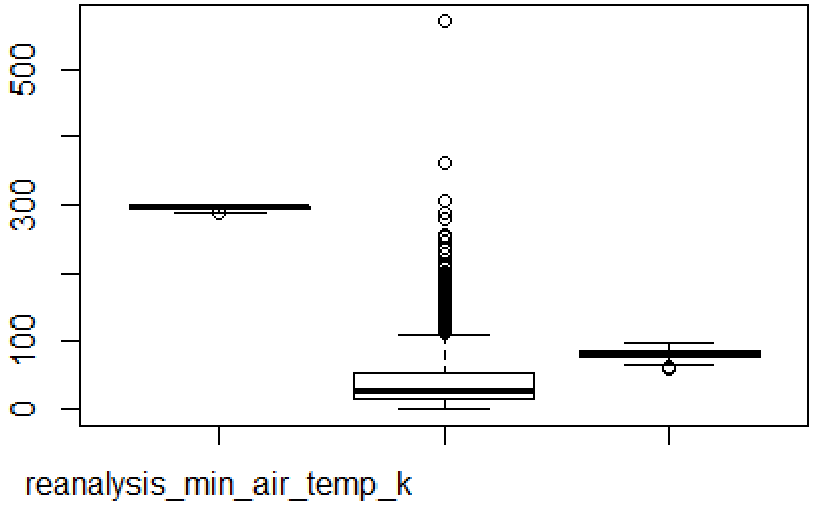
```
train$precipitation_amt_mm <- squish(train$precipitation_amt_mm,
quantile(train$precipitation_amt_mm, c(.05, .95)))
boxplot(train[7:10])
```



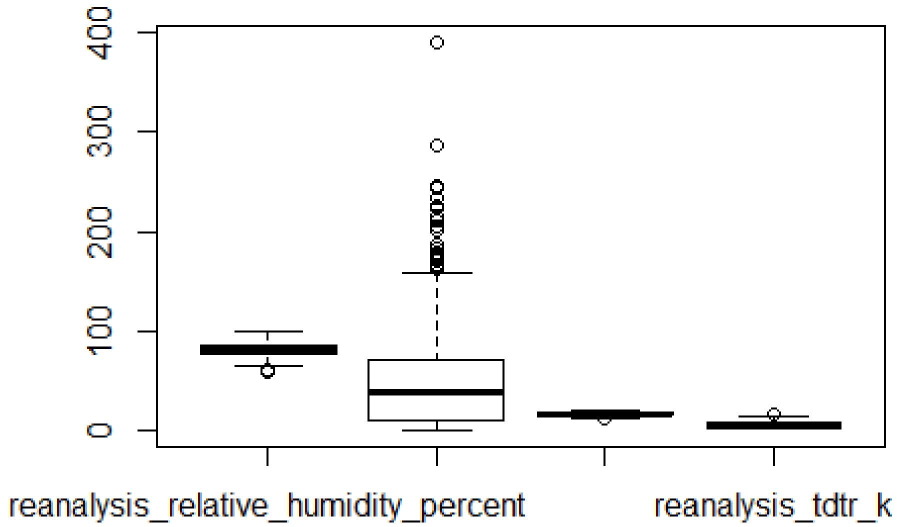
```r
boxplot(train[10:13])
```



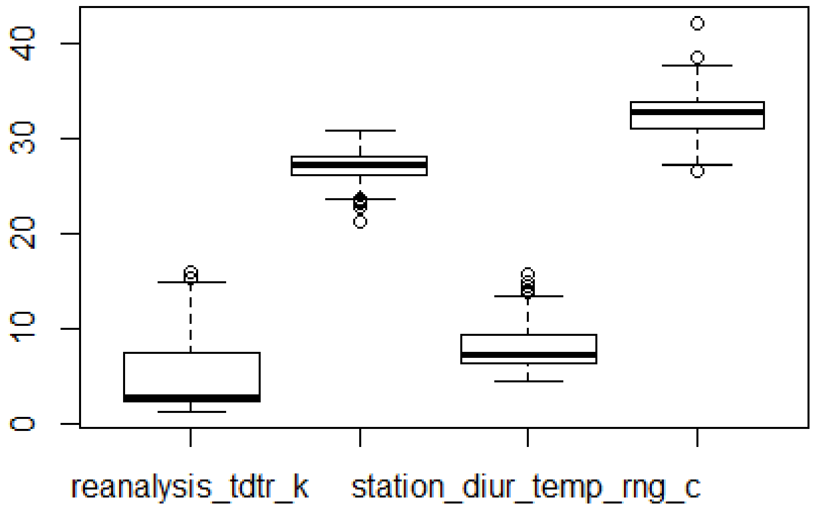
```r
boxplot(train[13:15])
```



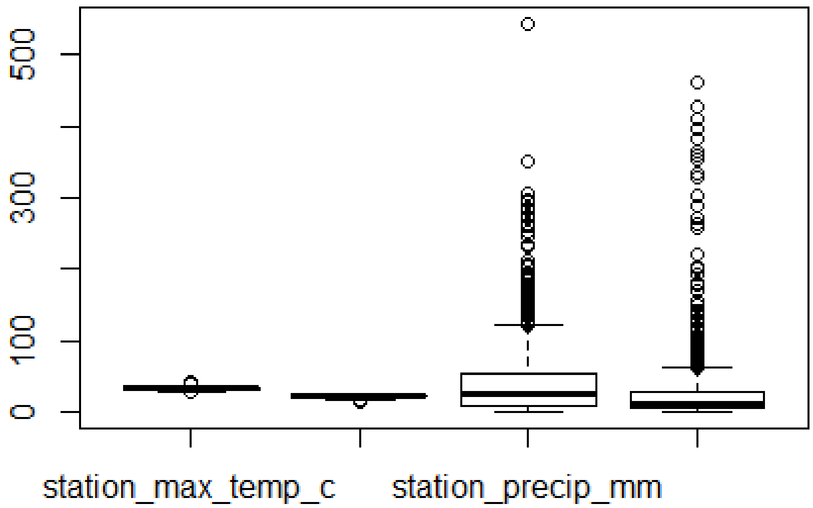
```r
train$reanalysis_precip_amt_kg_per_m2 <-
squish(train$reanalysis_precip_amt_kg_per_m2,
quantile(train$reanalysis_precip_amt_kg_per_m2, c(.05, .95)))
boxplot(train[15:18])
```



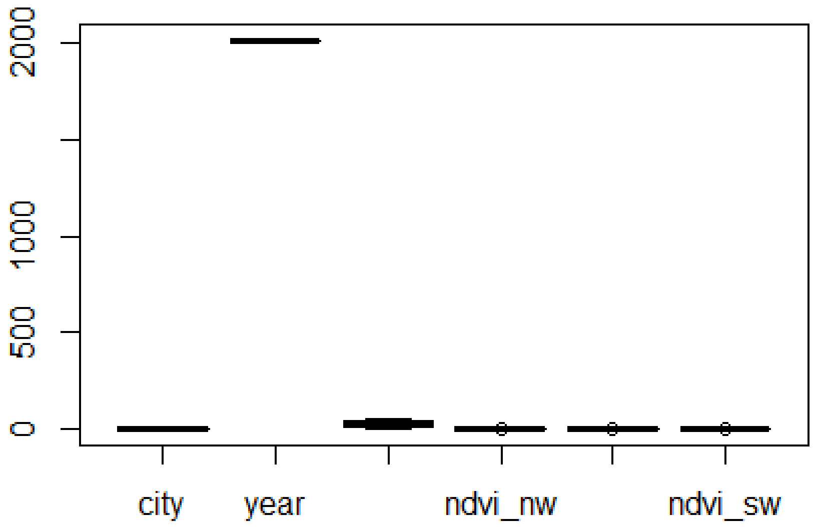
```r
train$reanalysis_sat_precip_amt_mm <-
squish(train$reanalysis_sat_precip_amt_mm,
quantile(train$reanalysis_sat_precip_amt_mm, c(.05, .95)))
boxplot(train[18:21])
```



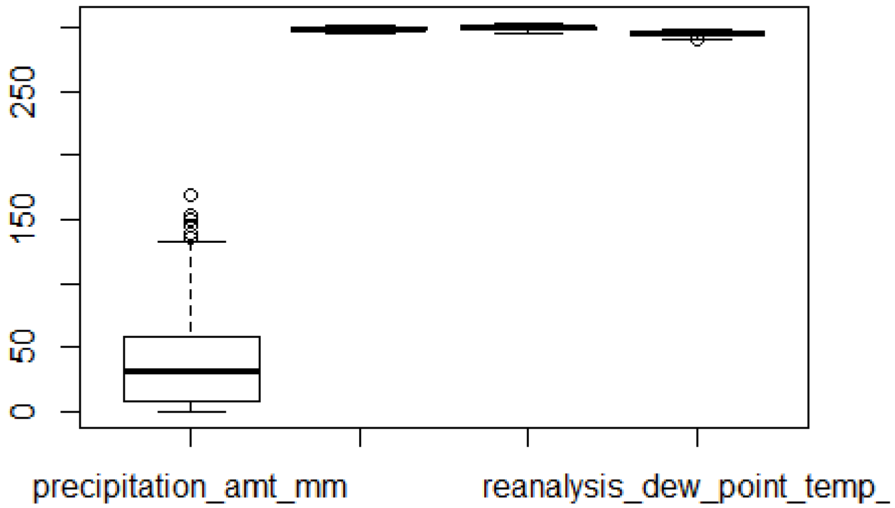
```
boxplot(train[21:24])
```



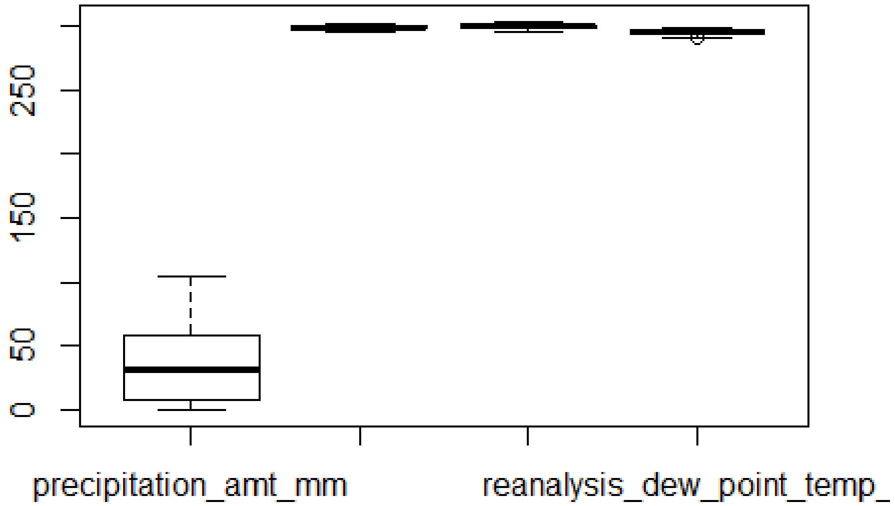
```
train$station_precip_mm <- squish(train$station_precip_mm,
quantile(train$station_precip_mm, c(.05, .95)))

train$week_start_date <- NULL
test$week_start_date <- NULL

##Checking the Outliers for the Values
boxplot(test[1:6])
```



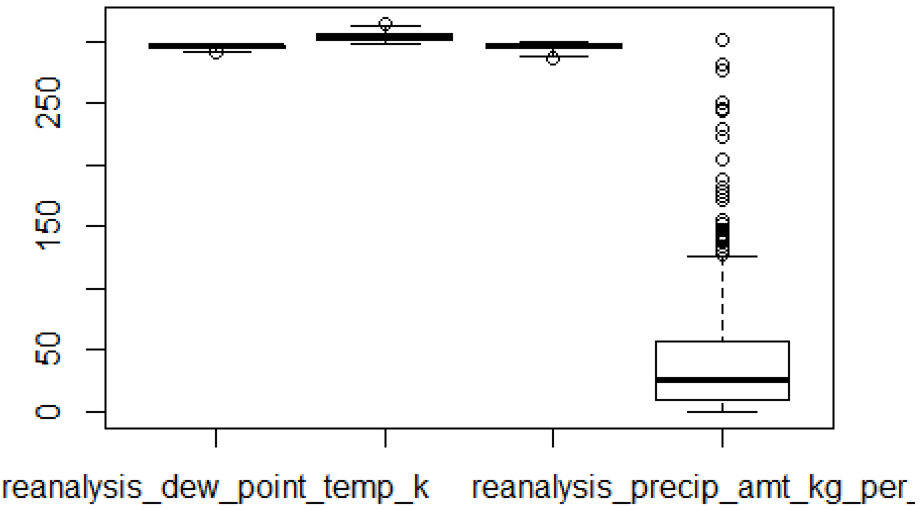
```
boxplot(test[7:10])
```



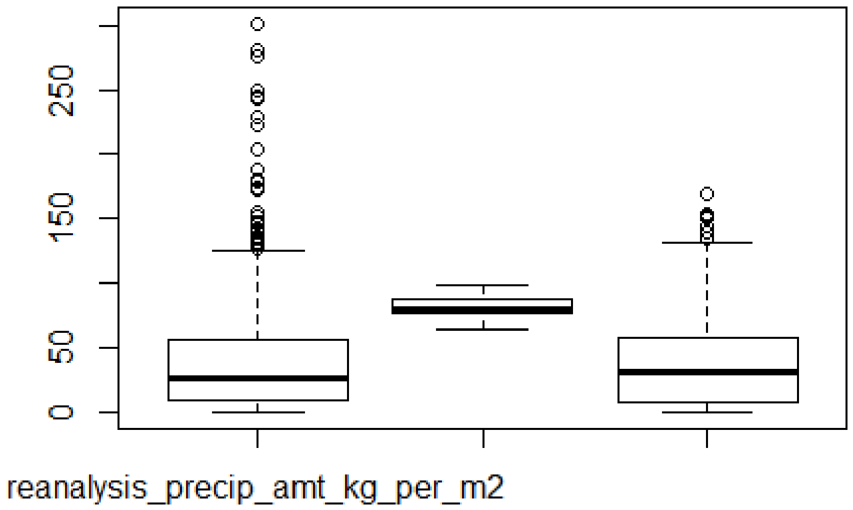
```
test$precipitation_amt_mm <- squish(test$precipitation_amt_mm,
quantile(test$precipitation_amt_mm, c(.05, .95)))
boxplot(test[7:10])
```



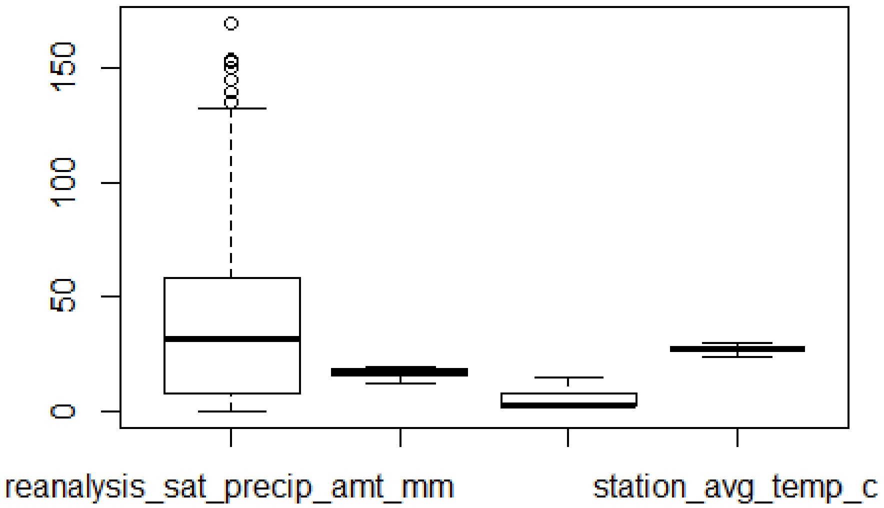
```r
boxplot(test[10:13])
```



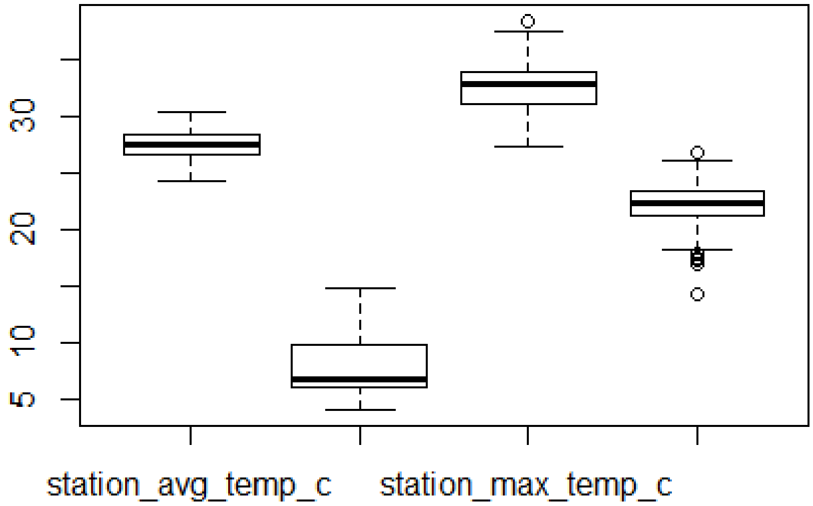
```r
boxplot(test[13:15])
```



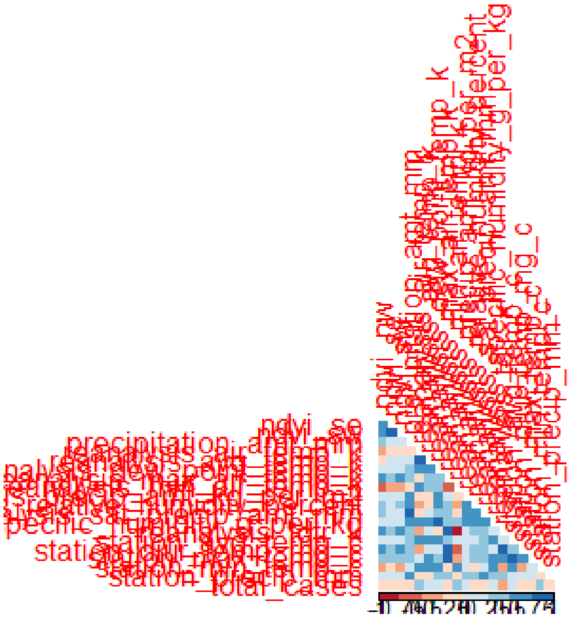
```
test$reanalysis_precip_amt_kg_per_m2 <-
squish(test$reanalysis_precip_amt_kg_per_m2,
quantile(test$reanalysis_precip_amt_kg_per_m2, c(.05, .95)))
boxplot(test[15:18])
```



```
test$reanalysis_sat_precip_amt_mm <-
squish(test$reanalysis_sat_precip_amt_mm,
quantile(test$reanalysis_sat_precip_amt_mm, c(.05, .95)))
boxplot(test[18:21])
```



```
#boxplot(test[21:24])
test$station_precip_mm <- squish(test$station_precip_mm,
quantile(test$station_precip_mm, c(.05, .95)))

## Ploting Some Data and Getting the Result on the basis of the same
train %>%
  dplyr::select(-city, -year, -weekofyear) %>%
  cor(use = 'pairwise.complete.obs') -> M1

corrplot(M1, type="lower", method="color",
         col=brewer.pal(n=8, name="RdBu"),diag=FALSE)
```



```
##Precisely none of these correlations are very strong.
##After seeing the Corelation we are moving forward to some of the Imputs
# reanalysis_specific_humidity_g_per_kg
#reanalysis_dew_point_temp_k
#station_avg_temp_c
#station_min_temp_c
#renanlysis_min_temp_k
#reanalysis_air_temp_k



#############################################################################
#               Modelling                                                   #
############################################################################

test_new   <- test[,c("reanalysis_air_temp_k" ,"reanalysis_min_air_temp_k",
"reanalysis_specific_humidity_g_per_kg", "reanalysis_dew_point_temp_k",
"station_avg_temp_c" , "station_min_temp_c")]
train_new  <- train[,c("reanalysis_air_temp_k" ,"reanalysis_min_air_temp_k",
"reanalysis_specific_humidity_g_per_kg", "reanalysis_dew_point_temp_k",
"station_avg_temp_c" , "station_min_temp_c")]
train_pred <- train[,c("total_cases")]




fit <- glm.nb("total_cases ~ 1 + reanalysis_specific_humidity_g_per_kg +
reanalysis_air_temp_k + reanalysis_dew_point_temp_k + station_avg_temp_c +
station_min_temp_c" , data = train_new )
```

```r
summary(train_new)
```

```
##  reanalysis_air_temp_k reanalysis_min_air_temp_k
##  Min.   :294.6         Min.   :286.9            
##  1st Qu.:297.7         1st Qu.:293.9            
##  Median :298.7         Median :296.2            
##  Mean   :298.7         Mean   :295.7            
##  3rd Qu.:299.8         3rd Qu.:297.9            
##  Max.   :302.2         Max.   :299.9            
##  reanalysis_specific_humidity_g_per_kg reanalysis_dew_point_temp_k
##  Min.   :11.72                         Min.   :289.6              
##  1st Qu.:15.56                         1st Qu.:294.1              
##  Median :17.07                         Median :295.6              
##  Mean   :16.75                         Mean   :295.2              
##  3rd Qu.:17.97                         3rd Qu.:296.5              
##  Max.   :20.46                         Max.   :298.4              
##  station_avg_temp_c station_min_temp_c
##  Min.   :21.40      Min.   :14.7      
##  1st Qu.:26.33      1st Qu.:21.1      
##  Median :27.39      Median :22.2      
##  Mean   :27.19      Mean   :22.1      
##  3rd Qu.:28.13      3rd Qu.:23.3      
##  Max.   :30.80      Max.   :25.6
```

```r
sj_iq_test <- test[[1]]
sj_iq_test$predicted = predict(fit , test_new, type = 'response')
```

```
## Warning in sj_iq_test$predicted = predict(fit, test_new, type =
## "response"): Coercing LHS to a list
```

```r
fin_pred_val <- round(sj_iq_test$predicted)

submissions = read.csv("D:\\Kaggle Projects\\Dengue\\submission_format.csv",
header=TRUE, sep=",")

submissions$total_cases <- fin_pred_val

write.csv(submissions, 'D:\\Kaggle Projects\\Dengue\\predictions.csv',
row.names = FALSE)
```