Problem 5 - Final Exam

2022-12-04

I have neither given nor received aid on this examination, nor have I concealed any violation of the Honor Code.

A: Preprocessing

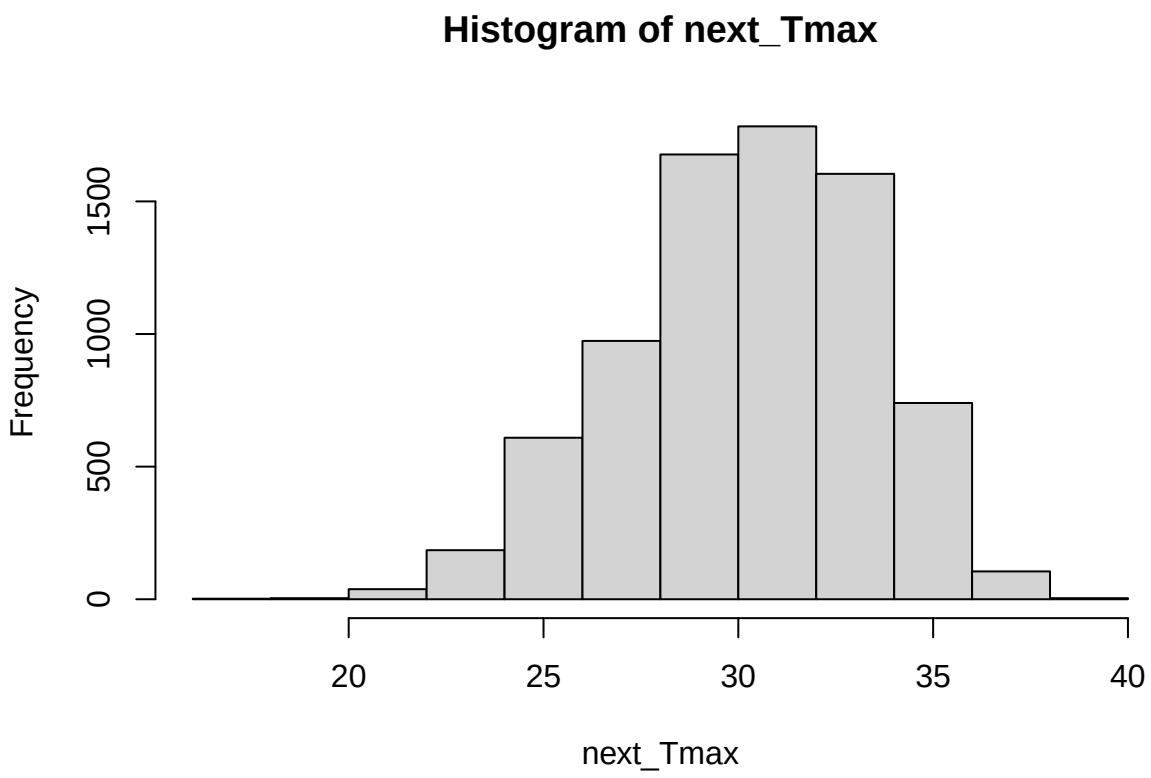
Read the Data

```
data_v0 <- read.csv("Bias_correction_ucl.csv")
```

Visualize the Target Variable

First of all, we just visualize the target variable to see if there are any anomalies. The data seems fine.

```
#plotting target variable distribution
next_Tmax<-as.numeric(data_v0$Next_Tmax , xlab="Next T max (Target)")
hist(next_Tmax)
```



Delete Missing Data (delete rows)

Since there are only $\frac{164}{7750}$ rows with missing values, we can delete them.

```
data <-  na.omit(data_v0)

n<- nrow(data)
deleted_rows<-nrow(data_v0)- n
print(paste("Deleted Rows: ",deleted_rows))
```

```
## [1] "Deleted Rows:  164"
```

Delete Next_Tmin Column

This column is not needed, so we can drop it.

```
#install.packages("dplyr")
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
```

```
data_v1 = select(data, -c("Next_Tmin"))
```

Change Column Types

Since the date here is for only the summer period (month 6 (june) to month 8 (august)), we can take only the years. We also change station to factor, since it's categorical.

```
data_v1$Date = as.integer(format(as.Date(data_v1$Date), "%Y"))
data_v1$Date = data_v1$Date
data_v1$station<-factor(data_v1$station)
```

Check the Data Types

```
str(data_v1)
```

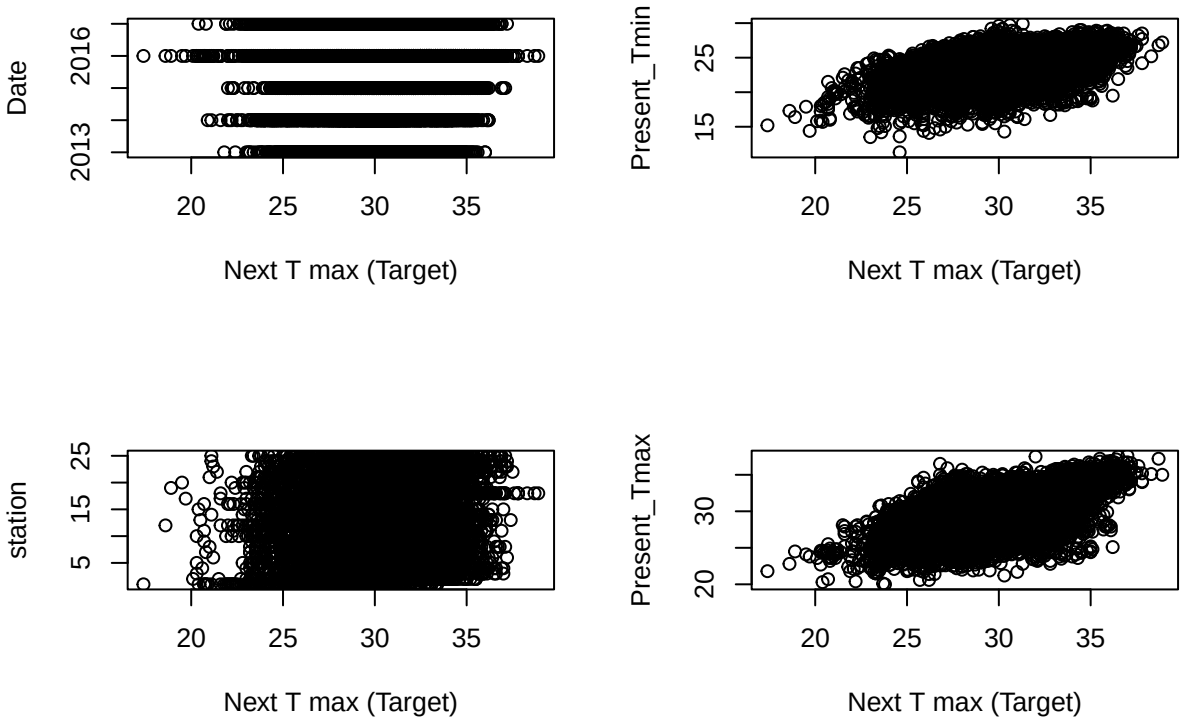
```
## 'data.frame':    7588 obs. of  24 variables:
##  $ station         : Factor w/ 25 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Date            : int  2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 ...
##  $ Present_Tmax    : num  28.7 31.9 31.6 32 31.4 31.9 31.4 32.1 31.4 31.6 ...
##  $ Present_Tmin    : num  21.4 21.6 23.3 23.4 21.9 23.5 24.4 23.6 22 20.5 ...
##  $ LDAPS_RHmin     : num  58.3 52.3 48.7 58.2 56.2 ...
##  $ LDAPS_RHmax     : num  91.1 90.6 84 96.5 90.2 ...
```

```
##  $ LDAPS_Tmax_lapse: num  28.1 29.9 30.1 29.7 29.1 ...
##  $ LDAPS_Tmin_lapse: num  23 24 24.6 23.3 23.5 ...
##  $ LDAPS_WS        : num  6.82 5.69 6.14 5.65 5.74 ...
##  $ LDAPS_LH        : num  69.5 51.9 20.6 65.7 108 ...
##  $ LDAPS_CC1       : num  0.234 0.226 0.209 0.216 0.151 ...
##  $ LDAPS_CC2       : num  0.204 0.252 0.257 0.226 0.25 ...
##  $ LDAPS_CC3       : num  0.162 0.159 0.204 0.161 0.179 ...
##  $ LDAPS_CC4       : num  0.131 0.128 0.142 0.134 0.17 ...
##  $ LDAPS_PPT1      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LDAPS_PPT2      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LDAPS_PPT3      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LDAPS_PPT4      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ lat             : num  37.6 37.6 37.6 37.6 37.6 ...
##  $ lon             : num  127 127 127 127 127 ...
##  $ DEM             : num  212.3 44.8 33.3 45.7 35 ...
##  $ Slope           : num  2.785 0.514 0.266 2.535 0.505 ...
##  $ Solar.radiation : num  5993 5869 5864 5857 5860 ...
##  $ Next_Tmax       : num  29.1 30.5 31.1 31.7 31.2 31.5 30.9 31.1 31.3 30.5 ...
##  - attr(*, "na.action")= 'omit' Named int [1:164] 226 272 301 451 465 628 832 857 882 914 ...
##   ..- attr(*, "names")= chr [1:164] "226" "272" "301" "451" ...
```

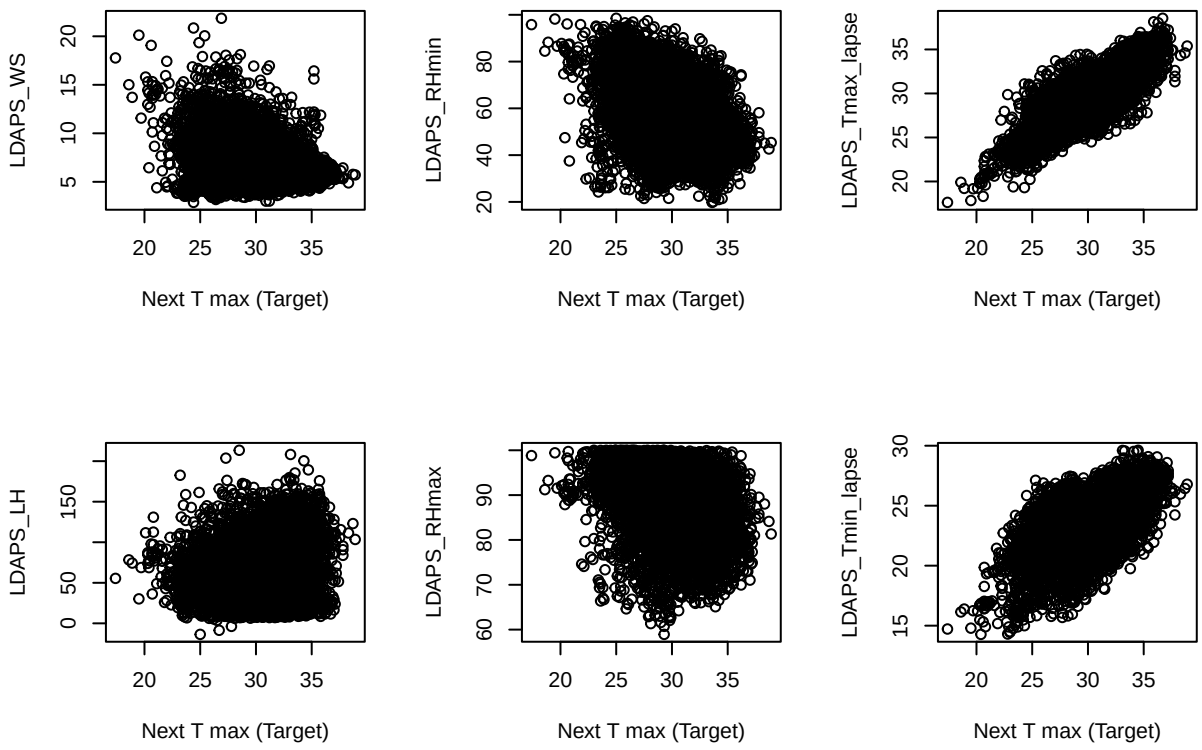
Scatter Plots of Features vs Target Variable

From the scatter plots below, we can find that some features seem to have a linear relationship with the target variable, while others seem to be random.

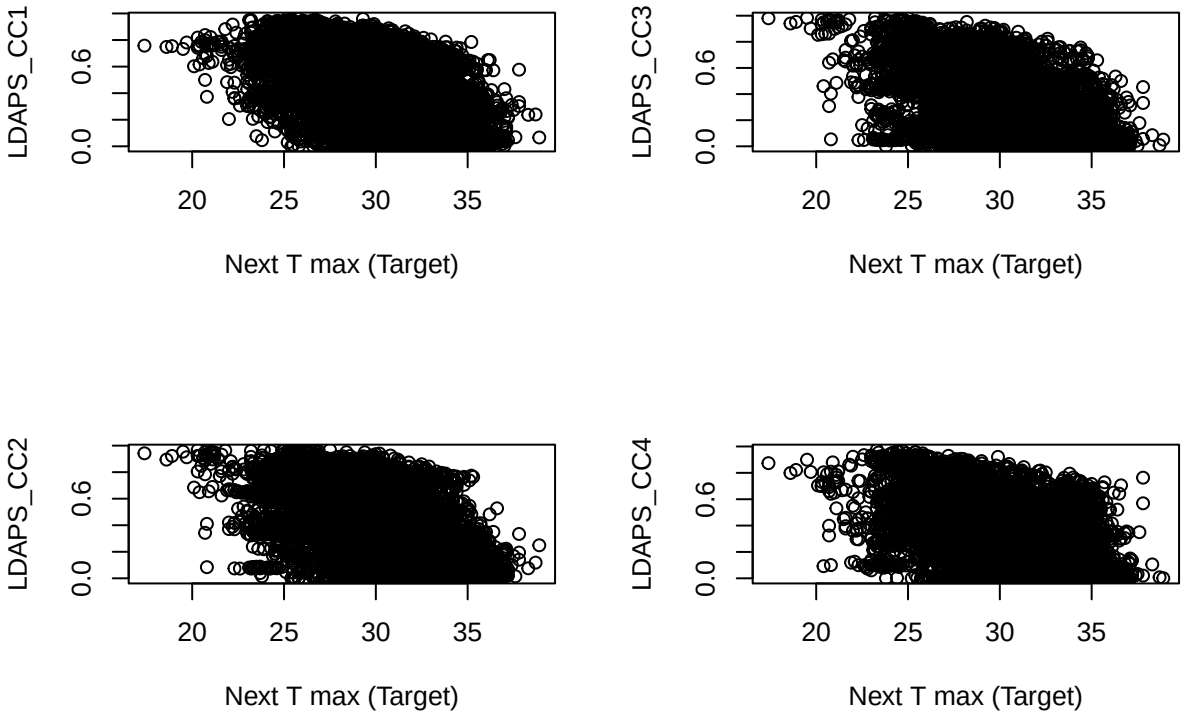
```
layout(matrix(1:4,2,2))
plot(data_v1$Next_Tmax,data_v1$Date ,  ylab="Date", xlab="Next T max (Target)")
plot(data_v1$Next_Tmax,data_v1$station ,  ylab="station", xlab="Next T max (Target)")
plot(data_v1$Next_Tmax,data_v1$Present_Tmin ,  ylab="Present_Tmin", xlab="Next T max (Target)")
plot(data_v1$Next_Tmax,data_v1$Present_Tmax ,  ylab="Present_Tmax", xlab="Next T max (Target)")
```



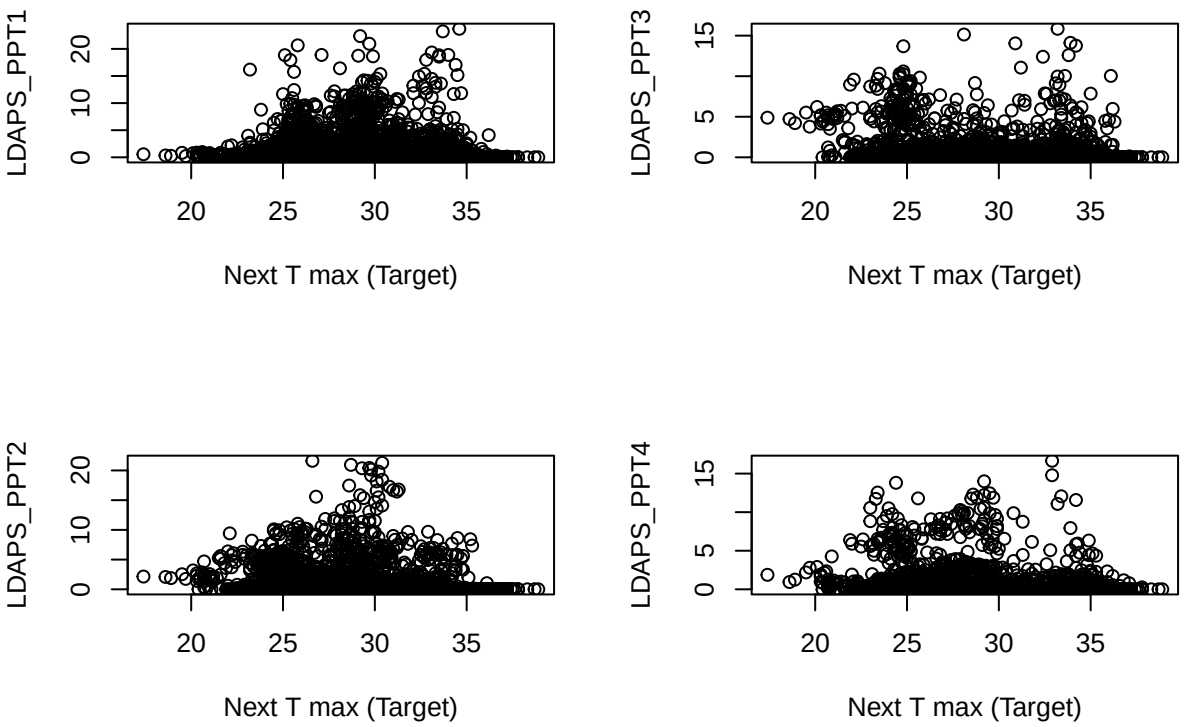

```
layout(matrix(1:6,2,3))
plot(data_v1$Next_Tmax,data_v1$LDAPS_WS ,  ylab="LDAPS_WS", xlab="Next T max (Target)")
plot(data_v1$Next_Tmax,data_v1$LDAPS_LH ,  ylab="LDAPS_LH", xlab="Next T max (Target)")
plot(data_v1$Next_Tmax,data_v1$LDAPS_RHmin ,  ylab="LDAPS_RHmin", xlab="Next T max (Target)")
plot(data_v1$Next_Tmax,data_v1$LDAPS_RHmax ,  ylab="LDAPS_RHmax", xlab="Next T max (Target)")
plot(data_v1$Next_Tmax,data_v1$LDAPS_Tmax_lapse ,  ylab="LDAPS_Tmax_lapse", xlab="Next T max (Target)")
plot(data_v1$Next_Tmax,data_v1$LDAPS_Tmin_lapse ,  ylab="LDAPS_Tmin_lapse", xlab="Next T max (Target)")
```



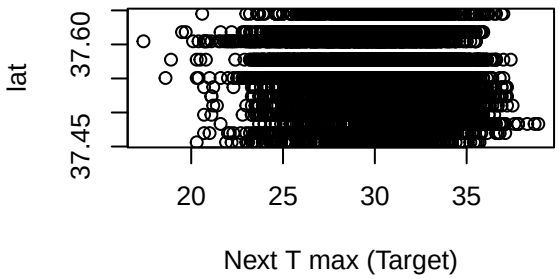

```
layout(matrix(1:4,2,2))
plot(data_v1$Next_Tmax,data_v1$LDAPS_CC1 ,  ylab="LDAPS_CC1", xlab="Next T max (Target)")
plot(data_v1$Next_Tmax,data_v1$LDAPS_CC2 ,  ylab="LDAPS_CC2", xlab="Next T max (Target)")
plot(data_v1$Next_Tmax,data_v1$LDAPS_CC3 ,  ylab="LDAPS_CC3", xlab="Next T max (Target)")
plot(data_v1$Next_Tmax,data_v1$LDAPS_CC4 ,  ylab="LDAPS_CC4", xlab="Next T max (Target)")
```

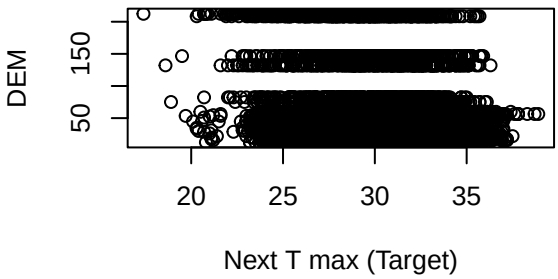


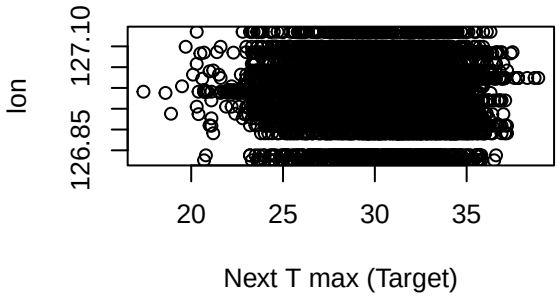

```
layout(matrix(1:4,2,2))
plot(data_v1$Next_Tmax,data_v1$LDAPS_PPT1 ,  ylab="LDAPS_PPT1", xlab="Next T max (Target)")
plot(data_v1$Next_Tmax,data_v1$LDAPS_PPT2 ,  ylab="LDAPS_PPT2", xlab="Next T max (Target)")
plot(data_v1$Next_Tmax,data_v1$LDAPS_PPT3 ,  ylab="LDAPS_PPT3", xlab="Next T max (Target)")
plot(data_v1$Next_Tmax,data_v1$LDAPS_PPT4 ,  ylab="LDAPS_PPT4", xlab="Next T max (Target)")
```

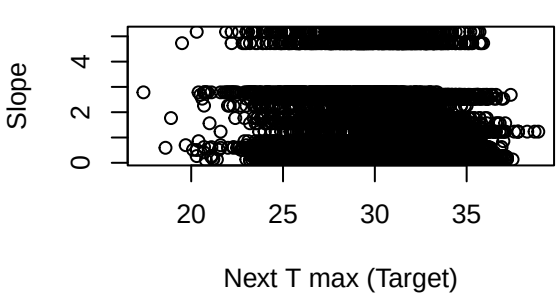



```
layout(matrix(1:4,2,2))
plot(data_v1$Next_Tmax,data_v1$lat ,  ylab="lat", xlab="Next T max (Target)")
plot(data_v1$Next_Tmax,data_v1$lon ,  ylab="lon", xlab="Next T max (Target)")
plot(data_v1$Next_Tmax,data_v1$DEM ,  ylab="DEM", xlab="Next T max (Target)")
plot(data_v1$Next_Tmax,data_v1$Slope ,  ylab="Slope", xlab="Next T max (Target)")
```

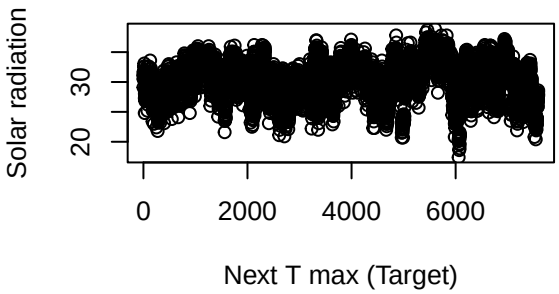









```
plot(data_v1$Next_Tmax,data_v1$`Solar radiation` ,  ylab="Solar radiation", xlab="Next T max (Target)")
```



ANOVA for Stations

We want to check if the stations have an effect on capturing the temperature data of the next day. In order to do that we performed analysis of variance. From the ANOVA table, the p_{value} is very small, which means that we accept that the stations have an effect on the temperature, and must be included in the linear model.

```
station.ANOVA<-aov(Next_Tmax ~station,data=data_v1)
summary(station.ANOVA)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## station       24   4170  173.73   18.96 <2e-16 ***
## Residuals   7563  69298    9.16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Drop Unnecessary Columns

We can drop lat, lon, DEM and Slope, because they are the same as stations. In fact these features constitute a station, so if we don't remove them, we will find in the anova table that their coefficient are marked as NA. Also, we will get the following warning "Warning: prediction from a rank-deficient fit may be misleading", when trying to predict the validation set, because there are coefficients that are perfectly correlated.

```
data_v1 = select(data_v1, -c("lat","lon","Slope","DEM"))
```

B: Split Data

We're splitting the data into 3 parts, based on the date.

```r
train_v1 <- data_v1[1:(n*0.60),]
valid <- data_v1[(n*0.60):(n*0.80),]
test <- data_v1[(n*0.80):n,]

#number of observations per splitted part
ntrain<-nrow(train_v1)
ntest<-nrow(test)
nvalid<-nrow(valid)

print(paste("Training Rows: ",ntrain, ", Validation Rows: ", nvalid, ", Testing Rows: ", ntest))
```

```
## [1] "Training Rows:  4552 , Validation Rows:  1518 , Testing Rows:  1518"
```

C: Model 1: Initial Regression Model

Fitting of a Multiple Linear Regression Model on the Training Set using All the Features

```r
init_model<-lm(train_v1$Next_Tmax~. , data=train_v1)
summary(init_model)
```

```
##
## Call:
## lm(formula = train_v1$Next_Tmax ~ ., data = train_v1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2715 -0.7902  0.0169  0.8007  4.7616
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -7.848e+02  5.470e+01 -14.347  < 2e-16 ***
## station2          8.414e-01  1.442e-01   5.837 5.70e-09 ***
## station3          6.299e-01  1.616e-01   3.898 9.84e-05 ***
## station4          1.703e+00  1.466e-01  11.614  < 2e-16 ***
## station5          1.062e+00  1.464e-01   7.255 4.70e-13 ***
## station6          1.156e+00  1.508e-01   7.667 2.15e-14 ***
## station7          1.173e+00  1.503e-01   7.806 7.27e-15 ***
## station8          1.180e+00  1.486e-01   7.944 2.45e-15 ***
## station9          1.691e+00  1.430e-01  11.822  < 2e-16 ***
## station10         1.466e+00  1.417e-01  10.343  < 2e-16 ***
## station11         1.183e+00  1.494e-01   7.920 2.96e-15 ***
## station12         1.230e+00  1.524e-01   8.074 8.64e-16 ***
## station13         1.082e+00  1.563e-01   6.921 5.10e-12 ***
## station14         1.111e+00  1.597e-01   6.958 3.95e-12 ***
## station15         9.307e-01  1.533e-01   6.070 1.38e-09 ***
## station16         6.349e-01  1.416e-01   4.484 7.51e-06 ***
## station17         9.288e-01  1.470e-01   6.320 2.87e-10 ***
## station18         2.421e+00  1.472e-01  16.445  < 2e-16 ***
## station19         1.102e+00  1.479e-01   7.451 1.10e-13 ***
## station20         2.211e+00  1.438e-01  15.376  < 2e-16 ***
## station21         3.875e-01  1.606e-01   2.413 0.015881 *
## station22         1.267e+00  1.459e-01   8.682  < 2e-16 ***
```

```
## station23         1.841e+00  1.461e-01  12.602  < 2e-16 ***
## station24         1.361e+00  1.513e-01   8.998  < 2e-16 ***
## station25         1.253e+00  1.600e-01   7.832 5.96e-15 ***
## Date              3.916e-01  2.712e-02  14.441  < 2e-16 ***
## Present_Tmax      5.738e-02  1.166e-02   4.922 8.88e-07 ***
## Present_Tmin      6.663e-03  1.575e-02   0.423 0.672366
## LDAPS_RHmin       3.632e-02  3.630e-03  10.006  < 2e-16 ***
## LDAPS_RHmax      -2.131e-02  4.086e-03  -5.215 1.92e-07 ***
## LDAPS_Tmax_lapse  6.906e-01  1.873e-02  36.876  < 2e-16 ***
## LDAPS_Tmin_lapse  1.868e-01  2.350e-02   7.950 2.35e-15 ***
## LDAPS_WS         -1.048e-01  1.054e-02  -9.936  < 2e-16 ***
## LDAPS_LH          5.296e-03  1.209e-03   4.379 1.22e-05 ***
## LDAPS_CC1        -1.794e+00  1.456e-01 -12.320  < 2e-16 ***
## LDAPS_CC2        -6.766e-01  1.792e-01  -3.777 0.000161 ***
## LDAPS_CC3        -7.287e-01  1.778e-01  -4.099 4.23e-05 ***
## LDAPS_CC4        -1.572e+00  1.381e-01 -11.385  < 2e-16 ***
## LDAPS_PPT1       -2.413e-02  1.118e-02  -2.159 0.030931 *
## LDAPS_PPT2        1.619e-01  1.299e-02  12.465  < 2e-16 ***
## LDAPS_PPT3        1.345e-02  2.123e-02   0.634 0.526405
## LDAPS_PPT4        5.430e-02  2.320e-02   2.340 0.019319 *
## Solar.radiation   3.087e-05  5.259e-05   0.587 0.557260
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.318 on 4509 degrees of freedom
## Multiple R-squared:  0.772,  Adjusted R-squared:  0.7699
## F-statistic: 363.5 on 42 and 4509 DF,  p-value: < 2.2e-16
```

From the anova table, we find that the p_{value} of the following coefficients: "Present_Tmin", "LDAPS_PPT3" and "Solar.radiation" is higher than 0.05 (if we assume $\alpha = 0.05$). So we accept that they are not significant. Also, from this table we find that the Adjusted R-squared is 0.7699, which means that 76.99% of the errors are explained by the model.

Prediction (Multiple Linear Regression Model) of the Validation Set

```
valid_prediction_v1 <-predict(init_model,valid)
```

RMSE Function

$$RMSE = \sqrt{\frac{\sum_1^n (y_i - \hat{y_i})^2}{n}} = \sqrt{\frac{SSE}{n}}$$

```
rmse <- function(actual, predicted, length) {
  sse = sum((actual - predicted )^2)
  rmse <- sqrt(sse/length)
  return(rmse)
}
```

RMSE of Validation Set

The $RMSE_{validation} = 1.3852$

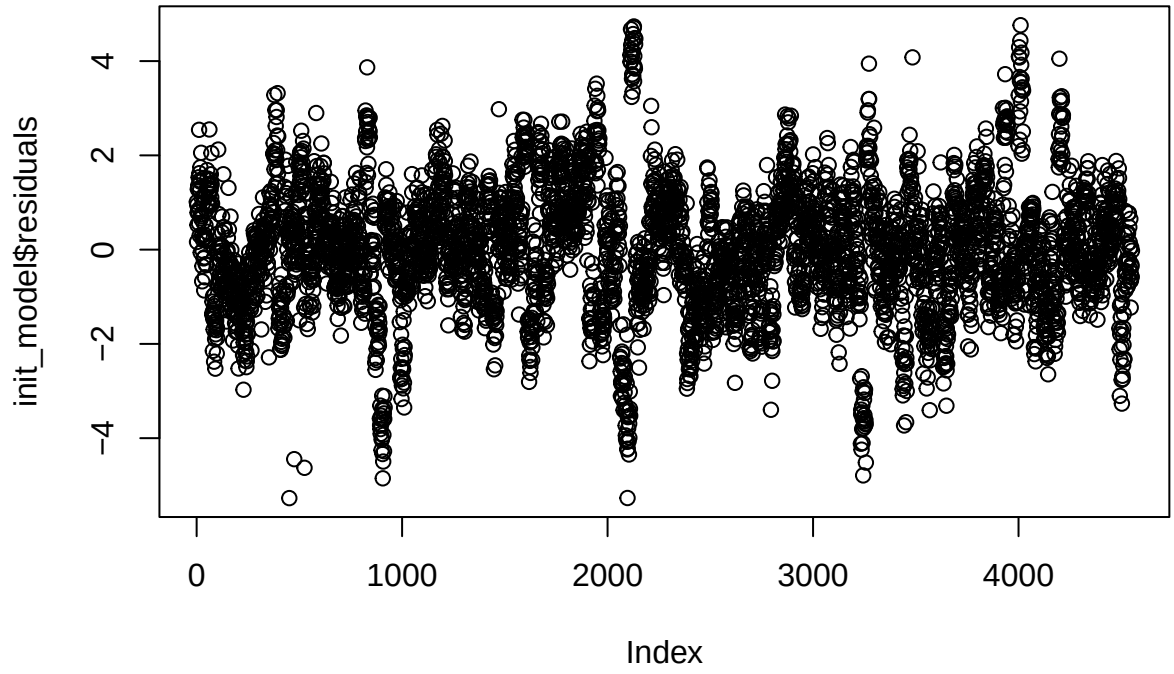
```
rmse_valid_v1 <- rmse(valid$Next_Tmax, valid_prediction_v1, nvalid)
round(rmse_valid_v1,4)
```

```
## [1] 1.3852
```

Initial Model Residuals

The residuals don't seem to be cyclical. Using Shapiro.test, we find that the p-value = 2.966e-15, which is less than $\alpha = 0.05$, therefore we accept that the residuals are not normally distributed for the intial model.

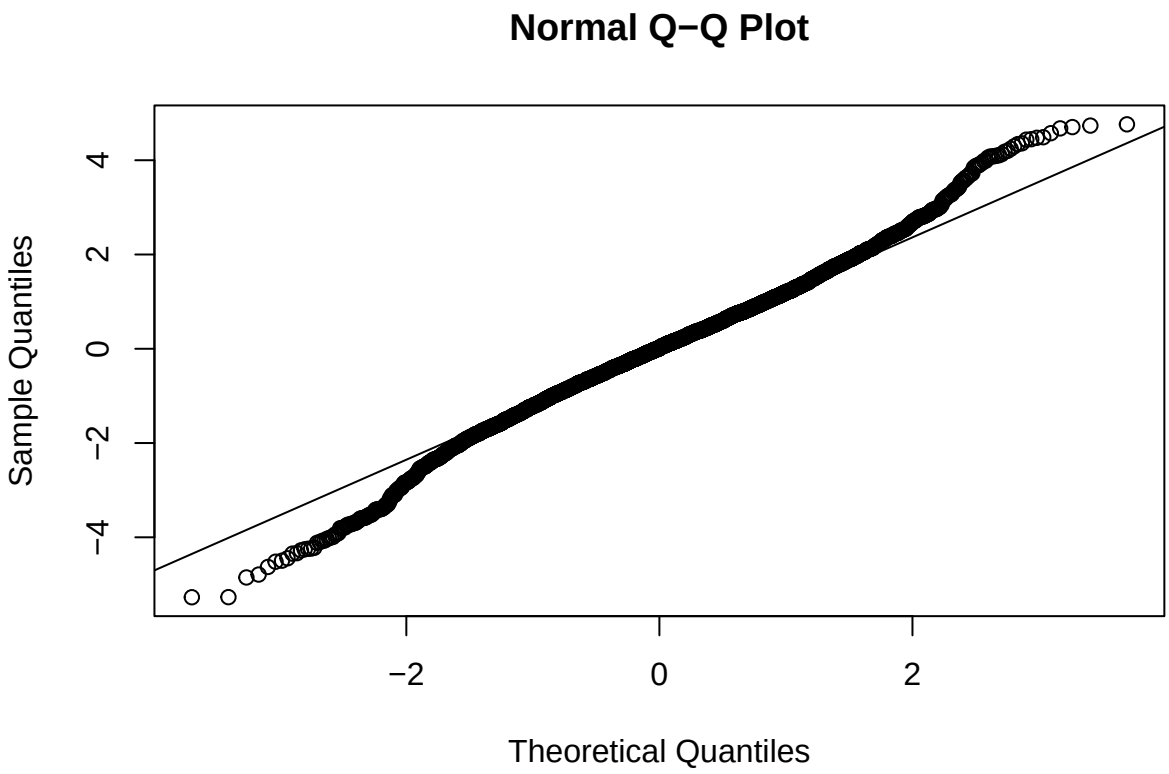
```
plot(init_model$residuals)
```




```
shapiro.test(init_model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  init_model$residuals
## W = 0.99207, p-value = 2.966e-15
```

```
{
  qqnorm(init_model$residuals)
  qqline(init_model$residuals)
}
```



D: Model Improvement

Model 3

From the previous anova table, we found that "Present_Tmin", "LDAPS_PPT3" and "Solar.radiation" were not significat, so we drop them.

```
data_v2 = select(data_v1, -c("Present_Tmin", "LDAPS_PPT3", "Solar.radiation"))
```

```
train_v2 <- data_v2[1:(n*0.60),]
```

```
model_v2<-lm(train_v2$Next_Tmax~. , data=train_v2)
summary(model_v2)
```

```
##
## Call:
## lm(formula = train_v2$Next_Tmax ~ ., data = train_v2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2626 -0.7909  0.0134  0.8009  4.7730
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)      -7.856e+02  5.429e+01 -14.472  < 2e-16 ***
## station2          8.412e-01  1.440e-01   5.843 5.48e-09 ***
## station3          6.390e-01  1.595e-01   4.006 6.28e-05 ***
## station4          1.712e+00  1.444e-01  11.852  < 2e-16 ***
## station5          1.062e+00  1.456e-01   7.291 3.61e-13 ***
## station6          1.168e+00  1.480e-01   7.896 3.58e-15 ***
## station7          1.179e+00  1.488e-01   7.925 2.85e-15 ***
## station8          1.190e+00  1.462e-01   8.139 5.11e-16 ***
## station9          1.695e+00  1.423e-01  11.914  < 2e-16 ***
## station10         1.468e+00  1.412e-01  10.393  < 2e-16 ***
## station11         1.192e+00  1.470e-01   8.111 6.42e-16 ***
## station12         1.235e+00  1.519e-01   8.129 5.55e-16 ***
## station13         1.093e+00  1.531e-01   7.144 1.05e-12 ***
## station14         1.120e+00  1.577e-01   7.105 1.40e-12 ***
## station15         9.379e-01  1.520e-01   6.171 7.40e-10 ***
## station16         6.322e-01  1.415e-01   4.468 8.07e-06 ***
## station17         9.236e-01  1.468e-01   6.293 3.41e-10 ***
## station18         2.430e+00  1.450e-01  16.753  < 2e-16 ***
## station19         1.108e+00  1.472e-01   7.527 6.25e-14 ***
## station20         2.219e+00  1.421e-01  15.619  < 2e-16 ***
## station21         3.988e-01  1.578e-01   2.527 0.011550 *
## station22         1.273e+00  1.445e-01   8.807  < 2e-16 ***
## station23         1.850e+00  1.439e-01  12.858  < 2e-16 ***
## station24         1.373e+00  1.482e-01   9.266  < 2e-16 ***
## station25         1.265e+00  1.571e-01   8.053 1.03e-15 ***
## Date              3.921e-01  2.690e-02  14.574  < 2e-16 ***
## Present_Tmax      5.794e-02  1.120e-02   5.173 2.40e-07 ***
## LDAPS_RHmin       3.689e-02  3.554e-03  10.379  < 2e-16 ***
## LDAPS_RHmax      -2.122e-02  4.041e-03  -5.252 1.58e-07 ***
## LDAPS_Tmax_lapse  6.944e-01  1.793e-02  38.721  < 2e-16 ***
## LDAPS_Tmin_lapse  1.877e-01  2.165e-02   8.667  < 2e-16 ***
## LDAPS_WS         -1.032e-01  1.018e-02 -10.136  < 2e-16 ***
## LDAPS_LH          5.390e-03  1.192e-03   4.522 6.29e-06 ***
## LDAPS_CC1        -1.771e+00  1.410e-01 -12.562  < 2e-16 ***
## LDAPS_CC2        -6.831e-01  1.781e-01  -3.834 0.000128 ***
## LDAPS_CC3        -7.140e-01  1.761e-01  -4.054 5.12e-05 ***
## LDAPS_CC4        -1.571e+00  1.378e-01 -11.399  < 2e-16 ***
## LDAPS_PPT1       -2.464e-02  1.098e-02  -2.243 0.024950 *
## LDAPS_PPT2        1.632e-01  1.286e-02  12.687  < 2e-16 ***
## LDAPS_PPT4        5.371e-02  2.318e-02   2.317 0.020542 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.318 on 4512 degrees of freedom
## Multiple R-squared:  0.7719, Adjusted R-squared:   0.77
## F-statistic: 391.6 on 39 and 4512 DF,  p-value: < 2.2e-16
```

From the anova table, we accept that all of the coefficients are significant, because their $p_{value} < \alpha = 0.05$. We also find that the Adjusted R-squared is now 0.77.

RMSE

The $RMSE_{valid} = 1.3834$ for this mode, which is 0.002 lower than the initial model.

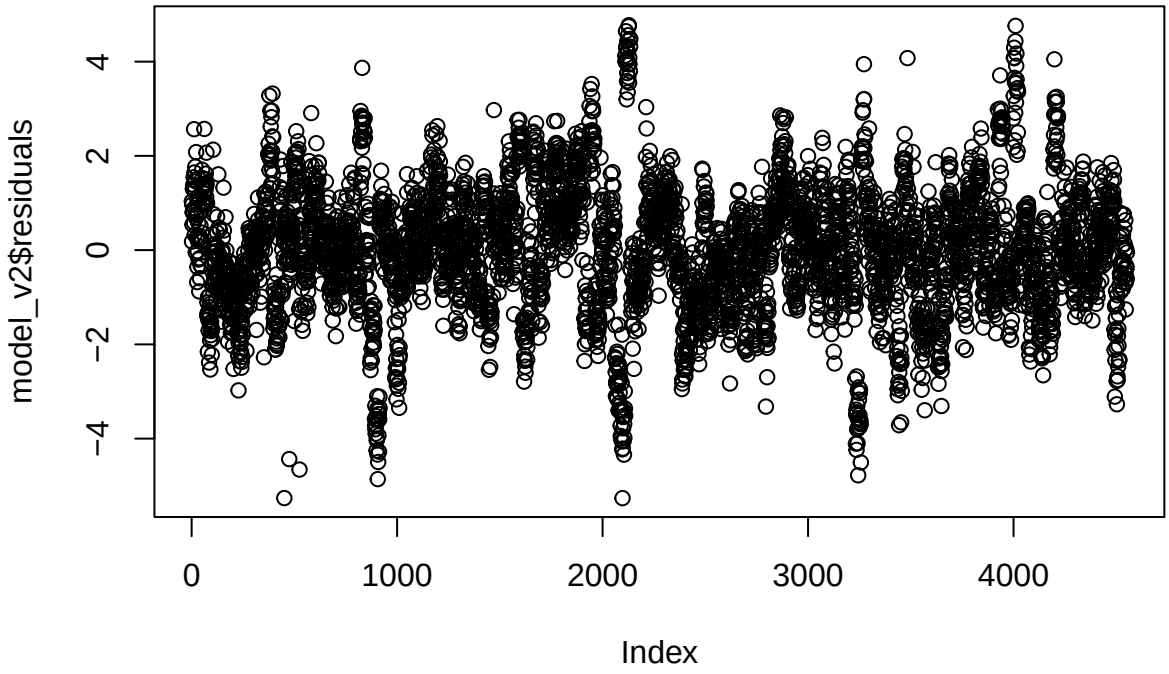
```
valid_prediction_v2 <-predict(model_v2,valid)
rmse_valid_v2 <- rmse(valid$Next_Tmax, valid_prediction_v2, nvalid)
round(rmse_valid_v2,4)
```

```
## [1] 1.3834
```

Residuals

The residuals are very similar to the previous model.

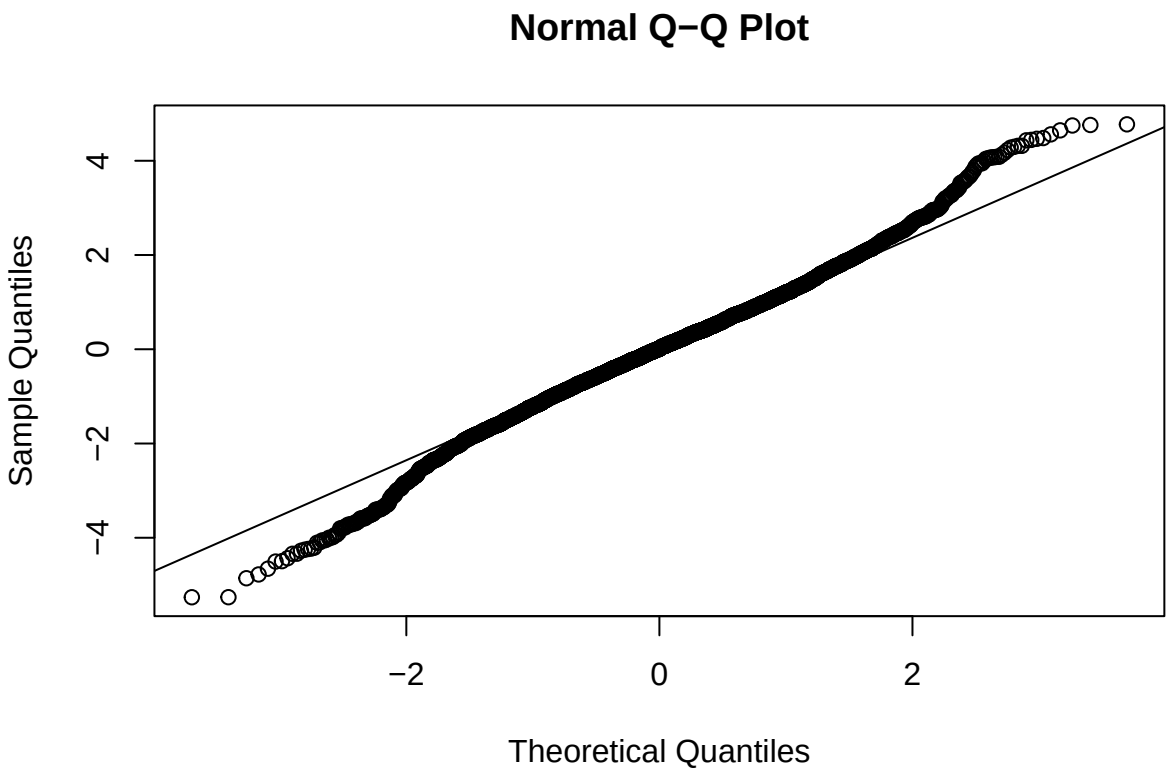
```
plot(model_v2$residuals)
```




```
shapiro.test(model_v2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model_v2$residuals
## W = 0.99224, p-value = 4.607e-15
```

```
{
  qqnorm(model_v2$residuals)
  qqline(model_v2$residuals)
}
```



Model 3

In this model, we're averaging all the possible features of model 2.

```
#model 3 includes average of CC and average of PPT's
train_v3<-train_v2
valid_v3<-valid


col1<-train_v3$LDAPS_CC1
col2<-train_v3$LDAPS_CC2
col3<-train_v3$LDAPS_CC3
col4<-train_v3$LDAPS_CC4
train_v3$CC_avg<-(col1+col2+col3+col4)/4


col1<-valid_v3$LDAPS_CC1
col2<-valid_v3$LDAPS_CC2
col3<-valid_v3$LDAPS_CC3
col4<-valid_v3$LDAPS_CC4
valid_v3$CC_avg<-(col1+col2+col3+col4)/4

col1<-train_v3$LDAPS_PPT1
col2<-train_v3$LDAPS_PPT2
col4<-train_v3$LDAPS_PPT4
train_v3$PPT_avg<-(col1+col2+col4)/3
```

```r
col1<-valid_v3$LDAPS_PPT1
col2<-valid_v3$LDAPS_PPT2
col4<-valid_v3$LDAPS_PPT4
valid_v3$PPT_avg<-(col1+col2+col4)/3

col1<-train_v3$LDAPS_RHmin
col2<-train_v3$LDAPS_RHmax
train_v3$RH_avg<-(col1+col2)/2

col1<-valid_v3$LDAPS_RHmin
col2<-valid_v3$LDAPS_RHmax
valid_v3$RH_avg<-(col1+col2)/2


col1<-train_v3$LDAPS_Tmax_lapse
col2<-train_v3$LDAPS_Tmin_lapse
train_v3$lapse_avg<-(col1+col2)/2

col1<-valid_v3$LDAPS_Tmax_lapse
col2<-valid_v3$LDAPS_Tmin_lapse
valid_v3$lapse_avg<-(col1+col2)/2


model_v3 = lm(Next_Tmax~station+Date+RH_avg+lapse_avg+LDAPS_WS+LDAPS_LH+CC_avg+PPT_avg,train_v3)
summary(model_v3)
```

```
##
## Call:
## lm(formula = Next_Tmax ~ station + Date + RH_avg + lapse_avg +
##     LDAPS_WS + LDAPS_LH + CC_avg + PPT_avg, data = train_v3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1939 -0.7640  0.0029  0.7680  5.2131
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.929e+02  5.632e+01 -15.855  < 2e-16 ***
## station2     8.863e-01  1.492e-01   5.939 3.09e-09 ***
## station3     5.165e-01  1.642e-01   3.145  0.00167 **
## station4     2.079e+00  1.469e-01  14.151  < 2e-16 ***
## station5     1.247e+00  1.518e-01   8.214 2.77e-16 ***
## station6     1.097e+00  1.530e-01   7.170 8.73e-13 ***
## station7     1.292e+00  1.537e-01   8.402  < 2e-16 ***
## station8     1.111e+00  1.511e-01   7.349 2.36e-13 ***
## station9     1.818e+00  1.471e-01  12.356  < 2e-16 ***
## station10    1.447e+00  1.464e-01   9.889  < 2e-16 ***
## station11    1.170e+00  1.520e-01   7.695 1.72e-14 ***
## station12    1.078e+00  1.572e-01   6.855 8.07e-12 ***
## station13    1.062e+00  1.572e-01   6.757 1.59e-11 ***
## station14    9.249e-01  1.623e-01   5.697 1.30e-08 ***
## station15    7.830e-01  1.574e-01   4.973 6.83e-07 ***
## station16    6.840e-01  1.478e-01   4.628 3.80e-06 ***
```

```
## station17    1.245e+00  1.524e-01   8.168 4.03e-16 ***
## station18    2.614e+00  1.485e-01  17.600  < 2e-16 ***
## station19    1.012e+00  1.524e-01   6.640 3.51e-11 ***
## station20    2.447e+00  1.460e-01  16.761  < 2e-16 ***
## station21    2.533e-01  1.629e-01   1.555  0.12006
## station22    1.304e+00  1.501e-01   8.692  < 2e-16 ***
## station23    1.834e+00  1.480e-01  12.391  < 2e-16 ***
## station24    1.255e+00  1.531e-01   8.197 3.17e-16 ***
## station25    1.109e+00  1.616e-01   6.861 7.74e-12 ***
## Date         4.467e-01  2.789e-02  16.014  < 2e-16 ***
## RH_avg      -7.006e-03  3.617e-03  -1.937  0.05282 .
## lapse_avg    9.287e-01  1.186e-02  78.313  < 2e-16 ***
## LDAPS_WS    -9.797e-02  1.028e-02  -9.532  < 2e-16 ***
## LDAPS_LH     3.057e-03  1.128e-03   2.711  0.00674 **
## CC_avg      -4.633e+00  1.622e-01 -28.562  < 2e-16 ***
## PPT_avg      1.855e-01  2.086e-02   8.890  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.385 on 4520 degrees of freedom
## Multiple R-squared:  0.7479, Adjusted R-squared:  0.7462
## F-statistic: 432.6 on 31 and 4520 DF,  p-value: < 2.2e-16
```

The "RH_avg" is not significant to this model, and the Adjusted R-squared = 0.7462 which is less than the previous models.

RMSE The $RMSE_{valid} = 1.31$ which is less than the previous models.

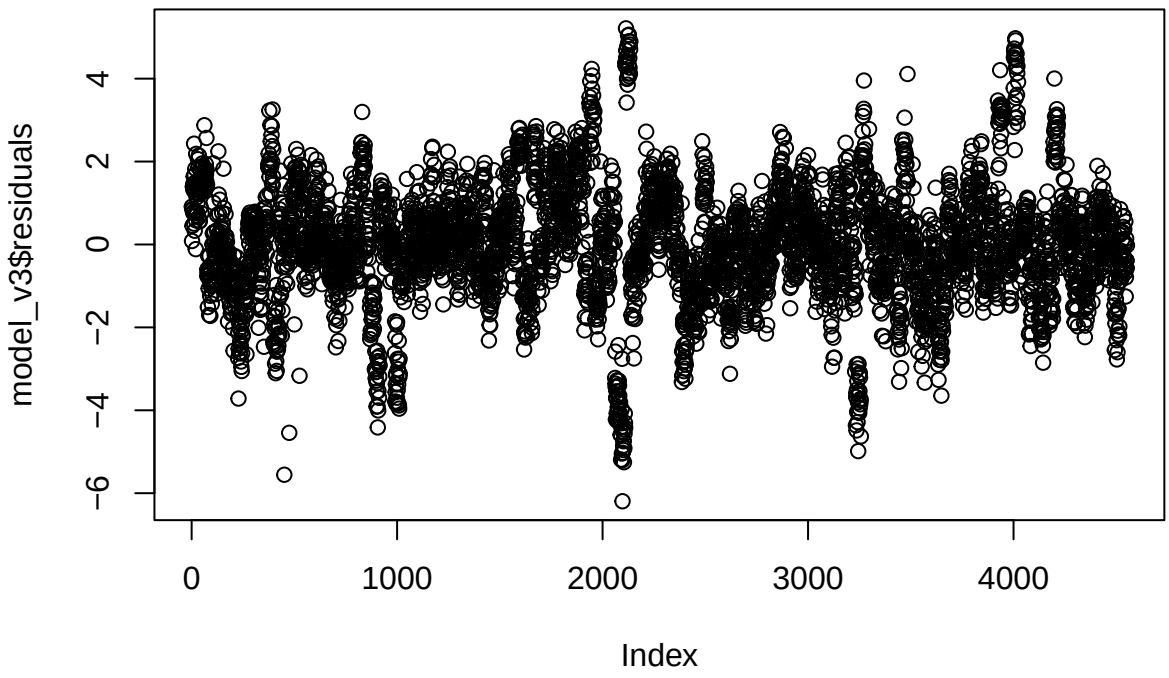
```
valid_v3<-valid_v3[c("station","Date","RH_avg","lapse_avg","LDAPS_WS","LDAPS_LH","CC_avg","PPT_avg")]

valid_prediction_v3 <-predict(model_v3,valid_v3)
rmse_valid_v3 <- rmse(valid$Next_Tmax, valid_prediction_v3,nvalid)
round(rmse_valid_v3,4)
```

```
## [1] 1.31
```

Residuals

The residuals are similar to the previous models.

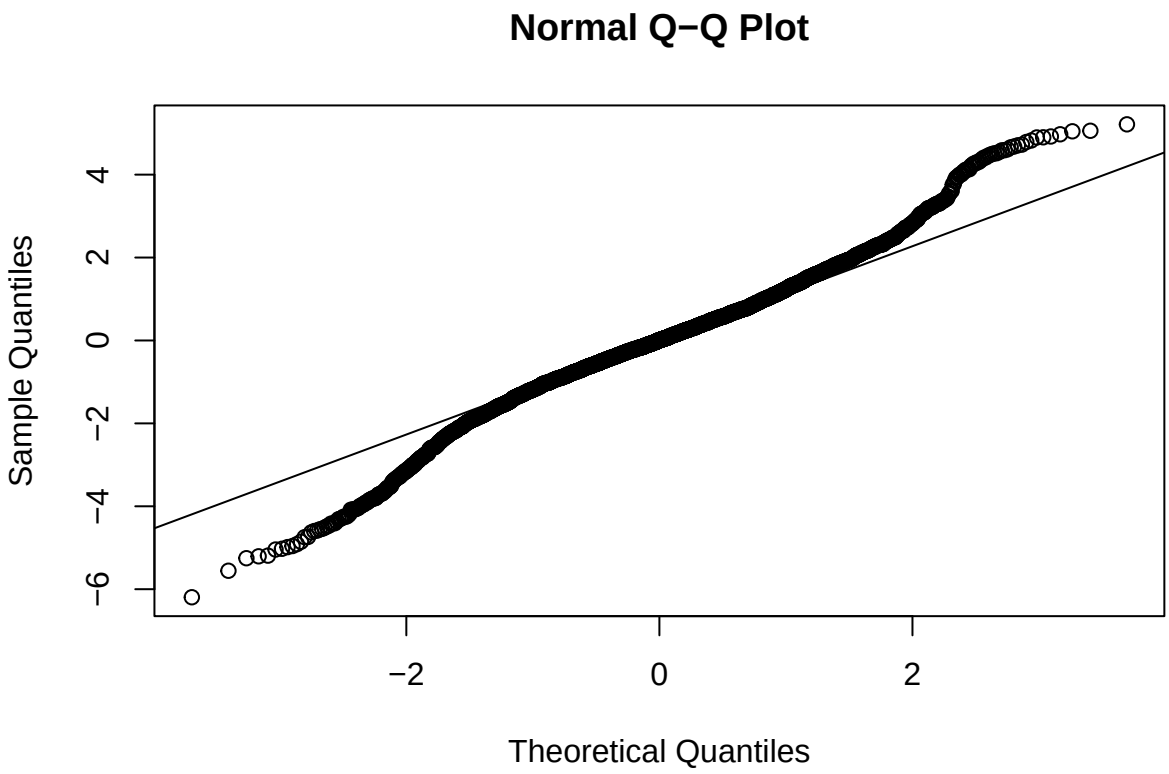
```
plot(model_v3$residuals)
```




```
shapiro.test(model_v3$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model_v3$residuals
## W = 0.98276, p-value < 2.2e-16
```

```
{
  qqnorm(model_v3$residuals)
  qqline(model_v3$residuals)
}
```



Model 4

In this model, we only choose features that would contribute to the maximum temperature using the scatter plots generated earlier & by examining the data description. The goal is to include the smallest number of variables that have a linear relashionship.

```
features <-c("LDAPS_LH","LDAPS_Tmin_lapse","LDAPS_CC2","LDAPS_PPT2","LDAPS_PPT1","LDAPS_CC1","LDAPS_RHma
data_base = select(data_v2, -c("LDAPS_LH","LDAPS_Tmin_lapse","LDAPS_CC2","LDAPS_PPT2","LDAPS_PPT1","LDAI
train_base <- data_base[1:(n*0.60),]

model_base<-lm(train_base$Next_Tmax~. , data=train_base)
summary(model_base)
```

```
##
## Call:
## lm(formula = train_base$Next_Tmax ~ ., data = train_base)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8861 -0.7967  0.0418  0.8550  5.4367
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -4.575e+02  5.406e+01  -8.464  < 2e-16 ***
## station2          4.780e-01  1.539e-01   3.105 0.001912 **
## station3          2.477e-02  1.549e-01   0.160 0.872985
```

```
## station4          1.523e+00  1.541e-01   9.881  < 2e-16 ***
## station5          9.346e-01  1.540e-01   6.067 1.41e-09 ***
## station6          7.279e-01  1.555e-01   4.682 2.93e-06 ***
## station7          1.299e+00  1.548e-01   8.390  < 2e-16 ***
## station8          7.625e-01  1.536e-01   4.965 7.13e-07 ***
## station9          1.522e+00  1.542e-01   9.868  < 2e-16 ***
## station10         1.136e+00  1.522e-01   7.464 1.00e-13 ***
## station11         7.032e-01  1.554e-01   4.525 6.19e-06 ***
## station12         7.236e-01  1.532e-01   4.725 2.38e-06 ***
## station13         4.332e-01  1.556e-01   2.785 0.005380 **
## station14         5.723e-01  1.545e-01   3.704 0.000215 ***
## station15         5.268e-01  1.549e-01   3.401 0.000676 ***
## station16         5.508e-01  1.536e-01   3.587 0.000338 ***
## station17         6.817e-01  1.550e-01   4.397 1.12e-05 ***
## station18         2.088e+00  1.566e-01  13.339  < 2e-16 ***
## station19         7.044e-01  1.540e-01   4.573 4.93e-06 ***
## station20         2.070e+00  1.532e-01  13.507  < 2e-16 ***
## station21        -1.560e-01  1.551e-01  -1.006 0.314342
## station22         1.004e+00  1.559e-01   6.441 1.31e-10 ***
## station23         1.699e+00  1.551e-01  10.959  < 2e-16 ***
## station24         1.017e+00  1.549e-01   6.562 5.91e-11 ***
## station25         6.936e-01  1.553e-01   4.465 8.19e-06 ***
## Date              2.290e-01  2.682e-02   8.541  < 2e-16 ***
## Present_Tmax      1.766e-01  9.599e-03  18.396  < 2e-16 ***
## LDAPS_Tmax_lapse  7.227e-01  1.217e-02  59.378  < 2e-16 ***
## LDAPS_WS         -9.906e-02  1.021e-02  -9.701  < 2e-16 ***
## LDAPS_CC3        -1.347e+00  1.110e-01 -12.139  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.439 on 4522 degrees of freedom
## Multiple R-squared:  0.7276, Adjusted R-squared:  0.7259
## F-statistic: 416.6 on 29 and 4522 DF,  p-value: < 2.2e-16
```

From the anova table we find that all the coefficients are significant, and that the Adjusted R-squared = 0.7259, less than the previous models.

RMSE

tHE $RMSE_{valid} = 1.2427$ which is less than the previous models.

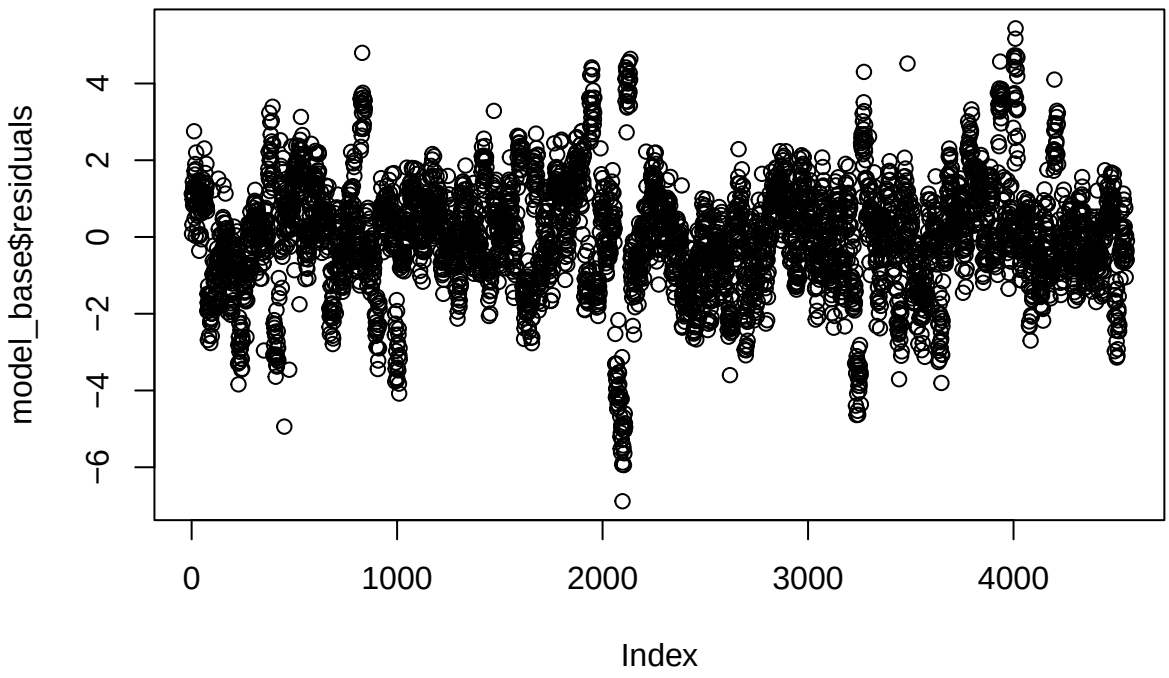
```
valid_prediction_base<-predict(model_base,valid)
rmse_valid_base <- rmse(valid$Next_Tmax, valid_prediction_base, nvalid)
round(rmse_valid_base,4)
```

```
## [1] 1.2427
```

Residuals

The residuals are the same as the previous models.

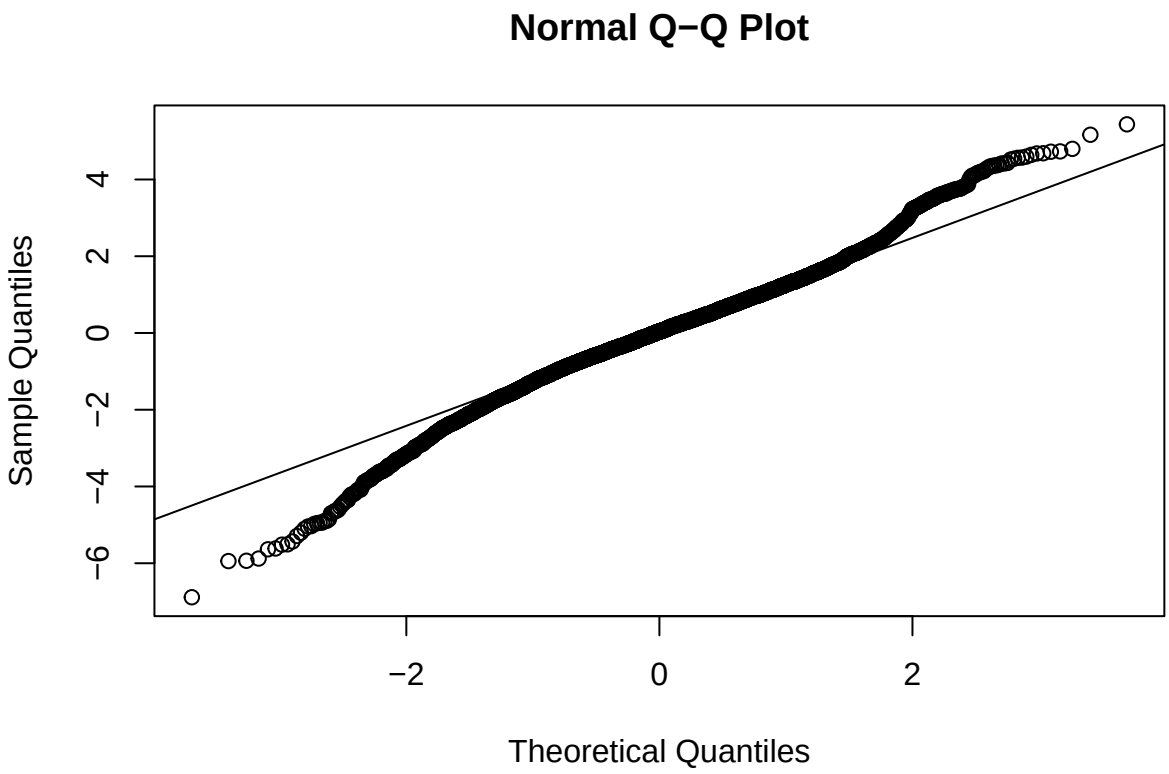
```
plot(model_base$residuals)
```




```
shapiro.test(model_base$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model_base$residuals
## W = 0.98539, p-value < 2.2e-16
```

```
{
  qqnorm(model_base$residuals)
  qqline(model_base$residuals)
}
```



Model 5: Model 4 forward propagation

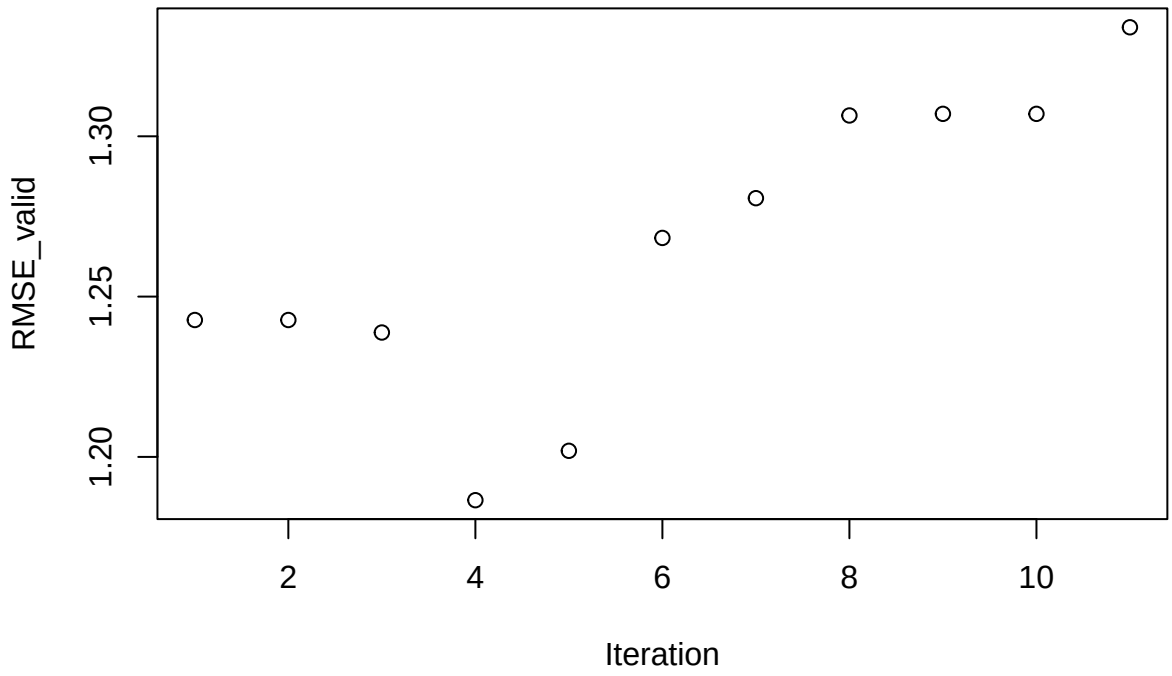
Now, using the previous model, we will do a forward propagation, and calculate each time the $RMSE_{valid}$

```
i<-1
rmse_list<-round(rmse_valid_base,4)

for (i in 1:length(features)){

  data_base = select(data_v2, -features[i:length(features)])
  train_base <- data_base[1:(n*0.60),]

  model_base<-lm(train_base$Next_Tmax~. , data=train_base)
  valid_prediction_base<-predict(model_base,valid)
  rmse_valid_base <- rmse(valid$Next_Tmax, valid_prediction_base, nvalid)
  rmse_list<-c(rmse_list, round(rmse_valid_base,4))
  print(paste("Iteration", i+1," , RMSE=", round(rmse_valid_base,4)))
  print(features[i:length(features)])
  }
```

```
## [1] "Iteration 2  , RMSE= 1.2427"
##  [1] "LDAPS_LH"          "LDAPS_Tmin_lapse" "LDAPS_CC2"         "LDAPS_PPT2"
##  [5] "LDAPS_PPT1"        "LDAPS_CC1"        "LDAPS_RHmax"       "LDAPS_PPT4"
##  [9] "LDAPS_CC4"         "LDAPS_RHmin"
## [1] "Iteration 3  , RMSE= 1.2388"
## [1] "LDAPS_Tmin_lapse" "LDAPS_CC2"        "LDAPS_PPT2"       "LDAPS_PPT1"
## [5] "LDAPS_CC1"        "LDAPS_RHmax"      "LDAPS_PPT4"       "LDAPS_CC4"
```

```
## [9] "LDAPS_RHmin"
## [1] "Iteration 4  , RMSE= 1.1865"
## [1] "LDAPS_CC2"   "LDAPS_PPT2"  "LDAPS_PPT1"  "LDAPS_CC1"   "LDAPS_RHmax"
## [6] "LDAPS_PPT4"  "LDAPS_CC4"   "LDAPS_RHmin"
## [1] "Iteration 5  , RMSE= 1.2019"
## [1] "LDAPS_PPT2"  "LDAPS_PPT1"  "LDAPS_CC1"   "LDAPS_RHmax" "LDAPS_PPT4"
## [6] "LDAPS_CC4"   "LDAPS_RHmin"
## [1] "Iteration 6  , RMSE= 1.2683"
## [1] "LDAPS_PPT1"  "LDAPS_CC1"   "LDAPS_RHmax" "LDAPS_PPT4"  "LDAPS_CC4"
## [6] "LDAPS_RHmin"
## [1] "Iteration 7  , RMSE= 1.2807"
## [1] "LDAPS_CC1"   "LDAPS_RHmax" "LDAPS_PPT4"  "LDAPS_CC4"   "LDAPS_RHmin"
## [1] "Iteration 8  , RMSE= 1.3065"
## [1] "LDAPS_RHmax" "LDAPS_PPT4"  "LDAPS_CC4"   "LDAPS_RHmin"
## [1] "Iteration 9  , RMSE= 1.307"
## [1] "LDAPS_PPT4"  "LDAPS_CC4"   "LDAPS_RHmin"
## [1] "Iteration 10  , RMSE= 1.307"
## [1] "LDAPS_CC4"   "LDAPS_RHmin"
## [1] "Iteration 11  , RMSE= 1.334"
## [1] "LDAPS_RHmin"
```

```
plot(rmse_list, xlab="Iteration", ylab="RMSE_valid")
```



We find that the iteration 4 has the least $RMSE_{valid}$ value, therefore, we take its features for the improved model.

```
data_improved = select(data_v2, -c("LDAPS_CC2" ,"LDAPS_PPT2" , "LDAPS_PPT1" , "LDAPS_CC1" ,  "LDAPS_RHma

train_improved <- data_improved[1:(n*0.60),]

improved_model<-lm(train_improved$Next_Tmax~. , data=train_improved)
valid_prediction_improved <-predict(improved_model,valid)
rmse_valid_improved <- rmse(valid$Next_Tmax, valid_prediction_improved, nvalid)
round(rmse_valid_improved,4)
```

```
## [1] 1.1865
```

```
summary(improved_model)
```

```
##
## Call:
## lm(formula = train_improved$Next_Tmax ~ ., data = train_improved)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9355 -0.7672  0.0453  0.8293  5.1982
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -6.513e+02  5.641e+01 -11.546  < 2e-16 ***
## station2          4.775e-01  1.535e-01   3.111 0.001875 **
## station3         -2.959e-03  1.657e-01  -0.018 0.985758
## station4          1.624e+00  1.526e-01  10.642  < 2e-16 ***
## station5          9.554e-01  1.564e-01   6.108 1.09e-09 ***
## station6          6.908e-01  1.557e-01   4.436 9.39e-06 ***
## station7          1.210e+00  1.594e-01   7.589 3.90e-14 ***
## station8          7.256e-01  1.549e-01   4.685 2.89e-06 ***
## station9          1.506e+00  1.526e-01   9.865  < 2e-16 ***
## station10         1.169e+00  1.512e-01   7.729 1.33e-14 ***
## station11         6.882e-01  1.550e-01   4.439 9.27e-06 ***
## station12         7.049e-01  1.592e-01   4.428 9.73e-06 ***
## station13         4.587e-01  1.601e-01   2.865 0.004189 **
## station14         4.991e-01  1.639e-01   3.044 0.002346 **
## station15         4.387e-01  1.591e-01   2.758 0.005836 **
## station16         5.196e-01  1.520e-01   3.418 0.000637 ***
## station17         7.859e-01  1.569e-01   5.009 5.68e-07 ***
## station18         2.131e+00  1.550e-01  13.746  < 2e-16 ***
## station19         6.548e-01  1.556e-01   4.208 2.62e-05 ***
## station20         2.153e+00  1.517e-01  14.195  < 2e-16 ***
## station21        -2.048e-01  1.643e-01  -1.247 0.212557
## station22         9.677e-01  1.542e-01   6.277 3.78e-10 ***
## station23         1.601e+00  1.536e-01  10.426  < 2e-16 ***
## station24         9.245e-01  1.560e-01   5.925 3.36e-09 ***
## station25         6.288e-01  1.629e-01   3.860 0.000115 ***
## Date              3.250e-01  2.798e-02  11.617  < 2e-16 ***
## Present_Tmax      1.394e-01  1.030e-02  13.532  < 2e-16 ***
## LDAPS_Tmax_lapse  6.519e-01  1.412e-02  46.168  < 2e-16 ***
## LDAPS_Tmin_lapse  1.653e-01  1.598e-02  10.339  < 2e-16 ***
```

```
## LDAPS_WS         -1.106e-01  1.070e-02 -10.337  < 2e-16 ***
## LDAPS_LH         -1.868e-04  1.192e-03  -0.157 0.875443
## LDAPS_CC3        -1.832e+00  1.198e-01 -15.299  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.421 on 4520 degrees of freedom
## Multiple R-squared:  0.7344, Adjusted R-squared:  0.7326
## F-statistic: 403.2 on 31 and 4520 DF,  p-value: < 2.2e-16
```

Using the anova table, we foind that LDAPS_LH is not significant. We can drop it, and have our final improved model.

Model 6

```
improved_data_2 = select(data_improved, -c("LDAPS_LH"))
improved_train_2 <- improved_data_2[1:(n*0.60),]

improved_model_final<-lm(improved_train_2$Next_Tmax~. , data=improved_train_2)
summary(improved_model_final)
```

```
##
## Call:
## lm(formula = improved_train_2$Next_Tmax ~ ., data = improved_train_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9395 -0.7648  0.0438  0.8298  5.1966
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -6.508e+02  5.633e+01 -11.554  < 2e-16 ***
## station2          4.808e-01  1.520e-01   3.163 0.001573 **
## station3          7.016e-03  1.530e-01   0.046 0.963436
## station4          1.623e+00  1.525e-01  10.645  < 2e-16 ***
## station5          9.497e-01  1.521e-01   6.243 4.70e-10 ***
## station6          6.948e-01  1.536e-01   4.524 6.22e-06 ***
## station7          1.203e+00  1.531e-01   7.855 4.96e-15 ***
## station8          7.305e-01  1.517e-01   4.815 1.52e-06 ***
## station9          1.504e+00  1.523e-01   9.874  < 2e-16 ***
## station10         1.172e+00  1.504e-01   7.792 8.13e-15 ***
## station11         6.916e-01  1.535e-01   4.507 6.75e-06 ***
## station12         7.126e-01  1.513e-01   4.711 2.53e-06 ***
## station13         4.657e-01  1.537e-01   3.031 0.002454 **
## station14         5.084e-01  1.527e-01   3.330 0.000876 ***
## station15         4.454e-01  1.531e-01   2.909 0.003643 **
## station16         5.180e-01  1.517e-01   3.415 0.000643 ***
## station17         7.808e-01  1.534e-01   5.090 3.72e-07 ***
## station18         2.129e+00  1.547e-01  13.767  < 2e-16 ***
## station19         6.598e-01  1.522e-01   4.336 1.48e-05 ***
## station20         2.154e+00  1.515e-01  14.216  < 2e-16 ***
## station21        -1.955e-01  1.532e-01  -1.276 0.201907
## station22         9.667e-01  1.540e-01   6.277 3.78e-10 ***
## station23         1.600e+00  1.534e-01  10.430  < 2e-16 ***
```

```
## station24         9.291e-01  1.532e-01   6.064 1.43e-09 ***
## station25         6.373e-01  1.535e-01   4.152 3.35e-05 ***
## Date              3.248e-01  2.794e-02  11.624  < 2e-16 ***
## Present_Tmax      1.391e-01  1.010e-02  13.766  < 2e-16 ***
## LDAPS_Tmax_lapse  6.514e-01  1.373e-02  47.438  < 2e-16 ***
## LDAPS_Tmin_lapse  1.659e-01  1.543e-02  10.751  < 2e-16 ***
## LDAPS_WS         -1.111e-01  1.015e-02 -10.954  < 2e-16 ***
## LDAPS_CC3        -1.830e+00  1.184e-01 -15.447  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.421 on 4521 degrees of freedom
## Multiple R-squared:  0.7344, Adjusted R-squared:  0.7327
## F-statistic: 416.7 on 30 and 4521 DF,  p-value: < 2.2e-16
```

The final model has no unsignificant coefficients, and its Adjusted R-squared is 0.7327.

RMSE

The final $RMSE_{valid} = 1.1865$, which is the lowest value that we got.

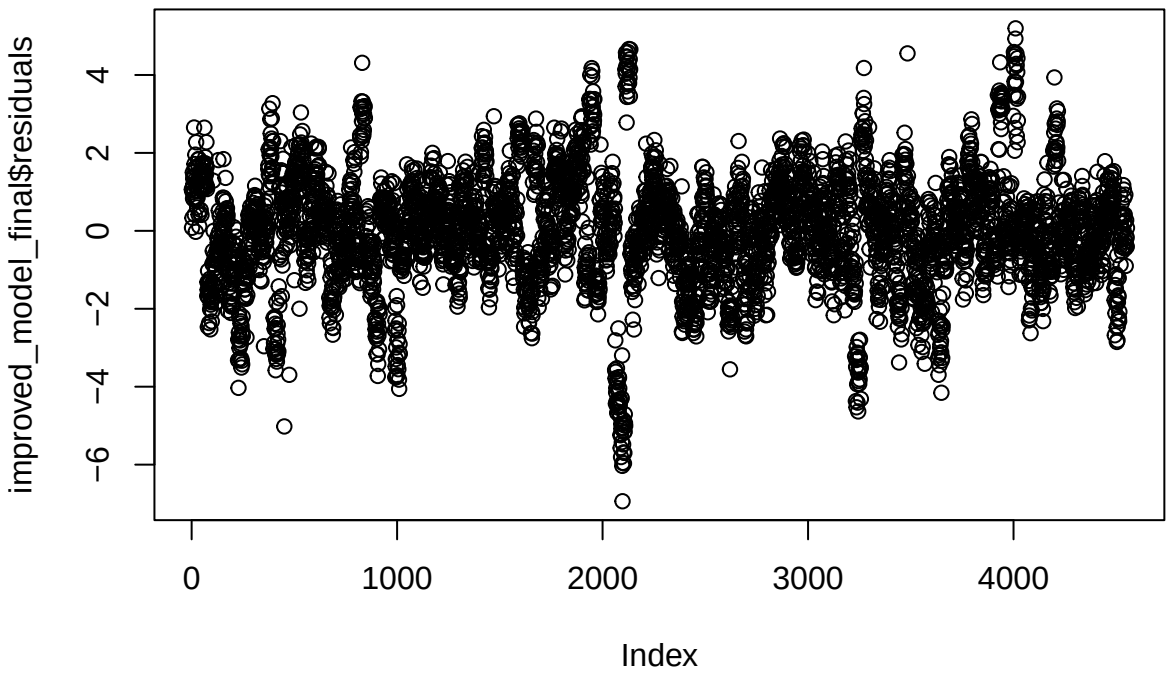
```
valid_prediction_improved_2 <-predict(improved_model_final,valid)
rmse_valid_improved_2 <- rmse(valid$Next_Tmax, valid_prediction_improved_2, nvalid)
round(rmse_valid_improved_2,4)
```

```
## [1] 1.1865
```

Residuals

The residuals did not change.

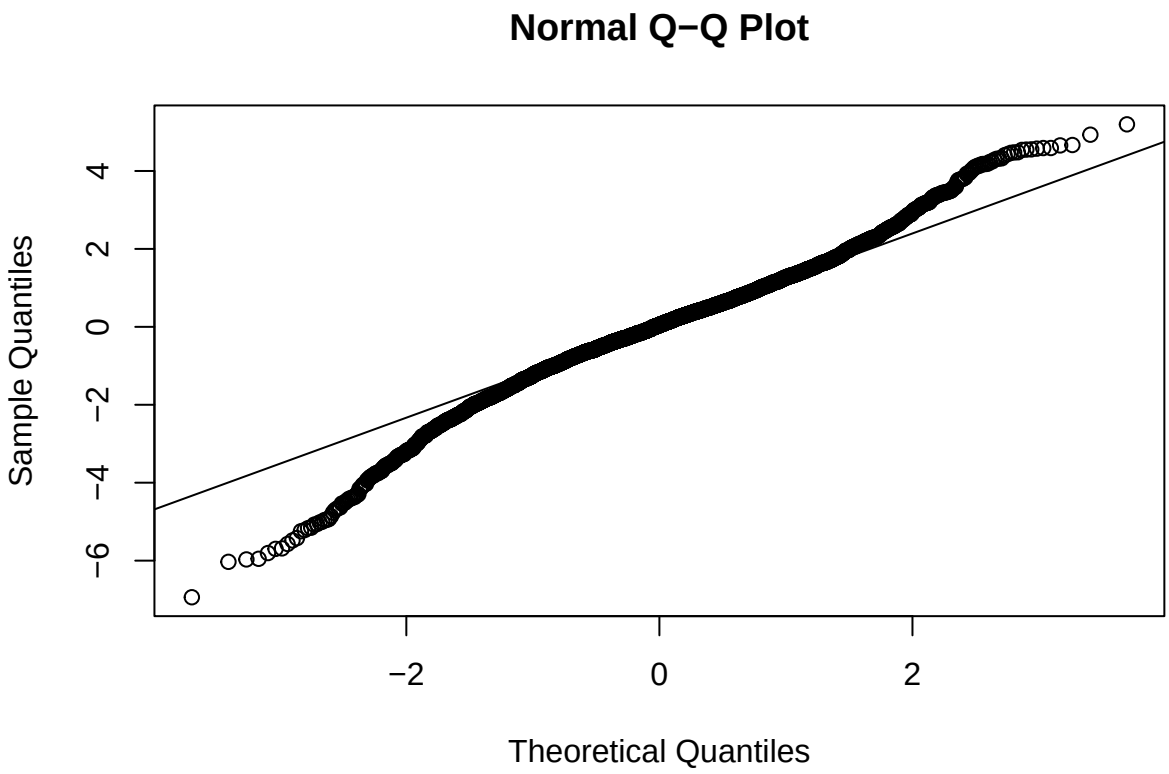
```
plot(improved_model_final$residuals)
```




```
shapiro.test(improved_model_final$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  improved_model_final$residuals
## W = 0.98291, p-value < 2.2e-16
```

```
{
  qqnorm(improved_model_final$residuals)
  qqline(improved_model_final$residuals)
}
```



E: Test Initial and Improved Models

We find that the $RMSE_{test,initial} = 2.1399 > RMSE_{test,improved} = 1.9263$, so we can confirm the model has improved.

```
test_initial <-predict(init_model,test)
rmse_test_initial<- rmse(test$Next_Tmax, test_initial, ntest)
print(paste("Initial Model RMSE test:",round(rmse_test_initial,4)))
```

```
## [1] "Initial Model RMSE test: 2.1399"
```

```
test_improved <-predict(improved_model_final,test)
rmse_test_improved <- rmse(test$Next_Tmax, test_improved, ntest)
print(paste("Improved Model RMSE test:",round(rmse_test_improved,4)))
```

```
## [1] "Improved Model RMSE test: 1.9263"
```