ADA Assignment 2 Jerbi Olfa 2021713094

Used dataset: Bias correction of numerical prediction model temperature forecast Data Set

Source :
https://archive.ics.uci.edu/ml/datasets/Bias+correction+of+numerical+prediction+model+temperature+forecast

Used program: R

Dataset description: "This data is for the purpose of bias correction of next-day maximum and minimum air temperatures forecast of the LDAPS model operated by the Korea Meteorological Administration over Seoul, South Korea. This data consists of summer data from 2013 to 2017. The input data is largely composed of the LDAPS model's next-day forecast data, in-situ maximum and minimum temperatures of present-day, and geographic auxiliary variables. There are two outputs (i.e. next-day maximum and minimum air temperatures) in this data. Hindcast validation was conducted for the period from 2015 to 2017."

Dataset exploration

##Data summary

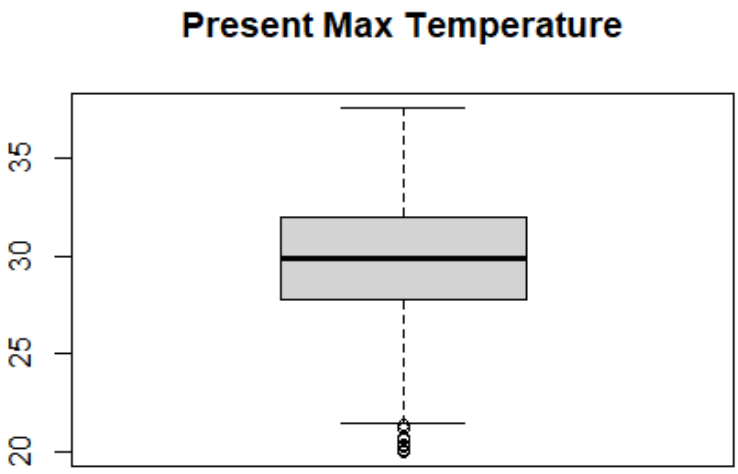
```
summary(data)
```

```
##     station        Date             Present_Tmax    Present_Tmin
##  Min.   : 1   Min.   :2013-06-30   Min.   :20.00   Min.   :11.30
##  1st Qu.: 7   1st Qu.:2014-07-15   1st Qu.:27.80   1st Qu.:21.70
##  Median :13   Median :2015-07-30   Median :29.90   Median :23.40
##  Mean   :13   Mean   :2015-07-30   Mean   :29.77   Mean   :23.23
##  3rd Qu.:19   3rd Qu.:2016-08-15   3rd Qu.:32.00   3rd Qu.:24.90
##  Max.   :25   Max.   :2017-08-30   Max.   :37.60   Max.   :29.90
##  NA's   :2    NA's   :2            NA's   :70      NA's   :70
##   LDAPS_RHmin     LDAPS_RHmax     LDAPS_Tmax_lapse LDAPS_Tmin_lapse
##  Min.   :19.79   Min.   : 58.94   Min.   :17.62    Min.   :14.27
##  1st Qu.:45.96   1st Qu.: 84.22   1st Qu.:27.67    1st Qu.:22.09
##  Median :55.04   Median : 89.79   Median :29.70    Median :23.76
##  Mean   :56.76   Mean   : 88.37   Mean   :29.61    Mean   :23.51
##  3rd Qu.:67.19   3rd Qu.: 93.74   3rd Qu.:31.71    3rd Qu.:25.15
##  Max.   :98.52   Max.   :100.00   Max.   :38.54    Max.   :29.62
##  NA's   :75      NA's   :75       NA's   :75       NA's   :75
##     LDAPS_WS         LDAPS_LH        LDAPS_CC1        LDAPS_CC2
##  Min.   : 2.883   Min.   :-13.60   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 5.679   1st Qu.: 37.27   1st Qu.:0.1467   1st Qu.:0.1406
##  Median : 6.547   Median : 56.87   Median :0.3157   Median :0.3124
##  Mean   : 7.098   Mean   : 62.51   Mean   :0.3688   Mean   :0.3561
```

```
##  3rd Qu.: 8.032   3rd Qu.: 84.22   3rd Qu.:0.5755   3rd Qu.:0.5587
##  Max.   :21.858   Max.   :213.41   Max.   :0.9673   Max.   :0.9684
##  NA's   :75       NA's   :75       NA's   :75       NA's   :75
##    LDAPS_CC3        LDAPS_CC4         LDAPS_PPT1          LDAPS_PPT2
##  Min.   :0.0000   Min.   :0.00000   Min.   : 0.00000   Min.   : 0.00000
##  1st Qu.:0.1014   1st Qu.:0.08153   1st Qu.: 0.00000   1st Qu.: 0.00000
##  Median :0.2626   Median :0.22766   Median : 0.00000   Median : 0.00000
##  Mean   :0.3184   Mean   :0.29919   Mean   : 0.59199   Mean   : 0.48500
##  3rd Qu.:0.4967   3rd Qu.:0.49949   3rd Qu.: 0.05252   3rd Qu.: 0.01836
##  Max.   :0.9838   Max.   :0.97471   Max.   :23.70154   Max.   :21.62166
##  NA's   :75       NA's   :75        NA's   :75         NA's   :75
##    LDAPS_PPT3        LDAPS_PPT4            lat             lon
##  Min.   : 0.0000   Min.   : 0.00000   Min.   :37.46   Min.   :126.8
##  1st Qu.: 0.0000   1st Qu.: 0.00000   1st Qu.:37.51   1st Qu.:126.9
##  Median : 0.0000   Median : 0.00000   Median :37.55   Median :127.0
##  Mean   : 0.2782   Mean   : 0.26941   Mean   :37.54   Mean   :127.0
##  3rd Qu.: 0.0079   3rd Qu.: 0.00004   3rd Qu.:37.58   3rd Qu.:127.0
##  Max.   :15.8412   Max.   :16.65547   Max.   :37.65   Max.   :127.1
##  NA's   :75        NA's   :75
##       DEM              Slope         Solar.radiation   Next_Tmax
##  Min.   : 12.37   Min.   :0.09847   Min.   :4330    Min.   :17.40
##  1st Qu.: 28.70   1st Qu.:0.27130   1st Qu.:4999    1st Qu.:28.20
##  Median : 45.72   Median :0.61800   Median :5436    Median :30.50
##  Mean   : 61.87   Mean   :1.25705   Mean   :5342    Mean   :30.27
##  3rd Qu.: 59.83   3rd Qu.:1.76780   3rd Qu.:5728    3rd Qu.:32.60
##  Max.   :212.34   Max.   :5.17823   Max.   :5993    Max.   :38.90
##                                                     NA's   :27
##    Next_Tmin
##  Min.   :11.30
##  1st Qu.:21.30
##  Median :23.10
##  Mean   :22.93
##  3rd Qu.:24.60
##  Max.   :29.80
##  NA's   :27
```

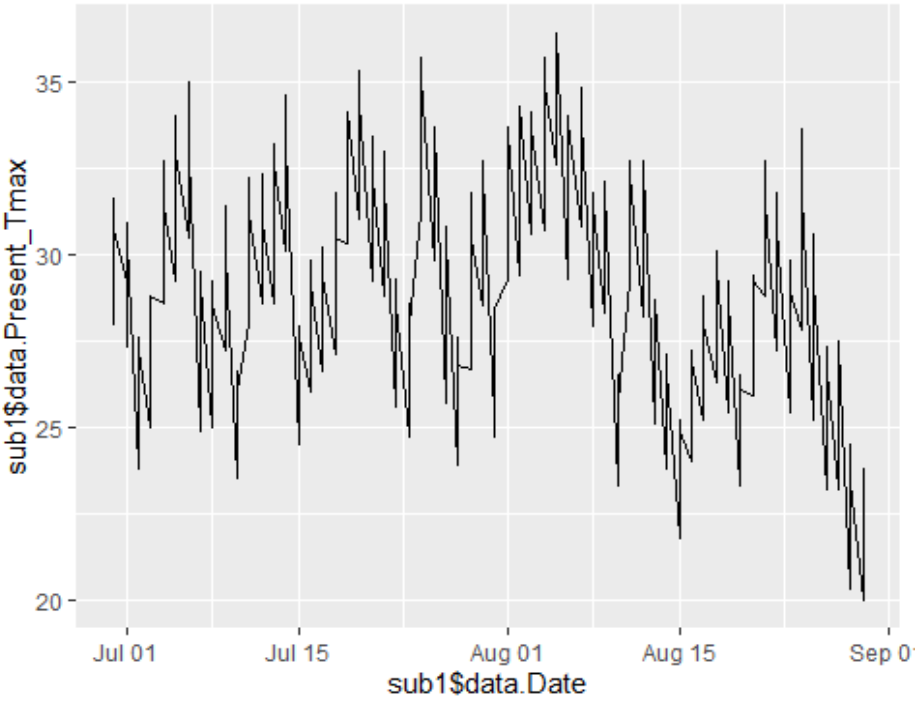
Descrptive stastics of variables

We start with boxplots for the present Max and Min temperature



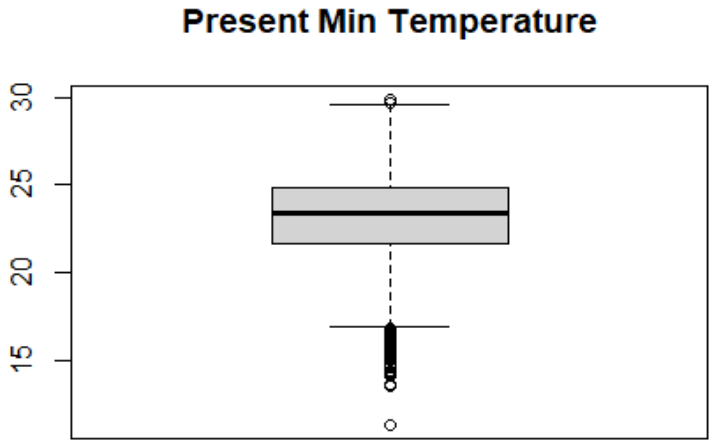
We can observe above that present Max temperature has an average of 30 degrees and it presents outliers where the temperature is close to 20 degrees. To explore the reason behind the existence of outliers we can trace Max temp in function of time for one of the given years:

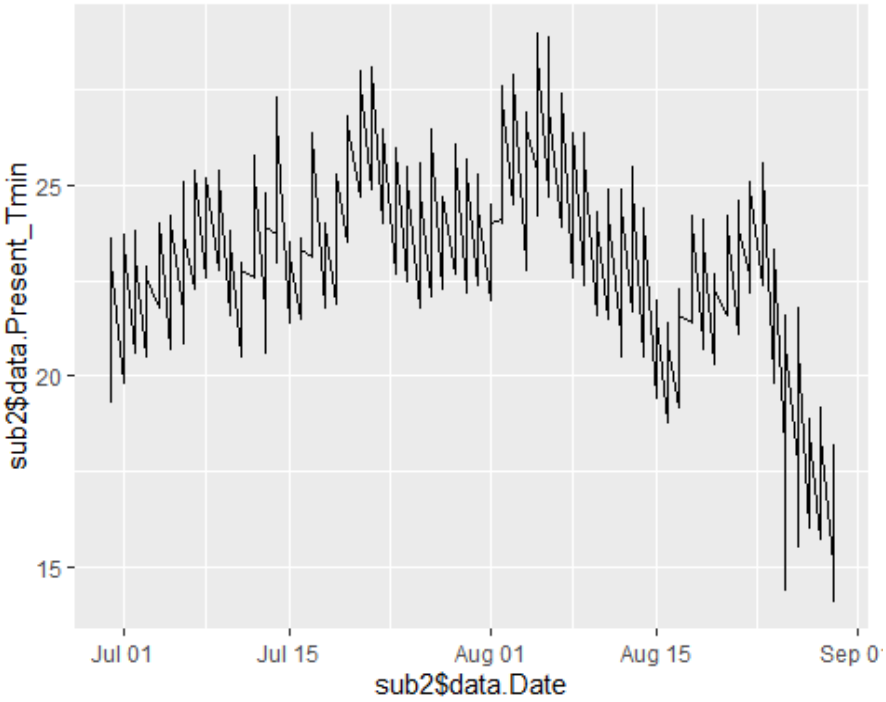
Rq: The last date present in the dataset is 8/30/2017, so we trace the temperature from 6/30/2017 to that most recent date to have a 1 summer's worth of data



We can observe that the drop in temperature starts around the beginning of September which marks the start of preparation of changing seasons (from summer to fall). After a hot summer, the earth takes time to cool down. South Korea being in the northern hemisphere, fall starts around the third week of September which can explain the gradual decrease of temperature starting from the end of August/ beginning of September, therefore explaining the presence of the outliers mentionned above.

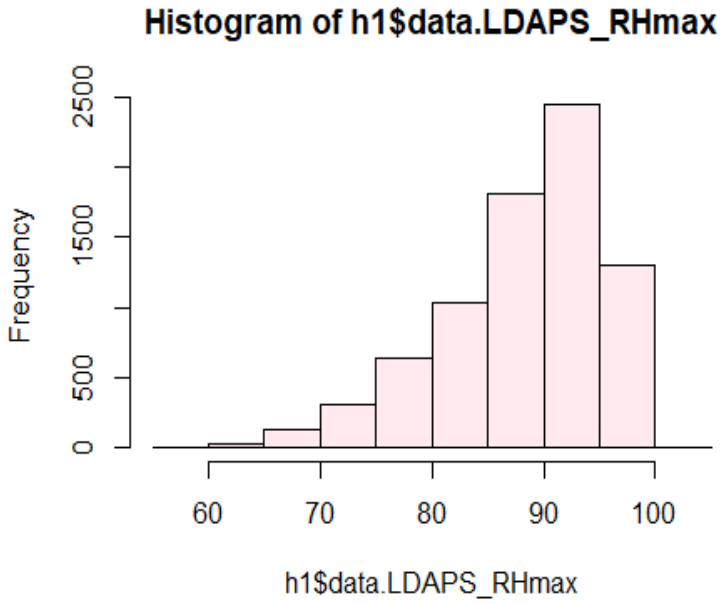
Let's also trace the boxplot and one summer's graph for the Min temperature:

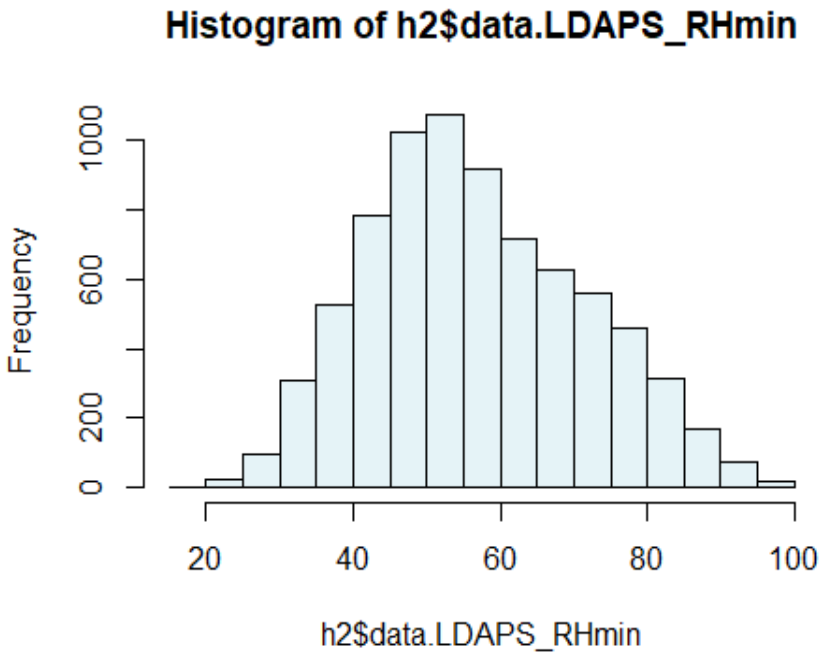




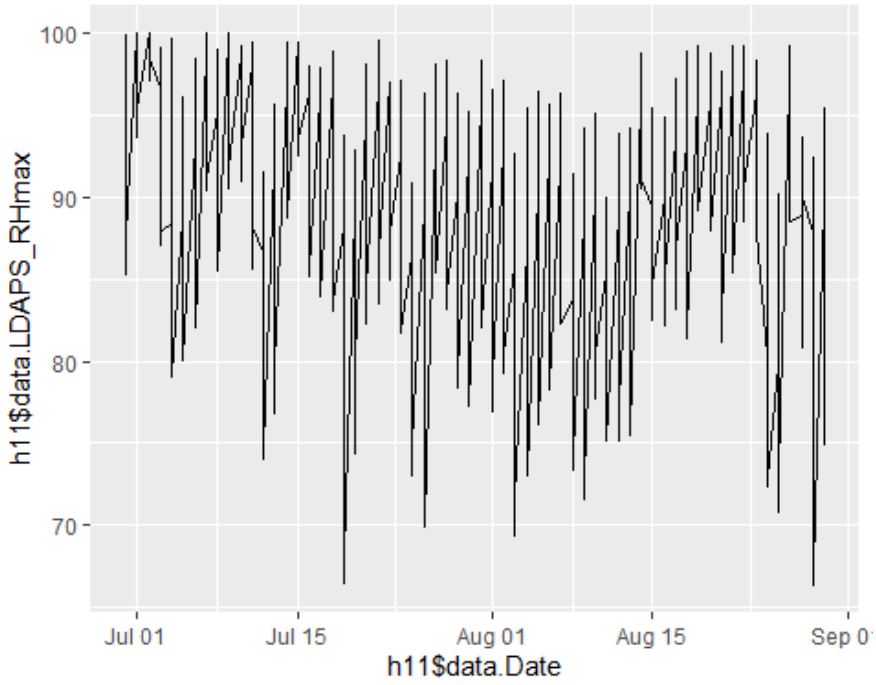
We can observe a similar trend for Min temperature where the average is around 23 degrees and there outliers around the value of 15 degrees. After tracing Min temperature in function of time we can again notice the decrease in temperature starts around September. This decrease can be explained in the same way as Max temperature.

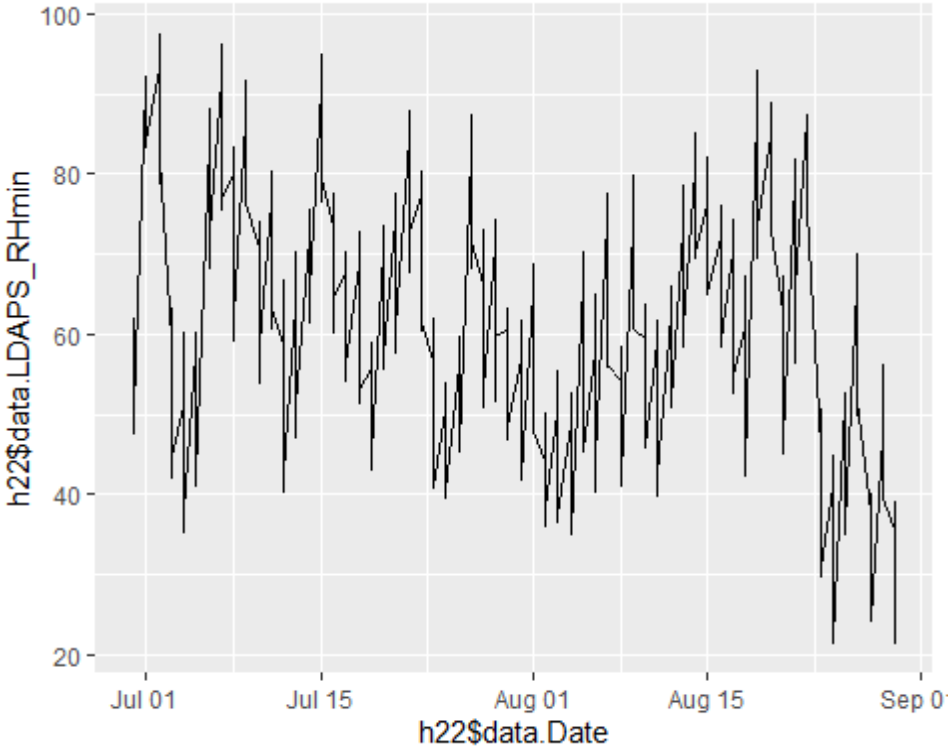
Let's now move to describing the Humidity factor:





When tracing the histograms of Max and Min humidity (traced above: red for Max and blue for min), we can find the following conclusions: -The maximum humidity varies between 65 and 100%, having an average of 90%. -The minimum humidity varies between 25 and 100%, having an average of 60%. We then trace one summer's worth of Max and Min humidity in function of time:





```
## [1] "variance of Max humidity = 51.7249230135225"

## [1] "variance of Min humidity = 215.153491123511"
```

We can observe that the humidity varies quite substantially (we calculate the variance for Max= 51.72 and Min=215.15) as well as being quite high. Seoul is known to have a wet and very humid climate during the summer season explaining the observed high values. We can also note that the humidity starts to decrease by the end of July and continues to do so till the end of August/ beginning of September, this marks the end of the rainy season in Korea (Jangma) that usually ends in middle to end of July.

Let's now check the correlation between some of our factors: 1-Correlation between present day Max temperature and next-day maximum air temperature applied lapse rate

```
## [1] "The correlation between present day Max temperature and next-day maxi
mum air temperature applied lapse rate  = 0.575288543419439"
```

1-Correlation between present day Min temperature and next-day minimum air temperature applied lapse rate

```
## [1] "The correlation between present day Min temperature and next-day mini
mum air temperature applied lapse rate  = 0.772920906770318"
```

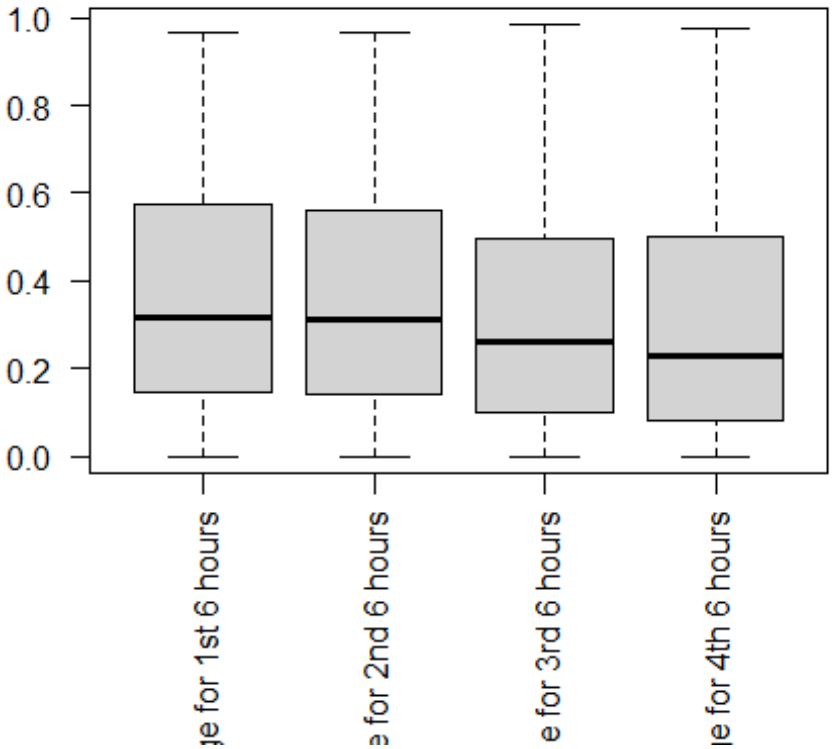
Observing the values above, we can see that the Min temperature of the present day is more highly correlated with the next day's Min temperature. This correlation is less pronounced for the for Max temperatures.

We do the same with wind speed too:

```
## [1] "The correlation between present day Min temperature and wind speed  =
-0.035110574143771"

## [1] "The correlation between present day Max temperature and wind speed  =
-0.122876054152989"
```

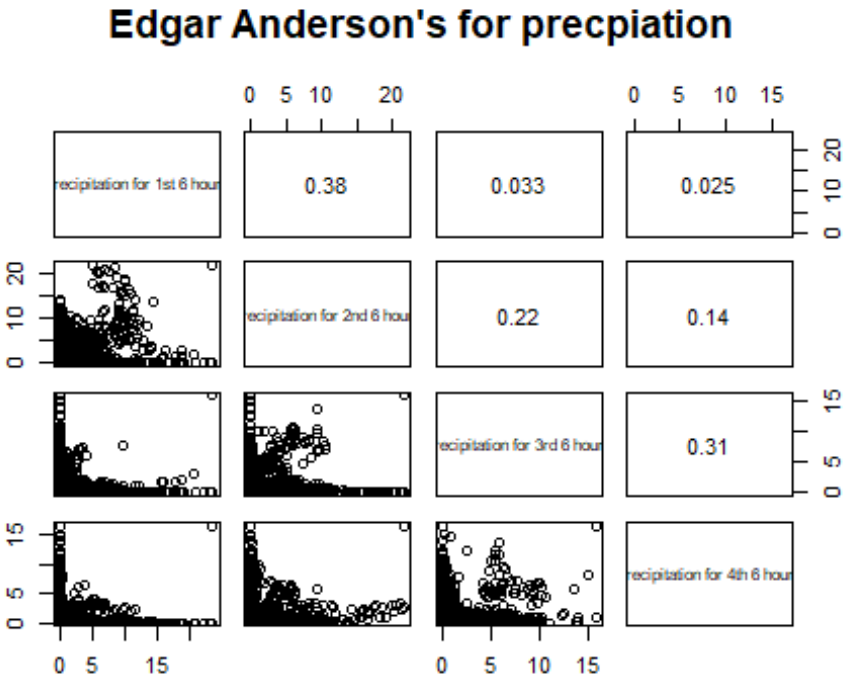
Correlation between temperatures and wind speed is low and negative, meaning present temperature has a low effect on next day's average wind speed.

For cloud cover:



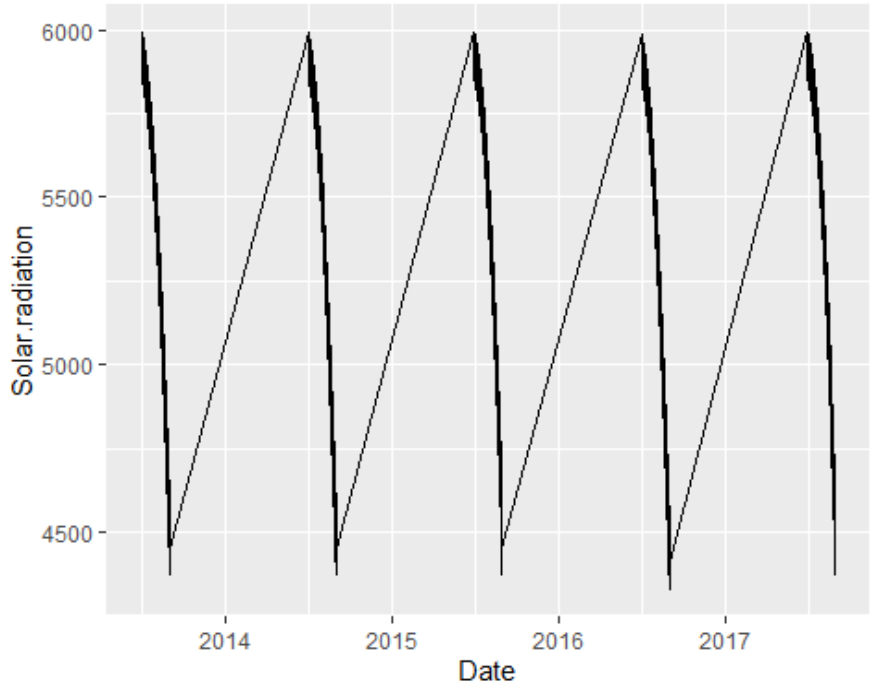
Through the boxplots above we can observe that the cloud coverage slowly decreases throughout the day.

For precipitation:

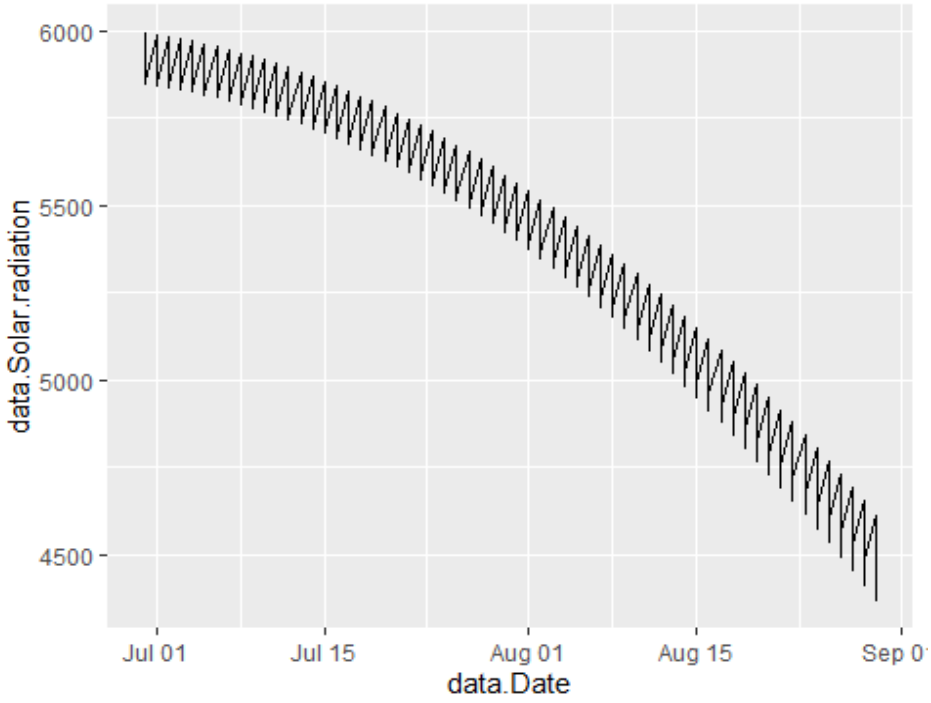


Through the graph above, we can observe that precipitations has a relatively low correlation: this correlation is at its highest between two consecutive periods (6 hours periods).

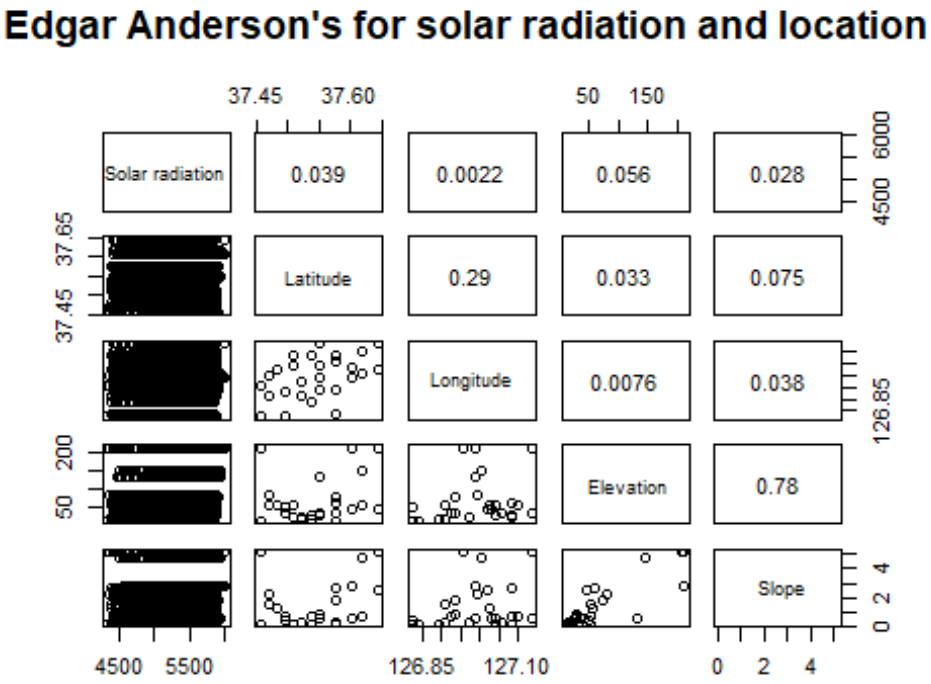
Let's now observe the solar radiation:



The solar radiation presents a yearly trend. Let's see when this trend peeks:



The solar radiation peeks in the beginning of July and constantly drops until reaching its minimum by the beginning of September. The amount of solar radiation that reaches any one spot on the Earth's surface varies according to: Geographic location Time of day Season Local landscape Local weather. Let's observe the correlation between the solar radiation and the geographic location then:



The low correlation between solar radiation and localization here can be explained by the fact that the present data is for Seoul only, meaning the variation of localization coordinates is limited. This could lead to masking the correlation between the two variables mentioned above.

Regression Models and Evaluation

For Max temperature

We start by using the function "regsubsets" on the dataset after excluding the next day's Min temperature. We start with the default settings that give us 8 estimated models:

```
## Subset selection object
## Call: regsubsets.formula(Next_Tmax ~ ., data = data[, -25])
## 23 Variables  (and intercept)
##                  Forced in Forced out
## station              FALSE      FALSE
## Date                 FALSE      FALSE
## Present_Tmax         FALSE      FALSE
```

```
## Present_Tmin         FALSE      FALSE
## LDAPS_RHmin          FALSE      FALSE
## LDAPS_RHmax          FALSE      FALSE
## LDAPS_Tmax_lapse     FALSE      FALSE
## LDAPS_Tmin_lapse     FALSE      FALSE
## LDAPS_WS             FALSE      FALSE
## LDAPS_LH             FALSE      FALSE
## LDAPS_CC1            FALSE      FALSE
## LDAPS_CC2            FALSE      FALSE
## LDAPS_CC3            FALSE      FALSE
## LDAPS_CC4            FALSE      FALSE
## LDAPS_PPT1           FALSE      FALSE
## LDAPS_PPT2           FALSE      FALSE
## LDAPS_PPT3           FALSE      FALSE
## LDAPS_PPT4           FALSE      FALSE
## lat                  FALSE      FALSE
## lon                  FALSE      FALSE
## DEM                  FALSE      FALSE
## Slope                FALSE      FALSE
## Solar.radiation      FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          station Date Present_Tmax Present_Tmin LDAPS_RHmin LDAPS_RHmax
## 1  ( 1 ) " "     " "  " "          " "          " "          " "
## 2  ( 1 ) " "     " "  "*"          " "          " "          " "
## 3  ( 1 ) " "     " "  "*"          " "          " "          " "
## 4  ( 1 ) " "     " "  "*"          " "          " "          " "
## 5  ( 1 ) " "     " "  "*"          " "          "*"          " "
## 6  ( 1 ) " "     " "  "*"          " "          "*"          " "
## 7  ( 1 ) " "     " "  "*"          " "          "*"          " "
## 8  ( 1 ) "*"     " "  "*"          " "          "*"          " "
##          LDAPS_Tmax_lapse LDAPS_Tmin_lapse LDAPS_WS LDAPS_LH LDAPS_CC1
## 1  ( 1 ) "*"              " "              " "      " "      " "
## 2  ( 1 ) "*"              " "              " "      " "      " "
## 3  ( 1 ) "*"              " "              " "      " "      " "
## 4  ( 1 ) "*"              " "              "*"      " "      " "
## 5  ( 1 ) "*"              " "              "*"      " "      " "
## 6  ( 1 ) "*"              " "              "*"      " "      "*"
## 7  ( 1 ) "*"              " "              "*"      "*"      "*"
## 8  ( 1 ) "*"              " "              "*"      "*"      "*"
##          LDAPS_CC2 LDAPS_CC3 LDAPS_CC4 LDAPS_PPT1 LDAPS_PPT2 LDAPS_PPT3
## 1  ( 1 ) " "       " "       " "       " "        " "        " "
## 2  ( 1 ) " "       " "       " "       " "        " "        " "
## 3  ( 1 ) " "       " "       "*"       " "        " "        " "
## 4  ( 1 ) " "       " "       "*"       " "        " "        " "
## 5  ( 1 ) " "       "*"       " "       " "        " "        " "
## 6  ( 1 ) " "       "*"       " "       " "        " "        " "
## 7  ( 1 ) " "       " "       "*"       " "        " "        " "
## 8  ( 1 ) " "       " "       "*"       " "        " "        " "
##          LDAPS_PPT4 lat lon DEM Slope Solar.radiation
```

```
## 1  ( 1 ) " "         " " " " " " " " " "    " "
## 2  ( 1 ) " "         " " " " " " " " " "    " "
## 3  ( 1 ) " "         " " " " " " " " " "    " "
## 4  ( 1 ) " "         " " " " " " " " " "    " "
## 5  ( 1 ) " "         " " " " " " " " " "    " "
## 6  ( 1 ) " "         " " " " " " " " " "    " "
## 7  ( 1 ) " "         " " " " " " " " " "    " "
## 8  ( 1 ) " "         " " " " " " " " " "    " "
```

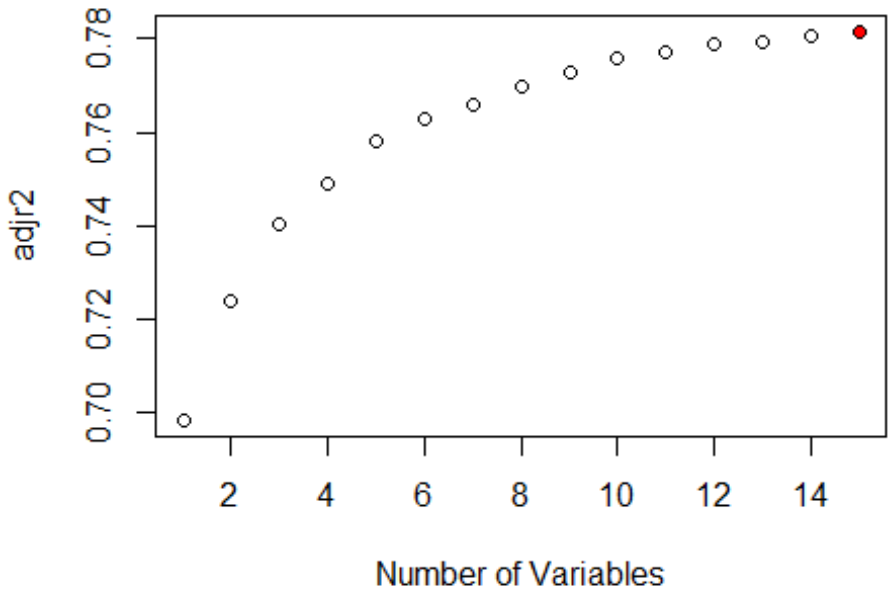
We can observe above the results of this first trial where for each model we can see the variables (that have * in front of them) that are included in each of the 8 models.

Given the relatively large number of variables we can also seek to build more models. Let's build 15 models for example.

```
## Subset selection object
## Call: regsubsets.formula(Next_Tmax ~ ., data = data[, -25], nvmax = 15)
## 23 Variables  (and intercept)
##                  Forced in Forced out
## station              FALSE      FALSE
## Date                 FALSE      FALSE
## Present_Tmax         FALSE      FALSE
## Present_Tmin         FALSE      FALSE
## LDAPS_RHmin          FALSE      FALSE
## LDAPS_RHmax          FALSE      FALSE
## LDAPS_Tmax_lapse     FALSE      FALSE
## LDAPS_Tmin_lapse     FALSE      FALSE
## LDAPS_WS             FALSE      FALSE
## LDAPS_LH             FALSE      FALSE
## LDAPS_CC1            FALSE      FALSE
## LDAPS_CC2            FALSE      FALSE
## LDAPS_CC3            FALSE      FALSE
## LDAPS_CC4            FALSE      FALSE
## LDAPS_PPT1           FALSE      FALSE
## LDAPS_PPT2           FALSE      FALSE
## LDAPS_PPT3           FALSE      FALSE
## LDAPS_PPT4           FALSE      FALSE
## lat                  FALSE      FALSE
## lon                  FALSE      FALSE
## DEM                  FALSE      FALSE
## Slope                FALSE      FALSE
## Solar.radiation      FALSE      FALSE
## 1 subsets of each size up to 15
## Selection Algorithm: exhaustive
##          station Date Present_Tmax Present_Tmin LDAPS_RHmin LDAPS_RHmax
## 1  ( 1 )  " "     " "  " "          " "          " "          " "
## 2  ( 1 )  " "     " "  "*"          " "          " "          " "
## 3  ( 1 )  " "     " "  "*"          " "          " "          " "
## 4  ( 1 )  " "     " "  "*"          " "          " "          " "
## 5  ( 1 )  " "     " "  "*"          " "          "*"          " "
```

```
## 6  ( 1 )  " "     " "  "*"          " "          "*"          " "
## 7  ( 1 )  " "     " "  "*"          " "          "*"          " "
## 8  ( 1 )  "*"     " "  "*"          " "          "*"          " "
## 9  ( 1 )  "*"     " "  "*"          " "          "*"          " "
## 10  ( 1 ) "*"     " "  "*"          " "          "*"          " "
## 11  ( 1 ) "*"     " "  "*"          " "          "*"          " "
## 12  ( 1 ) "*"     "*"  "*"          " "          "*"          " "
## 13  ( 1 ) "*"     "*"  "*"          " "          "*"          " "
## 14  ( 1 ) "*"     "*"  "*"          " "          "*"          " "
## 15  ( 1 ) "*"     "*"  "*"          " "          "*"          " "
##           LDAPS_Tmax_lapse LDAPS_Tmin_lapse LDAPS_WS LDAPS_LH LDAPS_CC1
## 1  ( 1 )  "*"              " "              " "       " "       " "
## 2  ( 1 )  "*"              " "              " "       " "       " "
## 3  ( 1 )  "*"              " "              " "       " "       " "
## 4  ( 1 )  "*"              " "              "*"       " "       " "
## 5  ( 1 )  "*"              " "              "*"       " "       " "
## 6  ( 1 )  "*"              " "              "*"       " "       "*"
## 7  ( 1 )  "*"              " "              "*"       "*"       "*"
## 8  ( 1 )  "*"              " "              "*"       "*"       "*"
## 9  ( 1 )  "*"              " "              "*"       "*"       "*"
## 10  ( 1 ) "*"              " "              "*"       "*"       "*"
## 11  ( 1 ) "*"              " "              "*"       "*"       "*"
## 12  ( 1 ) "*"              " "              "*"       "*"       "*"
## 13  ( 1 ) "*"              "*"              "*"       "*"       "*"
## 14  ( 1 ) "*"              " "              "*"       "*"       "*"
## 15  ( 1 ) "*"              "*"              "*"       "*"       "*"
##           LDAPS_CC2 LDAPS_CC3 LDAPS_CC4 LDAPS_PPT1 LDAPS_PPT2 LDAPS_PPT3
## 1  ( 1 )  " "       " "       " "       " "        " "        " "
## 2  ( 1 )  " "       " "       " "       " "        " "        " "
## 3  ( 1 )  " "       " "       "*"       " "        " "        " "
## 4  ( 1 )  " "       " "       "*"       " "        " "        " "
## 5  ( 1 )  " "       "*"       " "       " "        " "        " "
## 6  ( 1 )  " "       "*"       " "       " "        " "        " "
## 7  ( 1 )  " "       " "       "*"       " "        " "        " "
## 8  ( 1 )  " "       " "       "*"       " "        " "        " "
## 9  ( 1 )  " "       " "       "*"       " "        "*"        " "
## 10  ( 1 ) " "       " "       "*"       " "        "*"        " "
## 11  ( 1 ) " "       "*"       "*"       " "        "*"        " "
## 12  ( 1 ) " "       "*"       "*"       " "        "*"        " "
## 13  ( 1 ) " "       "*"       "*"       " "        "*"        " "
## 14  ( 1 ) " "       "*"       "*"       " "        "*"        " "
## 15  ( 1 ) " "       "*"       "*"       " "        "*"        " "
##           LDAPS_PPT4 lat lon DEM Slope Solar.radiation
## 1  ( 1 )  " "        " " " " " " " "   " "
## 2  ( 1 )  " "        " " " " " " " "   " "
## 3  ( 1 )  " "        " " " " " " " "   " "
## 4  ( 1 )  " "        " " " " " " " "   " "
## 5  ( 1 )  " "        " " " " " " " "   " "
## 6  ( 1 )  " "        " " " " " " " "   " "
## 7  ( 1 )  " "        " " " " " " " "   " "
```

```
## 8  ( 1 )  " "         " " " " " " " " " "   " "
## 9  ( 1 )  " "         " " " " " " " " " "   " "
## 10  ( 1 ) " "         " " "*" " " " " "   " "
## 11  ( 1 ) " "         " " "*" " " " " "   " "
## 12  ( 1 ) " "         " " "*" " " " " "   " "
## 13  ( 1 ) " "         " " "*" " " " " "   " "
## 14  ( 1 ) " "         " " "*" "*" "*"   " "
## 15  ( 1 ) " "         " " "*" "*" "*"   " "

## [1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"

##      (Intercept)          station             Date     Present_Tmax
##     2.431063e+02     1.899891e-02     2.397056e-04     1.506151e-01
##      LDAPS_RHmin LDAPS_Tmax_lapse LDAPS_Tmin_lapse         LDAPS_WS
##     2.295777e-02     6.345570e-01     1.099465e-01    -1.398551e-01
##         LDAPS_LH        LDAPS_CC1        LDAPS_CC3        LDAPS_CC4
##     7.569735e-03    -1.393321e+00    -1.147784e+00    -1.178040e+00
##       LDAPS_PPT2              lon              DEM            Slope
##     1.204170e-01    -1.909909e+00    -4.148559e-03     1.780115e-01

## [1] 15

## [1] "the best model is model number 15 with an adjusted R square= 0.781496
475131345"
```



After building the 15 models, we can check the model with the highest Adjusted R square (in other words the best model). We find that model number 15 is the best one. The

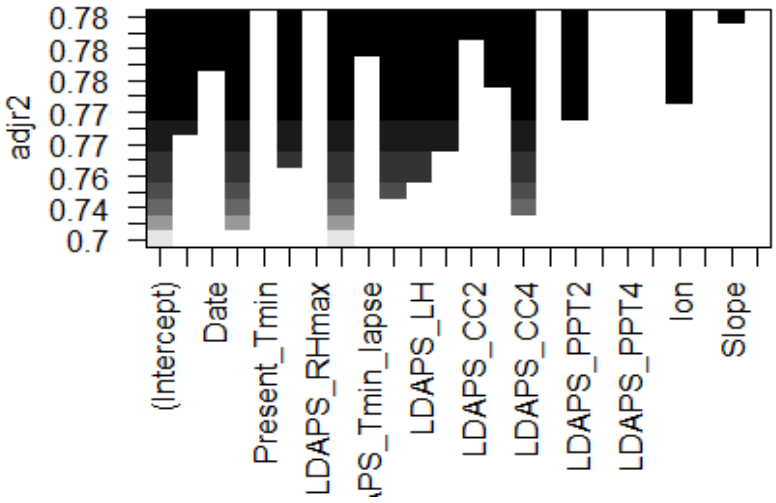
variables contained in this model are: Station, Date, Present Max Temperature, Next day's humidity, next-day maximum air temperature, next-day minimum air temperature, next day's wind speed,next-day average latent heat flux, split average cloud cover (for the 1st, 3rd and 4th 6 hours), precipitations for the 2nd 6 hours, longitude, elevation and slope. The coefficients for these variable are shown in the table above.

Let's now try this again with "forward" method:

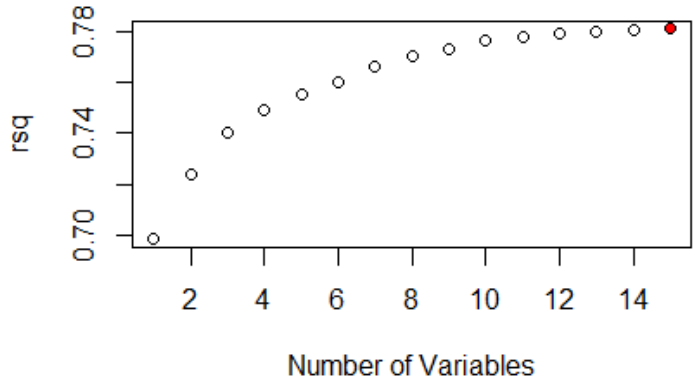
```
## Subset selection object
## Call: regsubsets.formula(Next_Tmax ~ ., data = data[, -25], nvmax = 15,
##     method = "forward")
## 23 Variables  (and intercept)
##                  Forced in Forced out
## station              FALSE      FALSE
## Date                 FALSE      FALSE
## Present_Tmax         FALSE      FALSE
## Present_Tmin         FALSE      FALSE
## LDAPS_RHmin          FALSE      FALSE
## LDAPS_RHmax          FALSE      FALSE
## LDAPS_Tmax_lapse     FALSE      FALSE
## LDAPS_Tmin_lapse     FALSE      FALSE
## LDAPS_WS             FALSE      FALSE
## LDAPS_LH             FALSE      FALSE
## LDAPS_CC1            FALSE      FALSE
## LDAPS_CC2            FALSE      FALSE
## LDAPS_CC3            FALSE      FALSE
## LDAPS_CC4            FALSE      FALSE
## LDAPS_PPT1           FALSE      FALSE
## LDAPS_PPT2           FALSE      FALSE
## LDAPS_PPT3           FALSE      FALSE
## LDAPS_PPT4           FALSE      FALSE
## lat                  FALSE      FALSE
## lon                  FALSE      FALSE
## DEM                  FALSE      FALSE
## Slope                FALSE      FALSE
## Solar.radiation      FALSE      FALSE
## 1 subsets of each size up to 15
## Selection Algorithm: forward
##           station Date Present_Tmax Present_Tmin LDAPS_RHmin LDAPS_RHmax
## 1  ( 1 )  " "     " "  " "          " "          " "         " "
## 2  ( 1 )  " "     " "  "*"          " "          " "         " "
## 3  ( 1 )  " "     " "  "*"          " "          " "         " "
## 4  ( 1 )  " "     " "  "*"          " "          " "         " "
## 5  ( 1 )  " "     " "  "*"          " "          " "         " "
## 6  ( 1 )  " "     " "  "*"          " "          "*"         " "
## 7  ( 1 )  " "     " "  "*"          " "          "*"         " "
## 8  ( 1 )  "*"     " "  "*"          " "          "*"         " "
## 9  ( 1 )  "*"     " "  "*"          " "          "*"         " "
## 10  ( 1 ) "*"     " "  "*"          " "          "*"         " "
## 11  ( 1 ) "*"     " "  "*"          " "          "*"         " "
```

```
## 12  ( 1 ) "*"     "*"  "*"          " "          "*"          " "
## 13  ( 1 ) "*"     "*"  "*"          " "          "*"          " "
## 14  ( 1 ) "*"     "*"  "*"          " "          "*"          " "
## 15  ( 1 ) "*"     "*"  "*"          " "          "*"          " "
##           LDAPS_Tmax_lapse LDAPS_Tmin_lapse LDAPS_WS LDAPS_LH LDAPS_CC1
## 1  ( 1 )  "*"              " "              " "      " "      " "
## 2  ( 1 )  "*"              " "              " "      " "      " "
## 3  ( 1 )  "*"              " "              " "      " "      " "
## 4  ( 1 )  "*"              " "              "*"      " "      " "
## 5  ( 1 )  "*"              " "              "*"      "*"      " "
## 6  ( 1 )  "*"              " "              "*"      "*"      " "
## 7  ( 1 )  "*"              " "              "*"      "*"      "*"
## 8  ( 1 )  "*"              " "              "*"      "*"      "*"
## 9  ( 1 )  "*"              " "              "*"      "*"      "*"
## 10  ( 1 ) "*"              " "              "*"      "*"      "*"
## 11  ( 1 ) "*"              " "              "*"      "*"      "*"
## 12  ( 1 ) "*"              " "              "*"      "*"      "*"
## 13  ( 1 ) "*"              "*"              "*"      "*"      "*"
## 14  ( 1 ) "*"              "*"              "*"      "*"      "*"
## 15  ( 1 ) "*"              "*"              "*"      "*"      "*"
##           LDAPS_CC2 LDAPS_CC3 LDAPS_CC4 LDAPS_PPT1 LDAPS_PPT2 LDAPS_PPT3
## 1  ( 1 )  " "       " "       " "       " "        " "        " "
## 2  ( 1 )  " "       " "       " "       " "        " "        " "
## 3  ( 1 )  " "       " "       "*"       " "        " "        " "
## 4  ( 1 )  " "       " "       "*"       " "        " "        " "
## 5  ( 1 )  " "       " "       "*"       " "        " "        " "
## 6  ( 1 )  " "       " "       "*"       " "        " "        " "
## 7  ( 1 )  " "       " "       "*"       " "        " "        " "
## 8  ( 1 )  " "       " "       "*"       " "        " "        " "
## 9  ( 1 )  " "       " "       "*"       " "        "*"        " "
## 10  ( 1 ) " "       " "       "*"       " "        "*"        " "
## 11  ( 1 ) " "       "*"       "*"       " "        "*"        " "
## 12  ( 1 ) " "       "*"       "*"       " "        "*"        " "
## 13  ( 1 ) " "       "*"       "*"       " "        "*"        " "
## 14  ( 1 ) "*"       "*"       "*"       " "        "*"        " "
## 15  ( 1 ) "*"       "*"       "*"       " "        "*"        " "
##           LDAPS_PPT4 lat lon DEM Slope Solar.radiation
## 1  ( 1 )  " "        " " " " " " " "   " "
## 2  ( 1 )  " "        " " " " " " " "   " "
## 3  ( 1 )  " "        " " " " " " " "   " "
## 4  ( 1 )  " "        " " " " " " " "   " "
## 5  ( 1 )  " "        " " " " " " " "   " "
## 6  ( 1 )  " "        " " " " " " " "   " "
## 7  ( 1 )  " "        " " " " " " " "   " "
## 8  ( 1 )  " "        " " " " " " " "   " "
## 9  ( 1 )  " "        " " " " " " " "   " "
## 10  ( 1 ) " "        " " "*" " " " "   " "
## 11  ( 1 ) " "        " " "*" " " " "   " "
## 12  ( 1 ) " "        " " "*" " " " "   " "
## 13  ( 1 ) " "        " " "*" " " " "   " "
```

```
## 14  ( 1 ) " "         " " "*" " " " " "    " "
## 15  ( 1 ) " "         " " "*" " " "*"    " "
```




```
## [1] 15

## [1] "the best model is model number 15 with an R square= 0.780746271125755
"
```



In this case, the model 15. The variables contained in this model are: Station, Date, Present Max Temperature, Next day's humidity, next-day maximum air temperature, next-day minimum air temperature, next day's wind speed,next-day average latent heat flux, split average cloud cover (all four), precipitations for the 2nd 6 hours, longitude and slope. The coefficients for these variable are shown in the table above. The variables contained in the two models are pretty similar with only 2 variables different between the two.

Let's then fit the two models and perform an ANOVA test.

```
## Next_Tmax ~ station + Date + Present_Tmax + LDAPS_RHmin + LDAPS_Tmax_lapse
+
##     LDAPS_Tmin_lapse + LDAPS_WS + LDAPS_LH + LDAPS_CC1 + LDAPS_CC3 +
##     LDAPS_CC4 + LDAPS_PPT2 + lon + DEM + Slope
## <environment: 0x0000000021dccba0>

## Next_Tmax ~ station + Date + Present_Tmax + LDAPS_RHmin + LDAPS_Tmax_lapse
+
##     LDAPS_Tmin_lapse + LDAPS_WS + LDAPS_LH + LDAPS_CC1 + LDAPS_CC2 +
##     LDAPS_CC3 + LDAPS_CC4 + LDAPS_PPT2 + lon + Slope
## <environment: 0x0000000022541ed0>

## Analysis of Variance Table
##
## Response: data$Next_Tmax
##                          Df  Sum Sq Mean Sq    F value    Pr(>F)
## data$station              1   852.9   852.9   403.1112 < 2.2e-16 ***
## data$Date                 1   281.6   281.6   133.0911 < 2.2e-16 ***
## data$Present_Tmax         1 26398.7 26398.7 12476.6527 < 2.2e-16 ***
## data$LDAPS_RHmin          1  7764.0  7764.0  3669.4649 < 2.2e-16 ***
## data$LDAPS_Tmax_lapse     1 18084.7 18084.7  8547.2558 < 2.2e-16 ***
## data$LDAPS_Tmin_lapse     1    24.3    24.3    11.5022 0.0006987 ***
## data$LDAPS_WS             1   787.5   787.5   372.1718 < 2.2e-16 ***
## data$LDAPS_LH             1   865.7   865.7   409.1701 < 2.2e-16 ***
## data$LDAPS_CC1            1   632.5   632.5   298.9204 < 2.2e-16 ***
## data$LDAPS_CC3            1  1007.9  1007.9   476.3391 < 2.2e-16 ***
## data$LDAPS_CC4            1   200.8   200.8    94.9262 < 2.2e-16 ***
## data$LDAPS_PPT2           1   245.3   245.3   115.9315 < 2.2e-16 ***
## data$lon                  1   137.6   137.6    65.0115 8.591e-16 ***
## data$DEM                  1     6.2     6.2     2.9419 0.0863525 .
## data$Slope                1   156.7   156.7    74.0617 < 2.2e-16 ***
## Residuals              7572 16021.2     2.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Analysis of Variance Table
##
## Response: data$Next_Tmax
##                          Df  Sum Sq Mean Sq   F value    Pr(>F)
## data$station              1   852.9   852.9   400.938 < 2.2e-16 ***
## data$Date                 1   281.6   281.6   132.374 < 2.2e-16 ***
```

```
## data$Present_Tmax        1 26398.7 26398.7 12409.379 < 2.2e-16 ***
## data$LDAPS_RHmin         1  7764.0  7764.0  3649.679 < 2.2e-16 ***
## data$LDAPS_Tmax_lapse    1 18084.7 18084.7  8501.170 < 2.2e-16 ***
## data$LDAPS_Tmin_lapse    1    24.3    24.3    11.440 0.0007224 ***
## data$LDAPS_WS            1   787.5   787.5   370.165 < 2.2e-16 ***
## data$LDAPS_LH            1   865.7   865.7   406.964 < 2.2e-16 ***
## data$LDAPS_CC1           1   632.5   632.5   297.309 < 2.2e-16 ***
## data$LDAPS_CC2           1   182.8   182.8    85.912 < 2.2e-16 ***
## data$LDAPS_CC3           1   833.0   833.0   391.551 < 2.2e-16 ***
## data$LDAPS_CC4           1   204.1   204.1    95.962 < 2.2e-16 ***
## data$LDAPS_PPT2          1   288.1   288.1   135.441 < 2.2e-16 ***
## data$lon                 1   128.5   128.5    60.397 8.775e-15 ***
## data$Slope               1    31.2    31.2    14.650 0.0001305 ***
## Residuals             7572 16108.0     2.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When applying anova on each of the two models, we find that for the first model the variable “Dem” which stands for elevation has a high a p value. This variable is then discarded.We build a new model without that variable, perform anova for the new model and finally perform a second anova for the new model and the second model retained from the earlier trial.

```
## Analysis of Variance Table
##
## Response: data$Next_Tmax
##                          Df  Sum Sq Mean Sq   F value    Pr(>F)
## data$station              1   852.9   852.9   399.970 < 2.2e-16 ***
## data$Date                 1   281.6   281.6   132.054 < 2.2e-16 ***
## data$Present_Tmax         1 26398.7 26398.7 12379.441 < 2.2e-16 ***
## data$LDAPS_RHmin          1  7764.0  7764.0  3640.874 < 2.2e-16 ***
## data$LDAPS_Tmax_lapse     1 18084.7 18084.7  8480.660 < 2.2e-16 ***
## data$LDAPS_Tmin_lapse     1    24.3    24.3    11.412 0.0007332 ***
## data$LDAPS_WS             1   787.5   787.5   369.272 < 2.2e-16 ***
## data$LDAPS_LH             1   865.7   865.7   405.982 < 2.2e-16 ***
## data$LDAPS_CC1            1   632.5   632.5   296.591 < 2.2e-16 ***
## data$LDAPS_CC3            1  1007.9  1007.9   472.628 < 2.2e-16 ***
## data$LDAPS_CC4            1   200.8   200.8    94.187 < 2.2e-16 ***
## data$LDAPS_PPT2           1   245.3   245.3   115.028 < 2.2e-16 ***
## data$lon                  1   137.6   137.6    64.505 1.108e-15 ***
## data$Slope                1    35.0    35.0    16.407 5.162e-05 ***
## Residuals              7573 16149.1     2.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Analysis of Variance Table
##
## Model 1: data$Next_Tmax ~ data$station + data$Date + data$Present_Tmax +
##     data$LDAPS_RHmin + data$LDAPS_Tmax_lapse + data$LDAPS_Tmin_lapse +
##     data$LDAPS_WS + data$LDAPS_LH + data$LDAPS_CC1 + data$LDAPS_CC3 +
```

```
##     data$LDAPS_CC4 + data$LDAPS_PPT2 + data$lon + data$Slope
## Model 2: data$Next_Tmax ~ data$station + data$Date + data$Present_Tmax +
##     data$LDAPS_RHmin + data$LDAPS_Tmax_lapse + data$LDAPS_Tmin_lapse +
##     data$LDAPS_WS + data$LDAPS_LH + data$LDAPS_CC1 + data$LDAPS_CC2 +
##     data$LDAPS_CC3 + data$LDAPS_CC4 + data$LDAPS_PPT2 + data$lon +
##     data$Slope
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1   7573 16149
## 2   7572 16108  1    41.088 19.314 1.124e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observing the pvalue in the anova above, we can conclude that there is a significant difference between the two models.

Let's now see the summary the two models:

```
## [1] "Model 1"

##
## Call:
## lm(formula = data$Next_Tmax ~ data$station + data$Date + data$Present_Tmax
+
##     data$LDAPS_RHmin + data$LDAPS_Tmax_lapse + data$LDAPS_Tmin_lapse +
##     data$LDAPS_WS + data$LDAPS_LH + data$LDAPS_CC1 + data$LDAPS_CC3 +
##     data$LDAPS_CC4 + data$LDAPS_PPT2 + data$lon + data$Slope)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3832 -0.8445  0.0522  0.8989  6.1753
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.301e+02  2.786e+01   8.257  < 2e-16 ***
## data$station           2.494e-02  2.400e-03  10.390  < 2e-16 ***
## data$Date              2.327e-04  3.361e-05   6.925 4.73e-12 ***
## data$Present_Tmax      1.616e-01  8.771e-03  18.427  < 2e-16 ***
## data$LDAPS_RHmin       2.469e-02  2.490e-03   9.919  < 2e-16 ***
## data$LDAPS_Tmax_lapse  6.428e-01  1.464e-02  43.896  < 2e-16 ***
## data$LDAPS_Tmin_lapse  9.601e-02  1.743e-02   5.509 3.73e-08 ***
## data$LDAPS_WS         -1.472e-01  8.484e-03 -17.356  < 2e-16 ***
## data$LDAPS_LH          7.592e-03  5.764e-04  13.170  < 2e-16 ***
## data$LDAPS_CC1        -1.347e+00  9.114e-02 -14.783  < 2e-16 ***
## data$LDAPS_CC3        -1.170e+00  1.449e-01  -8.075 7.81e-16 ***
## data$LDAPS_CC4        -1.167e+00  1.108e-01 -10.531  < 2e-16 ***
## data$LDAPS_PPT2        1.198e-01  1.080e-02  11.095  < 2e-16 ***
## data$lon              -1.810e+00  2.194e-01  -8.250  < 2e-16 ***
## data$Slope             5.279e-02  1.303e-02   4.051 5.16e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.46 on 7573 degrees of freedom
##   (164 observations deleted due to missingness)
## Multiple R-squared:  0.7802, Adjusted R-squared:  0.7798
## F-statistic:  1920 on 14 and 7573 DF,  p-value: < 2.2e-16

## [1] "Model 2"

##
## Call:
## lm(formula = data$Next_Tmax ~ data$station + data$Date + data$Present_Tmax
+
##     data$LDAPS_RHmin + data$LDAPS_Tmax_lapse + data$LDAPS_Tmin_lapse +
##     data$LDAPS_WS + data$LDAPS_LH + data$LDAPS_CC1 + data$LDAPS_CC2 +
##     data$LDAPS_CC3 + data$LDAPS_CC4 + data$LDAPS_PPT2 + data$lon +
##     data$Slope)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4525 -0.8397  0.0463  0.8840  6.1809
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.231e+02  2.788e+01   8.004 1.38e-15 ***
## data$station           2.488e-02  2.398e-03  10.377  < 2e-16 ***
## data$Date              2.221e-04  3.366e-05   6.600 4.39e-11 ***
## data$Present_Tmax      1.642e-01  8.780e-03  18.704  < 2e-16 ***
## data$LDAPS_RHmin       2.655e-02  2.522e-03  10.525  < 2e-16 ***
## data$LDAPS_Tmax_lapse  6.315e-01  1.485e-02  42.532  < 2e-16 ***
## data$LDAPS_Tmin_lapse  1.044e-01  1.751e-02   5.964 2.57e-09 ***
## data$LDAPS_WS         -1.494e-01  8.487e-03 -17.599  < 2e-16 ***
## data$LDAPS_LH          7.106e-03  5.862e-04  12.121  < 2e-16 ***
## data$LDAPS_CC1        -1.078e+00  1.097e-01  -9.824  < 2e-16 ***
## data$LDAPS_CC2        -6.493e-01  1.477e-01  -4.395 1.12e-05 ***
## data$LDAPS_CC3        -9.849e-01  1.508e-01  -6.533 6.89e-11 ***
## data$LDAPS_CC4        -1.193e+00  1.108e-01 -10.769  < 2e-16 ***
## data$LDAPS_PPT2        1.315e-01  1.111e-02  11.837  < 2e-16 ***
## data$lon              -1.753e+00  2.195e-01  -7.987 1.58e-15 ***
## data$Slope             4.989e-02  1.303e-02   3.828  0.00013 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.459 on 7572 degrees of freedom
##   (164 observations deleted due to missingness)
## Multiple R-squared:  0.7807, Adjusted R-squared:  0.7803
## F-statistic:  1798 on 15 and 7572 DF,  p-value: < 2.2e-16
```

Observing the summaries of the two models, we can see that the second model present a slightly better R square and adjusted R square, meaning Model 2 is the better one.

For Min temperature

We start by using the function "regsubsets" on the dataset after excluding the next day's Max temperature. We start with the default settings that give us 8 estimated models:

```
## Subset selection object
## Call: regsubsets.formula(data$Next_Tmax ~ ., data = data[, -24])
## 24 Variables  (and intercept)
##                  Forced in Forced out
## station              FALSE      FALSE
## Date                 FALSE      FALSE
## Present_Tmax         FALSE      FALSE
## Present_Tmin         FALSE      FALSE
## LDAPS_RHmin          FALSE      FALSE
## LDAPS_RHmax          FALSE      FALSE
## LDAPS_Tmax_lapse     FALSE      FALSE
## LDAPS_Tmin_lapse     FALSE      FALSE
## LDAPS_WS             FALSE      FALSE
## LDAPS_LH             FALSE      FALSE
## LDAPS_CC1            FALSE      FALSE
## LDAPS_CC2            FALSE      FALSE
## LDAPS_CC3            FALSE      FALSE
## LDAPS_CC4            FALSE      FALSE
## LDAPS_PPT1           FALSE      FALSE
## LDAPS_PPT2           FALSE      FALSE
## LDAPS_PPT3           FALSE      FALSE
## LDAPS_PPT4           FALSE      FALSE
## lat                  FALSE      FALSE
## lon                  FALSE      FALSE
## DEM                  FALSE      FALSE
## Slope                FALSE      FALSE
## Solar.radiation      FALSE      FALSE
## Next_Tmin            FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          station Date Present_Tmax Present_Tmin LDAPS_RHmin LDAPS_RHmax
## 1  ( 1 ) " "     " "  " "          " "          " "          " "
## 2  ( 1 ) " "     " "  "*"          " "          " "          " "
## 3  ( 1 ) " "     " "  " "          " "          " "          " "
## 4  ( 1 ) " "     " "  " "          " "          " "          " "
## 5  ( 1 ) " "     " "  " "          " "          " "          " "
## 6  ( 1 ) " "     " "  " "          " "          " "          " "
## 7  ( 1 ) " "     " "  " "          " "          " "          " "
## 8  ( 1 ) " "     " "  "*"          " "          " "          " "
##          LDAPS_Tmax_lapse LDAPS_Tmin_lapse LDAPS_WS LDAPS_LH LDAPS_CC1
## 1  ( 1 ) "*"              " "              " "      " "      " "
## 2  ( 1 ) "*"              " "              " "      " "      " "
## 3  ( 1 ) "*"              " "              " "      " "      " "
## 4  ( 1 ) "*"              " "              " "      " "      "*"
## 5  ( 1 ) "*"              " "              " "      "*"      "*"
```

```
## 6  ( 1 ) "*"              " "              "*"      "*"      "*"
## 7  ( 1 ) "*"              " "              "*"      "*"      "*"
## 8  ( 1 ) "*"              " "              "*"      "*"      "*"
##          LDAPS_CC2 LDAPS_CC3 LDAPS_CC4 LDAPS_PPT1 LDAPS_PPT2 LDAPS_PPT3
## 1  ( 1 ) " "       " "       " "       " "        " "        " "
## 2  ( 1 ) " "       " "       " "       " "        " "        " "
## 3  ( 1 ) "*"       " "       " "       " "        " "        " "
## 4  ( 1 ) " "       " "       "*"       " "        " "        " "
## 5  ( 1 ) " "       " "       "*"       " "        " "        " "
## 6  ( 1 ) " "       " "       "*"       " "        " "        " "
## 7  ( 1 ) " "       " "       "*"       " "        "*"        " "
## 8  ( 1 ) " "       " "       "*"       " "        "*"        " "
##          LDAPS_PPT4 lat lon DEM Slope Solar.radiation Next_Tmin
## 1  ( 1 ) " "        " " " " " " " " "   " "              " "
## 2  ( 1 ) " "        " " " " " " " " "   " "              " "
## 3  ( 1 ) " "        " " " " " " " " "   " "              "*"
## 4  ( 1 ) " "        " " " " " " " " "   " "              "*"
## 5  ( 1 ) " "        " " " " " " " " "   " "              "*"
## 6  ( 1 ) " "        " " " " " " " " "   " "              "*"
## 7  ( 1 ) " "        " " " " " " " " "   " "              "*"
## 8  ( 1 ) " "        " " " " " " " " "   " "              "*"
```

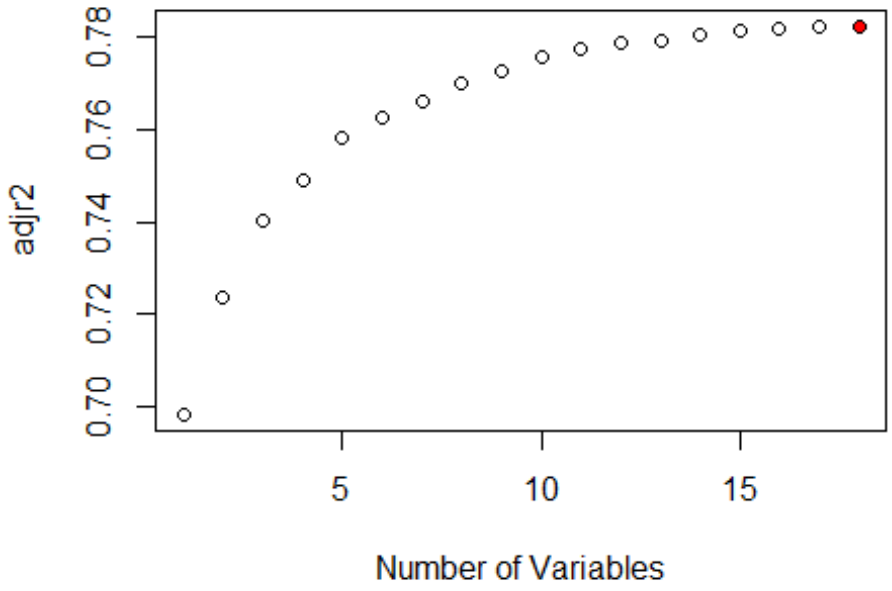
We can observe above the results of this first trial where for each model we can see the variables (that have * in front of them) that are included in each of the 8 models.

Given the relatively large number of variables we can also seek to build more models here too. Let's build 18 models for example.

```
## Subset selection object
## Call: regsubsets.formula(data$Next_Tmax ~ ., data = data[, -c(24, 25)],
##     nvmax = 18)
## 23 Variables  (and intercept)
##                  Forced in Forced out
## station              FALSE      FALSE
## Date                 FALSE      FALSE
## Present_Tmax         FALSE      FALSE
## Present_Tmin         FALSE      FALSE
## LDAPS_RHmin          FALSE      FALSE
## LDAPS_RHmax          FALSE      FALSE
## LDAPS_Tmax_lapse     FALSE      FALSE
## LDAPS_Tmin_lapse     FALSE      FALSE
## LDAPS_WS             FALSE      FALSE
## LDAPS_LH             FALSE      FALSE
## LDAPS_CC1            FALSE      FALSE
## LDAPS_CC2            FALSE      FALSE
## LDAPS_CC3            FALSE      FALSE
## LDAPS_CC4            FALSE      FALSE
## LDAPS_PPT1           FALSE      FALSE
## LDAPS_PPT2           FALSE      FALSE
## LDAPS_PPT3           FALSE      FALSE
```

```
## LDAPS_PPT4           FALSE      FALSE
## lat                  FALSE      FALSE
## lon                  FALSE      FALSE
## DEM                  FALSE      FALSE
## Slope                FALSE      FALSE
## Solar.radiation      FALSE      FALSE
## 1 subsets of each size up to 18
## Selection Algorithm: exhaustive
##           station Date Present_Tmax Present_Tmin LDAPS_RHmin LDAPS_RHmax
## 1  ( 1 )  " "     " "  " "          " "          " "          " "
## 2  ( 1 )  " "     " "  "*"          " "          " "          " "
## 3  ( 1 )  " "     " "  "*"          " "          " "          " "
## 4  ( 1 )  " "     " "  "*"          " "          " "          " "
## 5  ( 1 )  " "     " "  "*"          " "          "*"          " "
## 6  ( 1 )  " "     " "  "*"          " "          "*"          " "
## 7  ( 1 )  " "     " "  "*"          " "          "*"          " "
## 8  ( 1 )  "*"     " "  "*"          " "          "*"          " "
## 9  ( 1 )  "*"     " "  "*"          " "          "*"          " "
## 10  ( 1 ) "*"     " "  "*"          " "          "*"          " "
## 11  ( 1 ) "*"     " "  "*"          " "          "*"          " "
## 12  ( 1 ) "*"     "*"  "*"          " "          "*"          " "
## 13  ( 1 ) "*"     "*"  "*"          " "          "*"          " "
## 14  ( 1 ) "*"     "*"  "*"          " "          "*"          " "
## 15  ( 1 ) "*"     "*"  "*"          " "          "*"          " "
## 16  ( 1 ) "*"     "*"  "*"          " "          "*"          " "
## 17  ( 1 ) "*"     "*"  "*"          " "          "*"          " "
## 18  ( 1 ) "*"     "*"  "*"          " "          "*"          " "
##           LDAPS_Tmax_lapse LDAPS_Tmin_lapse LDAPS_WS LDAPS_LH LDAPS_CC1
## 1  ( 1 )  "*"              " "              " "       " "       " "
## 2  ( 1 )  "*"              " "              " "       " "       " "
## 3  ( 1 )  "*"              " "              " "       " "       " "
## 4  ( 1 )  "*"              " "              "*"       " "       " "
## 5  ( 1 )  "*"              " "              "*"       " "       " "
## 6  ( 1 )  "*"              " "              "*"       " "       "*"
## 7  ( 1 )  "*"              " "              "*"       "*"       "*"
## 8  ( 1 )  "*"              " "              "*"       "*"       "*"
## 9  ( 1 )  "*"              " "              "*"       "*"       "*"
## 10  ( 1 ) "*"              " "              "*"       "*"       "*"
## 11  ( 1 ) "*"              " "              "*"       "*"       "*"
## 12  ( 1 ) "*"              " "              "*"       "*"       "*"
## 13  ( 1 ) "*"              "*"              "*"       "*"       "*"
## 14  ( 1 ) "*"              " "              "*"       "*"       "*"
## 15  ( 1 ) "*"              "*"              "*"       "*"       "*"
## 16  ( 1 ) "*"              "*"              "*"       "*"       "*"
## 17  ( 1 ) "*"              "*"              "*"       "*"       "*"
## 18  ( 1 ) "*"              "*"              "*"       "*"       "*"
##           LDAPS_CC2 LDAPS_CC3 LDAPS_CC4 LDAPS_PPT1 LDAPS_PPT2 LDAPS_PPT3
## 1  ( 1 )  " "       " "       " "       " "        " "        " "
## 2  ( 1 )  " "       " "       " "       " "        " "        " "
## 3  ( 1 )  " "       " "       "*"       " "        " "        " "
```

```
## 4  ( 1 )  " "       " "       "*"       " "       " "       " "
## 5  ( 1 )  " "       "*"       " "       " "       " "       " "
## 6  ( 1 )  " "       "*"       " "       " "       " "       " "
## 7  ( 1 )  " "       " "       "*"       " "       " "       " "
## 8  ( 1 )  " "       " "       "*"       " "       " "       " "
## 9  ( 1 )  " "       " "       "*"       " "       "*"       " "
## 10  ( 1 ) " "       " "       "*"       " "       "*"       " "
## 11  ( 1 ) " "       "*"       "*"       " "       "*"       " "
## 12  ( 1 ) " "       "*"       "*"       " "       "*"       " "
## 13  ( 1 ) " "       "*"       "*"       " "       "*"       " "
## 14  ( 1 ) " "       "*"       "*"       " "       "*"       " "
## 15  ( 1 ) " "       "*"       "*"       " "       "*"       " "
## 16  ( 1 ) "*"       "*"       "*"       " "       "*"       " "
## 17  ( 1 ) "*"       "*"       "*"       " "       "*"       " "
## 18  ( 1 ) "*"       "*"       "*"       " "       "*"       "*"
##           LDAPS_PPT4 lat lon DEM Slope Solar.radiation
## 1  ( 1 )  " "        " " " " " " " " "   " "
## 2  ( 1 )  " "        " " " " " " " " "   " "
## 3  ( 1 )  " "        " " " " " " " " "   " "
## 4  ( 1 )  " "        " " " " " " " " "   " "
## 5  ( 1 )  " "        " " " " " " " " "   " "
## 6  ( 1 )  " "        " " " " " " " " "   " "
## 7  ( 1 )  " "        " " " " " " " " "   " "
## 8  ( 1 )  " "        " " " " " " " " "   " "
## 9  ( 1 )  " "        " " " " " " " " "   " "
## 10  ( 1 ) " "        " " "*" " " " " "   " "
## 11  ( 1 ) " "        " " "*" " " " " "   " "
## 12  ( 1 ) " "        " " "*" " " " " "   " "
## 13  ( 1 ) " "        " " "*" " " " " "   " "
## 14  ( 1 ) " "        " " "*" "*" "*"   " "
## 15  ( 1 ) " "        " " "*" "*" "*"   " "
## 16  ( 1 ) " "        " " "*" "*" "*"   " "
## 17  ( 1 ) " "        "*" "*" "*" "*"   " "
## 18  ( 1 ) " "        "*" "*" "*" "*"   " "

## [1] 18

##      (Intercept)          station             Date     Present_Tmax
##     2.508116e+02     1.774941e-02     2.308506e-04     1.517126e-01
##      LDAPS_RHmin LDAPS_Tmax_lapse LDAPS_Tmin_lapse         LDAPS_WS
##     2.606578e-02     6.302143e-01     1.107050e-01    -1.394680e-01
##         LDAPS_LH        LDAPS_CC1        LDAPS_CC2        LDAPS_CC3
##     7.270654e-03    -1.146811e+00    -6.490823e-01    -8.961988e-01
##        LDAPS_CC4       LDAPS_PPT2       LDAPS_PPT3              lat
##    -1.220851e+00     1.344634e-01    -3.751698e-02    -8.859339e-01
##             lon              DEM            Slope
##    -1.707347e+00    -4.312373e-03     1.790148e-01

## [1] "the best model is model number 18 with an adjusted R square= 0.782306
351455501"
```



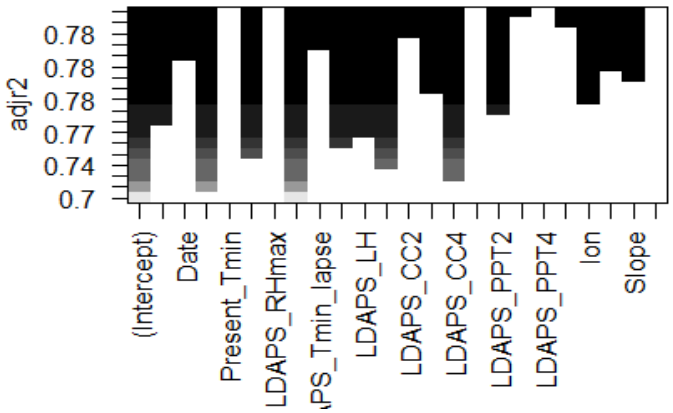
After building the 18 models, we can check the model with the highest Adjusted R square (in other words the best model). We find that model number 18 is the best one. The variables contained in this model are: Station, Date, Present Min Temperature, Next day's humidity, next-day maximum air temperature, next-day minimum air temperature, next day's wind speed,next-day average latent heat flux, split average cloud cover (all 4)), precipitations for the 2nd and 3rd 6 hours, longitude,latitude, elevation and slope. The coefficients for these variable are shown in the table above.

Let's now try this again with "backward" method:

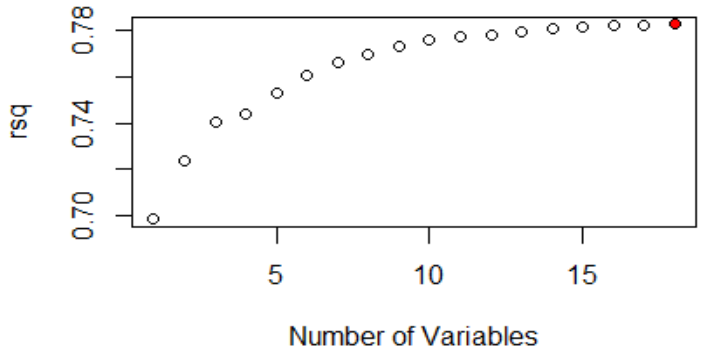
```
## Subset selection object
## Call: regsubsets.formula(Next_Tmax ~ ., data = data[, -25], nvmax = 18,
##     method = "backward")
## 23 Variables  (and intercept)
##                  Forced in Forced out
## station              FALSE      FALSE
## Date                 FALSE      FALSE
## Present_Tmax         FALSE      FALSE
## Present_Tmin         FALSE      FALSE
## LDAPS_RHmin          FALSE      FALSE
## LDAPS_RHmax          FALSE      FALSE
## LDAPS_Tmax_lapse     FALSE      FALSE
## LDAPS_Tmin_lapse     FALSE      FALSE
## LDAPS_WS             FALSE      FALSE
## LDAPS_LH             FALSE      FALSE
## LDAPS_CC1            FALSE      FALSE
```

```
## LDAPS_CC2            FALSE      FALSE
## LDAPS_CC3            FALSE      FALSE
## LDAPS_CC4            FALSE      FALSE
## LDAPS_PPT1           FALSE      FALSE
## LDAPS_PPT2           FALSE      FALSE
## LDAPS_PPT3           FALSE      FALSE
## LDAPS_PPT4           FALSE      FALSE
## lat                  FALSE      FALSE
## lon                  FALSE      FALSE
## DEM                  FALSE      FALSE
## Slope                FALSE      FALSE
## Solar.radiation      FALSE      FALSE
## 1 subsets of each size up to 18
## Selection Algorithm: backward
##           station Date Present_Tmax Present_Tmin LDAPS_RHmin LDAPS_RHmax
## 1  ( 1 )  " "     " "  " "          " "          " "         " "
## 2  ( 1 )  " "     " "  "*"          " "          " "         " "
## 3  ( 1 )  " "     " "  "*"          " "          " "         " "
## 4  ( 1 )  " "     " "  "*"          " "          " "         " "
## 5  ( 1 )  " "     " "  "*"          " "          "*"         " "
## 6  ( 1 )  " "     " "  "*"          " "          "*"         " "
## 7  ( 1 )  " "     " "  "*"          " "          "*"         " "
## 8  ( 1 )  "*"     " "  "*"          " "          "*"         " "
## 9  ( 1 )  "*"     " "  "*"          " "          "*"         " "
## 10  ( 1 ) "*"     " "  "*"          " "          "*"         " "
## 11  ( 1 ) "*"     " "  "*"          " "          "*"         " "
## 12  ( 1 ) "*"     " "  "*"          " "          "*"         " "
## 13  ( 1 ) "*"     " "  "*"          " "          "*"         " "
## 14  ( 1 ) "*"     "*"  "*"          " "          "*"         " "
## 15  ( 1 ) "*"     "*"  "*"          " "          "*"         " "
## 16  ( 1 ) "*"     "*"  "*"          " "          "*"         " "
## 17  ( 1 ) "*"     "*"  "*"          " "          "*"         " "
## 18  ( 1 ) "*"     "*"  "*"          " "          "*"         " "
##           LDAPS_Tmax_lapse LDAPS_Tmin_lapse LDAPS_WS LDAPS_LH LDAPS_CC1
## 1  ( 1 )  "*"              " "              " "      " "      " "
## 2  ( 1 )  "*"              " "              " "      " "      " "
## 3  ( 1 )  "*"              " "              " "      " "      " "
## 4  ( 1 )  "*"              " "              " "      " "      "*"
## 5  ( 1 )  "*"              " "              " "      " "      "*"
## 6  ( 1 )  "*"              " "              "*"      " "      "*"
## 7  ( 1 )  "*"              " "              "*"      "*"      "*"
## 8  ( 1 )  "*"              " "              "*"      "*"      "*"
## 9  ( 1 )  "*"              " "              "*"      "*"      "*"
## 10  ( 1 ) "*"              " "              "*"      "*"      "*"
## 11  ( 1 ) "*"              " "              "*"      "*"      "*"
## 12  ( 1 ) "*"              " "              "*"      "*"      "*"
## 13  ( 1 ) "*"              " "              "*"      "*"      "*"
## 14  ( 1 ) "*"              " "              "*"      "*"      "*"
## 15  ( 1 ) "*"              "*"              "*"      "*"      "*"
## 16  ( 1 ) "*"              "*"              "*"      "*"      "*"
```

```
## 17  ( 1 ) "*"             "*"             "*"      "*"      "*"
## 18  ( 1 ) "*"             "*"             "*"      "*"      "*"
##           LDAPS_CC2 LDAPS_CC3 LDAPS_CC4 LDAPS_PPT1 LDAPS_PPT2 LDAPS_PPT3
## 1  ( 1 )  " "       " "       " "       " "        " "        " "
## 2  ( 1 )  " "       " "       " "       " "        " "        " "
## 3  ( 1 )  " "       " "       "*"       " "        " "        " "
## 4  ( 1 )  " "       " "       "*"       " "        " "        " "
## 5  ( 1 )  " "       " "       "*"       " "        " "        " "
## 6  ( 1 )  " "       " "       "*"       " "        " "        " "
## 7  ( 1 )  " "       " "       "*"       " "        " "        " "
## 8  ( 1 )  " "       " "       "*"       " "        " "        " "
## 9  ( 1 )  " "       " "       "*"       " "        "*"        " "
## 10  ( 1 ) " "       " "       "*"       " "        "*"        " "
## 11  ( 1 ) " "       "*"       "*"       " "        "*"        " "
## 12  ( 1 ) " "       "*"       "*"       " "        "*"        " "
## 13  ( 1 ) " "       "*"       "*"       " "        "*"        " "
## 14  ( 1 ) " "       "*"       "*"       " "        "*"        " "
## 15  ( 1 ) " "       "*"       "*"       " "        "*"        " "
## 16  ( 1 ) "*"       "*"       "*"       " "        "*"        " "
## 17  ( 1 ) "*"       "*"       "*"       " "        "*"        " "
## 18  ( 1 ) "*"       "*"       "*"       " "        "*"        "*"
##           LDAPS_PPT4 lat lon DEM Slope Solar.radiation
## 1  ( 1 )  " "        " " " " " " " "   " "
## 2  ( 1 )  " "        " " " " " " " "   " "
## 3  ( 1 )  " "        " " " " " " " "   " "
## 4  ( 1 )  " "        " " " " " " " "   " "
## 5  ( 1 )  " "        " " " " " " " "   " "
## 6  ( 1 )  " "        " " " " " " " "   " "
## 7  ( 1 )  " "        " " " " " " " "   " "
## 8  ( 1 )  " "        " " " " " " " "   " "
## 9  ( 1 )  " "        " " " " " " " "   " "
## 10  ( 1 ) " "        " " "*" " " " "   " "
## 11  ( 1 ) " "        " " "*" " " " "   " "
## 12  ( 1 ) " "        " " "*" " " "*"   " "
## 13  ( 1 ) " "        " " "*" "*" "*"   " "
## 14  ( 1 ) " "        " " "*" "*" "*"   " "
## 15  ( 1 ) " "        " " "*" "*" "*"   " "
## 16  ( 1 ) " "        " " "*" "*" "*"   " "
## 17  ( 1 ) " "        "*" "*" "*" "*"   " "
## 18  ( 1 ) " "        "*" "*" "*" "*"   " "
```




```
## [1] 18

## [1] "the best model is model number 18 with an R square= 0.782822825117528
"
```




```
##      (Intercept)          station             Date     Present_Tmax
##     2.508116e+02     1.774941e-02     2.308506e-04     1.517126e-01
##      LDAPS_RHmin LDAPS_Tmax_lapse LDAPS_Tmin_lapse         LDAPS_WS
##     2.606578e-02     6.302143e-01     1.107050e-01    -1.394680e-01
##         LDAPS_LH        LDAPS_CC1        LDAPS_CC2        LDAPS_CC3
##     7.270654e-03    -1.146811e+00    -6.490823e-01    -8.961988e-01
##        LDAPS_CC4       LDAPS_PPT2       LDAPS_PPT3              lat
##    -1.220851e+00     1.344634e-01    -3.751698e-02    -8.859339e-01
```

```
##               lon               DEM             Slope
##    -1.707347e+00    -4.312373e-03     1.790148e-01
```

In this case, the model 18. The variables contained in this model are: Station, Date, Present Max Temperature, Next day's humidity, next-day maximum air temperature, next-day minimum air temperature, next day's wind speed,next-day average latent heat flux, split average cloud cover (all four), precipitations for the 2nd 6 hours, longitude, latitude, elevation and slope. The coefficients for these variable are shown in the table above. The variables contained in the two models are pretty similar with only 2 variables different between the two.

Let's then fit the two models and perform an ANOVA test.

```
## Next_Tmin ~ station + Date + Present_Tmax + LDAPS_RHmin + LDAPS_Tmax_lapse
+
##     LDAPS_Tmin_lapse + LDAPS_WS + LDAPS_LH + LDAPS_CC1 + LDAPS_CC2 +
##     LDAPS_CC3 + LDAPS_CC4 + LDAPS_PPT2 + LDAPS_PPT3 + lat + lon +
##     DEM + Slope
## <environment: 0x00000000491c0f18>

## Next_Tmin ~ station + Date + Present_Tmax + LDAPS_RHmin + LDAPS_Tmax_lapse
+
##     LDAPS_Tmin_lapse + LDAPS_WS + LDAPS_LH + LDAPS_CC1 + LDAPS_CC2 +
##     LDAPS_CC3 + LDAPS_CC4 + LDAPS_PPT2 + LDAPS_PPT3 + lat + lon +
##     DEM + Slope
## <environment: 0x00000000490b8f20>

## Analysis of Variance Table
##
## Response: data$Next_Tmin
##                          Df  Sum Sq Mean Sq    F value    Pr(>F)
## data$station              1   768.1   768.1   685.8293 < 2.2e-16 ***
## data$Date                 1    19.0    19.0    16.9199 3.940e-05 ***
## data$Present_Tmax         1 17413.2 17413.2 15547.5760 < 2.2e-16 ***
## data$LDAPS_RHmin          1  2495.7  2495.7  2228.3013 < 2.2e-16 ***
## data$LDAPS_Tmax_lapse     1 11734.1 11734.1 10476.9817 < 2.2e-16 ***
## data$LDAPS_Tmin_lapse     1  4940.8  4940.8  4411.4592 < 2.2e-16 ***
## data$LDAPS_WS             1     8.0     8.0     7.1249 0.0076184 **
## data$LDAPS_LH             1    49.8    49.8    44.4447 2.800e-11 ***
## data$LDAPS_CC1            1     2.7     2.7     2.3900 0.1221585
## data$LDAPS_CC2            1     0.0     0.0     0.0003 0.9870148
## data$LDAPS_CC3            1    69.1    69.1    61.7219 4.501e-15 ***
## data$LDAPS_CC4            1     0.8     0.8     0.6990 0.4031308
## data$LDAPS_PPT2           1    14.2    14.2    12.7110 0.0003658 ***
## data$LDAPS_PPT3           1    12.9    12.9    11.5306 0.0006881 ***
## data$lon                  1    19.9    19.9    17.7428 2.558e-05 ***
## data$DEM                  1   280.8   280.8   250.7555 < 2.2e-16 ***
## data$Slope                1   431.3   431.3   385.0493 < 2.2e-16 ***
## data$lat                  1    10.5    10.5     9.3361 0.0022546 **
## Residuals              7569  8477.2     1.1
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Analysis of Variance Table
##
## Response: data$Next_Tmin
##                          Df  Sum Sq Mean Sq    F value    Pr(>F)
## data$station              1   768.1   768.1   685.8293 < 2.2e-16 ***
## data$Date                 1    19.0    19.0    16.9199 3.940e-05 ***
## data$Present_Tmax         1 17413.2 17413.2 15547.5760 < 2.2e-16 ***
## data$LDAPS_RHmin          1  2495.7  2495.7  2228.3013 < 2.2e-16 ***
## data$LDAPS_Tmax_lapse     1 11734.1 11734.1 10476.9817 < 2.2e-16 ***
## data$LDAPS_Tmin_lapse     1  4940.8  4940.8  4411.4592 < 2.2e-16 ***
## data$LDAPS_WS             1     8.0     8.0     7.1249 0.0076184 **
## data$LDAPS_LH             1    49.8    49.8    44.4447 2.800e-11 ***
## data$LDAPS_CC1            1     2.7     2.7     2.3900 0.1221585
## data$LDAPS_CC2            1     0.0     0.0     0.0003 0.9870148
## data$LDAPS_CC3            1    69.1    69.1    61.7219 4.501e-15 ***
## data$LDAPS_CC4            1     0.8     0.8     0.6990 0.4031308
## data$LDAPS_PPT2           1    14.2    14.2    12.7110 0.0003658 ***
## data$LDAPS_PPT3           1    12.9    12.9    11.5306 0.0006881 ***
## data$lon                  1    19.9    19.9    17.7428 2.558e-05 ***
## data$Slope                1     0.0     0.0     0.0000 0.9951434
## data$lat                  1     0.3     0.3     0.2971 0.5857077
## data$DEM                  1   722.2   722.2   644.8437 < 2.2e-16 ***
## Residuals              7569  8477.2     1.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observing the anovas above, let's remove the insignificant variables (with high pvalue) and build new models:

```
## Analysis of Variance Table
##
## Response: data$Next_Tmin
##                          Df  Sum Sq Mean Sq    F value    Pr(>F)
## data$station              1   768.1   768.1   684.7452 < 2.2e-16 ***
## data$Date                 1    19.0    19.0    16.8931 3.996e-05 ***
## data$Present_Tmax         1 17413.2 17413.2 15522.9997 < 2.2e-16 ***
## data$LDAPS_RHmin          1  2495.7  2495.7  2224.7789 < 2.2e-16 ***
## data$LDAPS_Tmax_lapse     1 11734.1 11734.1 10460.4206 < 2.2e-16 ***
## data$LDAPS_Tmin_lapse     1  4940.8  4940.8  4404.4859 < 2.2e-16 ***
## data$LDAPS_WS             1     8.0     8.0     7.1136 0.0076664 **
## data$LDAPS_LH             1    49.8    49.8    44.3744 2.902e-11 ***
## data$LDAPS_CC3            1    63.7    63.7    56.7938 5.401e-14 ***
## data$LDAPS_PPT2           1     9.2     9.2     8.1895 0.0042250 **
## data$LDAPS_PPT3           1    13.4    13.4    11.9502 0.0005494 ***
## data$lon                  1    17.5    17.5    15.5714 8.016e-05 ***
## data$DEM                  1   283.4   283.4   252.6185 < 2.2e-16 ***
```

```
## data$Slope               1   428.2   428.2   381.6841 < 2.2e-16 ***
## data$lat                 1    10.1    10.1     8.9759 0.0027445 **
## Residuals             7572  8494.0     1.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Response: data$Next_Tmin
##                          Df  Sum Sq Mean Sq    F value    Pr(>F)
## data$station              1   768.1   768.1   651.3226 < 2.2e-16 ***
## data$Date                 1    19.0    19.0    16.0685 6.168e-05 ***
## data$Present_Tmax         1 17413.2 17413.2 14765.3172 < 2.2e-16 ***
## data$LDAPS_RHmin          1  2495.7  2495.7  2116.1868 < 2.2e-16 ***
## data$LDAPS_Tmax_lapse     1 11734.1 11734.1  9949.8442 < 2.2e-16 ***
## data$LDAPS_Tmin_lapse     1  4940.8  4940.8  4189.5015 < 2.2e-16 ***
## data$LDAPS_WS             1     8.0     8.0     6.7664 0.0093071 **
## data$LDAPS_LH             1    49.8    49.8    42.2085 8.723e-11 ***
## data$LDAPS_CC3            1    63.7    63.7    54.0217 2.190e-13 ***
## data$LDAPS_PPT2           1     9.2     9.2     7.7897 0.0052675 **
## data$LDAPS_PPT3           1    13.4    13.4    11.3669 0.0007514 ***
## data$lon                  1    17.5    17.5    14.8114 0.0001198 ***
## data$DEM                  1   283.4   283.4   240.2881 < 2.2e-16 ***
## Residuals              7574  8932.2     1.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Model 1: data$Next_Tmin ~ data$station + data$Date + data$Present_Tmax +
##     data$LDAPS_RHmin + data$LDAPS_Tmax_lapse + data$LDAPS_Tmin_lapse +
##     data$LDAPS_WS + data$LDAPS_LH + data$LDAPS_CC3 + data$LDAPS_PPT2 +
##     data$LDAPS_PPT3 + data$lon + data$DEM + data$Slope + data$lat
## Model 2: data$Next_Tmin ~ data$station + data$Date + data$Present_Tmax +
##     data$LDAPS_RHmin + data$LDAPS_Tmax_lapse + data$LDAPS_Tmin_lapse +
##     data$LDAPS_WS + data$LDAPS_LH + data$LDAPS_CC3 + data$LDAPS_PPT2 +
##     data$LDAPS_PPT3 + data$lon + data$DEM
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   7572 8494.0
## 2   7574 8932.2 -2   -438.23 195.33 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observing the p-value in the anova above, we can conclude that there is a significant difference between the two models.

Let's now see the summary the two models:

```
## [1] "Model 1"
```

```
##
## Call:
## lm(formula = data$Next_Tmin ~ data$station + data$Date + data$Present_Tmax
+
##     data$LDAPS_RHmin + data$LDAPS_Tmax_lapse + data$LDAPS_Tmin_lapse +
##     data$LDAPS_WS + data$LDAPS_LH + data$LDAPS_CC3 + data$LDAPS_PPT2 +
##     data$LDAPS_PPT3 + data$lon + data$DEM + data$Slope + data$lat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3174 -0.6299  0.0733  0.7295  3.4855
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.206e+02  2.065e+01   5.843 5.35e-09 ***
## data$station          -1.003e-03  1.865e-03  -0.538 0.590732
## data$Date              1.117e-04  2.435e-05   4.589 4.53e-06 ***
## data$Present_Tmax      8.038e-02  6.135e-03  13.101  < 2e-16 ***
## data$LDAPS_RHmin       2.324e-02  1.827e-03  12.719  < 2e-16 ***
## data$LDAPS_Tmax_lapse  7.440e-02  1.064e-02   6.994 2.90e-12 ***
## data$LDAPS_Tmin_lapse  7.930e-01  1.252e-02  63.361  < 2e-16 ***
## data$LDAPS_WS          3.230e-02  6.131e-03   5.269 1.41e-07 ***
## data$LDAPS_LH          1.483e-03  4.193e-04   3.537 0.000407 ***
## data$LDAPS_CC3        -5.729e-01  8.017e-02  -7.146 9.75e-13 ***
## data$LDAPS_PPT2       -2.475e-02  7.718e-03  -3.206 0.001349 **
## data$LDAPS_PPT3        3.401e-02  1.154e-02   2.947 0.003217 **
## data$lon              -7.413e-01  1.651e-01  -4.490 7.22e-06 ***
## data$DEM              -9.875e-03  3.895e-04 -25.355  < 2e-16 ***
## data$Slope             2.971e-01  1.509e-02  19.689  < 2e-16 ***
## data$lat              -7.965e-01  2.659e-01  -2.996 0.002745 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.059 on 7572 degrees of freedom
##   (164 observations deleted due to missingness)
## Multiple R-squared:  0.8183, Adjusted R-squared:  0.8179
## F-statistic:  2273 on 15 and 7572 DF,  p-value: < 2.2e-16

## [1] "Model 2"

##
## Call:
## lm(formula = data$Next_Tmin ~ data$station + data$Date + data$Present_Tmax
+
##     data$LDAPS_RHmin + data$LDAPS_Tmax_lapse + data$LDAPS_Tmin_lapse +
##     data$LDAPS_WS + data$LDAPS_LH + data$LDAPS_CC3 + data$LDAPS_PPT2 +
##     data$LDAPS_PPT3 + data$lon + data$DEM)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -5.5933 -0.6651  0.0664  0.7372  3.4773
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            7.498e+01  2.063e+01   3.634 0.000281 ***
## data$station           7.676e-03  1.831e-03   4.193 2.78e-05 ***
## data$Date              1.091e-04  2.496e-05   4.372 1.25e-05 ***
## data$Present_Tmax      9.985e-02  6.198e-03  16.110  < 2e-16 ***
## data$LDAPS_RHmin       2.948e-02  1.808e-03  16.301  < 2e-16 ***
## data$LDAPS_Tmax_lapse  9.065e-02  1.080e-02   8.393  < 2e-16 ***
## data$LDAPS_Tmin_lapse  7.539e-01  1.251e-02  60.285  < 2e-16 ***
## data$LDAPS_WS          3.170e-02  6.282e-03   5.046 4.61e-07 ***
## data$LDAPS_LH          1.332e-03  4.295e-04   3.102 0.001929 **
## data$LDAPS_CC3        -7.098e-01  8.163e-02  -8.695  < 2e-16 ***
## data$LDAPS_PPT2       -2.862e-02  7.907e-03  -3.619 0.000297 ***
## data$LDAPS_PPT3        3.651e-02  1.183e-02   3.086 0.002037 **
## data$lon              -6.212e-01  1.624e-01  -3.824 0.000132 ***
## data$DEM              -3.847e-03  2.482e-04 -15.501  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.086 on 7574 degrees of freedom
##   (164 observations deleted due to missingness)
## Multiple R-squared:  0.8089, Adjusted R-squared:  0.8086
## F-statistic:  2467 on 13 and 7574 DF,  p-value: < 2.2e-16
```

The first model has better R square and adjusted R square values and smaller RME , this model is then the better performing one.

Application of training/testing paradigm with features selection + Applying cross-validation paradigm with features selection + Applying the bootstrapping paradigm with features selection

For Max temperature

Let's now partition our data into a training and test sub datasets, fit the training data, try different models and finally choose the best performin one.

```
## [1] "MAE= 7.70778423176"

## [1] "R2= 0.608103404768984"

## [1] "RMSE= 7.98207716891311"
```

For this fitted model R square is considered relatively low. More modifications to the model can improve R2 as well as decrease both MAE and RMSE. Let's now perform a Leave one out cross validation:

```
## Linear Regression
##
```

```
## 5884 samples
##   15 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 5883, 5883, 5883, 5883, 5883, 5883, ...
## Resampling results:
##
##   RMSE      Rsquared  MAE
##   1.389434  0.78551   1.075558
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

As observed in the results above there is a great improvement in our model as R2 moves up to 0.785 and MAE and RMSE drop to around 1. Next, we try a K-cross validation with k=10:

```
## Linear Regression
##
## 5884 samples
##   15 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 5295, 5297, 5295, 5296, 5294, 5296, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   1.388393  0.7860251  1.075569
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

As observed in the results above there is a slight improvement from CF to K-CF as R2 moves up to 0.786 and MAE and RMSE decrease slightly too. Let's also try Repeated K-fold cross validation where the repetition is done 3 times:

```
## Linear Regression
##
## 5884 samples
##   15 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 5295, 5297, 5295, 5296, 5294, 5296, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   1.389116  0.7857316  1.075856
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

The results are pretty similar to the previous ones. In conclusion, the K-fold cross validation is overall the best performing model for predicting the Next day's Max Temperature.

We can also try bootstrapping:

```
## BOOTSTRAP OF LINEAR MODEL  (method = rows)
##
## Original Model Fit
## ------------------
## Call:
## lm(formula = Next_Tmax ~ station + Date + Present_Tmax + LDAPS_RHmin +
##     LDAPS_Tmax_lapse + LDAPS_Tmin_lapse + LDAPS_WS + LDAPS_LH +
##     LDAPS_CC1 + LDAPS_CC2 + LDAPS_CC3 + LDAPS_CC4 + LDAPS_PPT2 +
##     lon + Slope, data = datatr)
##
## Coefficients:
##      (Intercept)           station              Date      Present_Tmax
##        1.861e+02         2.500e-02         7.657e-04         1.660e-01
##      LDAPS_RHmin  LDAPS_Tmax_lapse  LDAPS_Tmin_lapse          LDAPS_WS
##        3.250e-02         6.599e-01         1.250e-01        -1.641e-01
##         LDAPS_LH         LDAPS_CC1         LDAPS_CC2         LDAPS_CC3
##        6.474e-03        -1.285e+00        -5.338e-01        -8.987e-01
##        LDAPS_CC4        LDAPS_PPT2               lon             Slope
##       -1.044e+00         9.990e-02        -1.544e+00         6.997e-02
##
## Bootstrap SD's:
##      (Intercept)           station              Date      Present_Tmax
##     2.966451e+01      2.576027e-03      4.434864e-05      1.120311e-02
##      LDAPS_RHmin  LDAPS_Tmax_lapse  LDAPS_Tmin_lapse          LDAPS_WS
##     3.212711e-03      1.874802e-02      1.954074e-02      9.550924e-03
##         LDAPS_LH         LDAPS_CC1         LDAPS_CC2         LDAPS_CC3
##     6.152668e-04      1.189784e-01      1.510809e-01      1.782718e-01
##        LDAPS_CC4        LDAPS_PPT2               lon             Slope
##     1.443037e-01      1.311967e-02      2.336785e-01      1.359854e-02
```

We can observe the coefficients as well as the SDs for this model where the number of bootstraps was 10000.

Let's now try non linear regression. For example, here we try a similar to model 1, only where the humidity for next day is squared:

```
## [1] "MAE= 7.71628994332065"

## [1] "R2= 0.608630463009126"

## [1] "RMSE= 7.99126138420446"
```

Observing the values of R2, RMSE and MAE this fourth model has a worse performance than all of our previous models Adding non linear variables to the model can be rather complicated when there is no clear or no known polynomial relation between the variable

and the thing we need to predict. We can also try variable interactions. For example we add an interaction variable where we multiply the two precipitations rates present in model 1:

```
## [1] "MAE= 7.70980650812993"

## [1] "R2= 0.608768945294719"

## [1] "RMSE= 7.98210228953212"
```

This new interaction model has worse results too. For interaction variables too, it is not simple to find the optimal ones. Adding some random interactions can worsen the performance of the models as seen above. Knowing more profound properties of the data variables can help with this task.

For Min temperature

We now do the same for Min Temperature:

```
## [1] "MAE= 0.792156871314049"

## [1] "R2= 0.866684693384554"

## [1] "RMSE= 1.00660963783085"
```

As seen through the three indicator above the regression model already has pretty decent results -compared to the regression model for Max Temperature. We have a relatively high R2=0.866 and rather low errors. Nevertheless, we move on to try different models too. Let's start with Leave one out cross validation:

```
## Linear Regression
##
## 5884 samples
##   15 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 5883, 5883, 5883, 5883, 5883, 5883, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   1.073207  0.7920487  0.8438548
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

The results of this second model are comparable to the first one (good too), however the first one has better R2 and slighly smaller errors. So far we retain the first model. Now we try K-fold cross-validation.

```
## Linear Regression
##
## 5884 samples
```

```
##   15 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 5295, 5295, 5296, 5295, 5295, 5297, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   1.072514  0.7924258  0.8437626
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

This 3rd model is quite similar to the second one. Again, for now, we retain the 1st model as best performing one.

For Repeated K-fold cross-validation:

```
## Linear Regression
##
## 5884 samples
##   15 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 5295, 5295, 5296, 5295, 5295, 5297, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   1.072663  0.7924018  0.843743
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

The Repeated K-fold cross-validation model too is more similar to the 2nd and 3rd ones. In conclusion, the first model is the best performing one for the prediction of the Next day's Min temperature.

We can also try bootstrapping:

```
## BOOTSTRAP OF LINEAR MODEL  (method = rows)
##
## Original Model Fit
## ------------------
## Call:
## lm(formula = Next_Tmin ~ station + Date + Present_Tmax + LDAPS_RHmin +
##     LDAPS_Tmax_lapse + LDAPS_Tmin_lapse + LDAPS_WS + LDAPS_LH +
##     LDAPS_CC3 + LDAPS_PPT2 + LDAPS_PPT3 + lon + DEM + Slope +
##     lat, data = datatr)
##
## Coefficients:
##      (Intercept)           station              Date      Present_Tmax
##        1.072e+02        -1.189e-03         2.472e-04         6.872e-02
```

```
##      LDAPS_RHmin  LDAPS_Tmax_lapse  LDAPS_Tmin_lapse          LDAPS_WS
##        2.458e-02         8.075e-02         8.099e-01         4.211e-02
##         LDAPS_LH         LDAPS_CC3        LDAPS_PPT2        LDAPS_PPT3
##        1.205e-03        -4.573e-01        -3.298e-02         4.513e-02
##              lon               DEM             Slope               lat
##       -5.873e-01        -1.011e-02         3.084e-01        -1.028e+00
##
## Bootstrap SD's:
##      (Intercept)           station              Date      Present_Tmax
##     2.312208e+01      2.125923e-03      3.509029e-05      7.256352e-03
##      LDAPS_RHmin  LDAPS_Tmax_lapse  LDAPS_Tmin_lapse          LDAPS_WS
##     2.151144e-03      1.247027e-02      1.427556e-02      8.532892e-03
##         LDAPS_LH         LDAPS_CC3        LDAPS_PPT2        LDAPS_PPT3
##     4.940586e-04      9.255928e-02      7.132178e-03      1.151981e-02
##              lon               DEM             Slope               lat
##     1.840505e-01      3.961476e-04      1.644182e-02      3.230663e-01
```

We can observe the coefficients as well as the SDs for this model where the number of bootstraps was 10000.

Let's now try non linear regression. For example, here we try a similar to model 1, only where the humidity for next day is squared:

```
## [1] "MAE= 0.797616456415802"

## [1] "R2= 0.864430495358172"

## [1] "RMSE= 1.01192488848523"
```

Observing the values of R2, RMSE and MAE this fourth model is similar to the first one. It has a slightly lower R2. Adding non linear variables to the model can be rather complicated when there is no clear or no known polynomial relation between the variable and the thing we need to predict. We can also try variable interactions. For example we add an interaction variable where we multiply the two precipitations rates present in model 1:

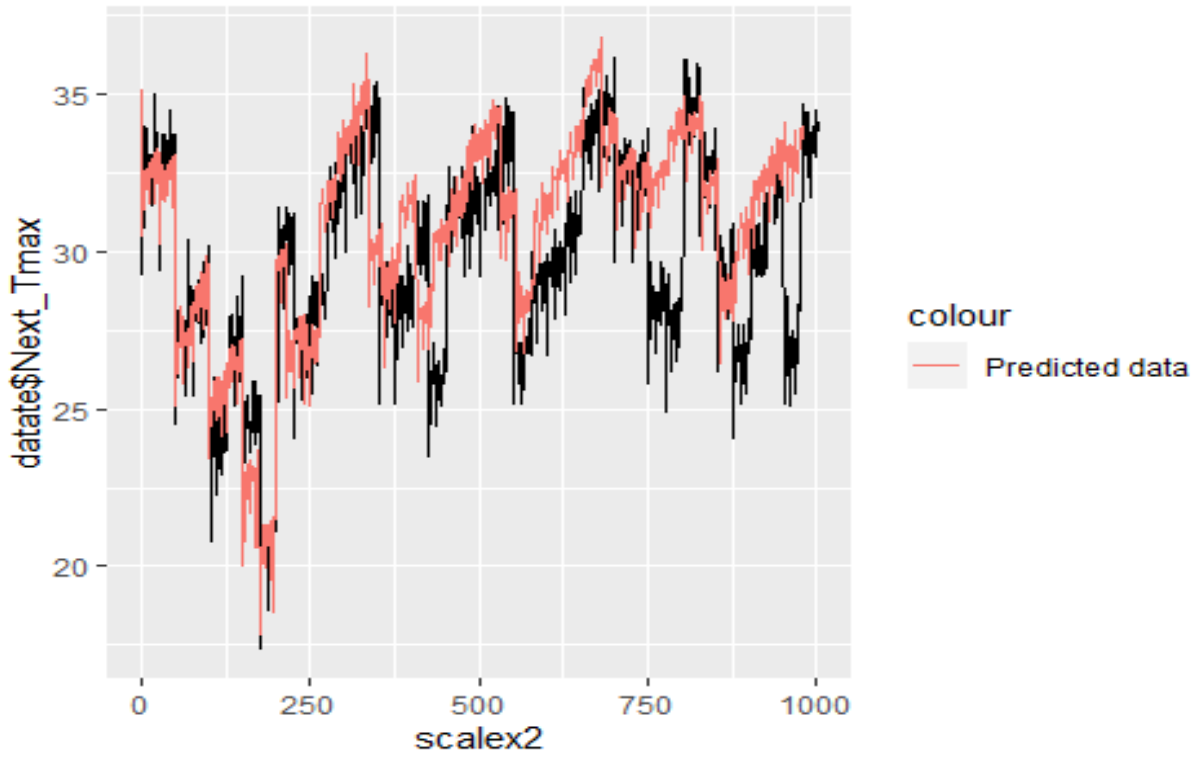
```
## [1] "MAE= 0.792841487669694"

## [1] "R2= 0.866503251306275"

## [1] "RMSE= 1.00713476134798"
```

This new interaction model has basically similar results to model 1 (slightly better errors). For interaction variables too, it is not simple to find the optimal ones.

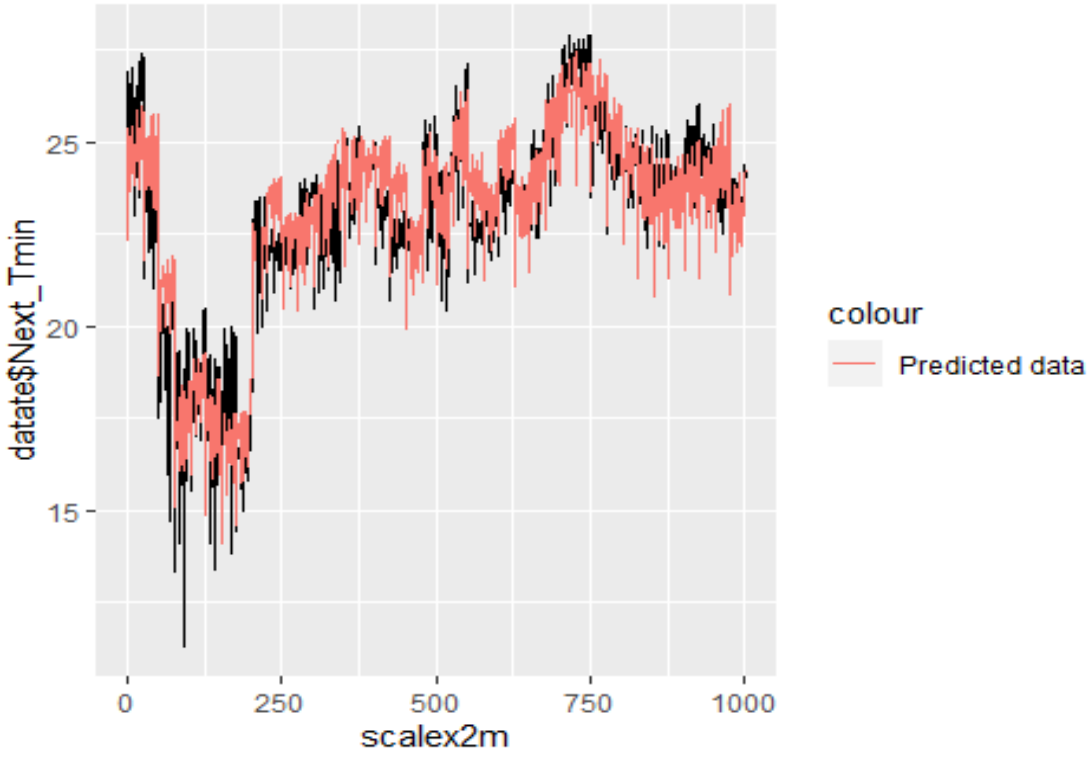
Conclusion & Best models' visualization

After trying multiple methods and models, we chose the best one for predicting each of the Max and Min temperatures. Let's now trace the graphs containing the predicted and actual data for each model.

*For max temperature:



And now for mean temperature:



Through the first graph above we can evaluate the performance of our model for predicting next day's Max temperature: we can notice that the two lines follow a very similar trend and overlap for the most part- indicating a good prediction model for our variable.

The accuracy of the model seems to drop in some parts, however. This can be especially seen around the end of the graph.

All in all, our best model for Max Temperature has a relatively good prediction, given the complexity of the variables and the simplicity of the presented model. More variables (for example variables from a few days ago not just present day) and more complex models (more interaction variables and non-linear ones) might make room for improvement for this model.

As for the best model for Tmin (second graph above), we can easily see that this model has a really good performance (an even better one than for Tmax), the gap between the two graphs of actual and predicted data is really minimal indicating a good accuracy of the built model.

Here, interaction variables seem to have a positive effect on predicting minimal temperatures: for this model the precipitations were multiplied and added as a new variable. Precipitations have an proportional relation with Min temperature as more precipitations would usually lower the ambient temperature. This fact was used to try and improve the model.

We must note though that this interaction model is very similar to the 1st regression one so the the effect of the added interaction variable should be perceived with caution. In this case too more variables and modifications could give an even better performance of models predicting next day's temperature.

Discussion

The prediction of the next day's min and max temperatures is one of the most needed and vital predictions for many sectors. As a result, an accurate prediction is of great importance.

Through the presented dataset, we tried to visualize and explore the many variables this database has: an understanding of these variables, their correlation, distribution, variance and many other factors -along with finding possible explanations to why they are- can help build more comprehensive models and therefore better performing ones.

After exploring the dataset and explaining some of the phenomenas stemming from that exploration, we moved towards building the regression models: multiple linear regression models, models with interaction and non-linear variables, models built on cross validation (LOCV, K-Fold, Repeated K-fold) and bootstrap models were built and compared. Using indicators (R square, MAE, RMSE), each model was evaluated and then for each case (Tmax and Tmin), we chose the best model.

As mentioned above for Tmax the K-Fold cross validation (K=10) model was the best performing one: RMSE=1.388393, Rsquared=0.7860251, MAE=1.075569. This model has acceptable results but there is room for improvement. A more in-depth study of variables and their interaction. We can also note that the prediction of Tmax can become more complicated due to environmental factors such as global warming, since higher temperatures keep increasing at an unprecedented and rather unknown level.

As for Tmin, the best model was the interaction model mentioned in the conclusion above. Here we had an even better prediction accuracy: MAE= 0.792841487669694, R2= 0.866503251306275, RMSE= 1.00713476134798. For Tmin, the best linear regression model, the non-linear model (humidity was squared) and the interaction model (precipitations multiplied) had very similar results. Different to the case of Tmax where interaction and non-linear models worsened the performance of the model, for Tmin there was no to little improvement. As a result, more exploration of the use of interaction and non-linear models can be of help, especially for variables such as humidity and precipitation that are know to have a direct effect on lowering ambient temperature.