Water Evaporation Forecasting

2022-07-25

Executive Summary

The Melbourne Water Corporation (MWC) is known for managing the supply of water in Melbourne, Australia to generate a sheet/report relating to water evaporation. However, due to the fluctuations in the climate of Melbourne, MWC is reconsidering its previous forecasting/estimation.

The Corporation wants to know whether specific changes in the climate will lead to a change in water evaporation. To examine the current scenario, one statistical model will be developed and implemented to make predictions on evaporation.

The report details the procedures and outcomes of such a model, including the specification that the Minimum temperature ranging 17-25 degrees Celsius with the Relative humidity ranging 20-40% at 9 am will be the reason for the occurrence of more than 10mm evaporation.

Methods

- **Software and Frameworks**
  - Software : RStudio
  - Frameworks for this section : tidyverse, dplyr, ggplot, lubridate
- **Features used in this section**
  - Month,
  - Day of the week,
  - Maximum temperature in degrees Celsius,
  - Minimum temperature in degrees Celsius, and
  - Relative humidity, as measured at 9am.
- **Bivariate Summaries**
  - January has the highest median evaporation whilst June has the lowest. April has the largest variability (IQR) in evaporation whilst June has the smallest. There are outliers in all the months except June and September.
  - Saturday has the highest median evaporation whilst Friday has the lowest. Saturday has the largest variability (IQR) in evaporation whilst Friday has the smallest. There are outliers on all the weekdays except Tuesday.
  - There appears to be a moderate, positive, linear relationship between the Maximum Temperature (Degree Celsius) and evaporation (in mm).
  - There appears to be a moderate, positive, linear relationship between the Minimum Temperature (Degree Celsius) and evaporation (in mm).
  - There appears to be a moderate, negative, linear relationship between the 9 am Relative Humidity and evaporation (in mm).

- **Model Selection**
  - Significant Features in model:
    * Month
    * Minimum temperature in degrees Celsius, and
    * Relative humidity, as measured at 9am.
- **Comparison between Bivariate Analysis and Model Selection**
  - In Bivariate Analysis, we choose 5 features to work with. The Maximum temperature shown capable relationship with evaporation. However, Maximum temperature is not having p-value that significant at the 5. Therefore, Maximum temperature is not in the statistical model anymore.
- **Model diagnostics**
  - Linearity looks reasonable - want to see random scatter around zero. Almost no change in trend as we go left to right in the residuals vs fitted plot.
  - Homoscedasticity looks reasonable - want to see no change in the vertical spread as we look from left to right. Almost no change in trend as we go left to right in the scale vs location plot.
  - Normality looks reasonable, despite few outliers - mostly follows the trend line.
  - Independence does look justified – The interaction of Relative humidity per month is low. We can verify that by observing the p-value of the interaction term.

Results

- **Model interpretation**
  - Target/Response variable : Evaporation in mm
  - Predictors : Month, Minimum temperature in degrees Celsius, and Relative humidity, as measured at 9am.
  - Considering the numerical variables/features only:
    * The intercept is 10.73 and the slopes are 0.37 and -0.148 with respect to Minimum temperature and Humidity at 9 am.
    * Interpretation of intercept: If the city's climate has 0-degree Celsius minimum temperature and 0% humidity at 9 am, we expect a 10.73% increase in evaporation.
    * Interpretation of slope: If the Minimum temperature increases by 1 degree Celsius, the evaporation is expected to increase by 0.37%. If the Relative Humidity increases by 1%, the evaporation is expected to decrease by 0.148%.
    * There is a statistically significant relationship between the Minimum temperature and Evaporation, and Relative Humidity at 9 am and Evaporation.
  - Considering the categorical variable only:
    * Month is the categorical variable having very small p-value, that indicates the variable is suitable with response variable, Furthermore p-value is in the significant criteria (5%).

Discussions

- **Prediction**
  - We can say with 95% confidence that:
  - February 29, 2020, if this day has a minimum temperature of 13.8 degrees and reaches a maximum of 23.2 degrees, and has 74% humidity at 9am is having modrate evaporation in between 4.7mm to 7mm.
  - December 25, 2020, if this day has a minimum temperature of 16.4 degrees and reaches a maximum of 31.9 degrees, and has 57% humidity at 9am is having high evaporation in between 7.4mm to 9.6mm.

- January 13, 2020, if this day has a minimum temperature of 26.5 degrees and reaches a maximum of 44.3 degrees, and has 35% humidity at 9am is having critically high evaporation in between 11.6mm to 17.7mm
- July 6, 2020, if this day has a minimum temperature of 6.8 degrees and reaches a maximum of 10.6 degrees, and has 76% humidity at 9am is having low evaporation in between 1.5mm to 3.3mm
- Hence,if the city's minimum temperature is higher than average and humidity at 9 am is lower than average then, probability of evaporation more than 10mm is possible.

Conclusion

The Melbourne Water Corporation (MWC) is known for managing the supply of water in Melbourne, Australia to generate a report relating to water evaporation. However, due to the recent change in the climate of Melbourne, MWC is reconsidering its previous forecasting/estimation.

MWC wants to know the factors affecting the evaporation based on the data of the previous financial year, to aid their management of their Cardinia Reservoir, in the city's South East.

After implementation of the statistical model, the event of evaporation (more than 10mm) will occur if the Minimum temperature of the city is between 17 to 25 degrees Celsius and the humidity of the city in the morning (at 9 am) is less than 35-40%.

The model is considering the city's humidity (at 9 am), minimum temperature, and specific month to forecast the evaporation (in mm) with a 95% confidence interval.

A limitation of the model that should be addressed is the model is not considering the maximum temperature and specific weekday of the city. Therefore, the model may give some fluctuations when forecasting evaporation.

Finally, considering the data of previous financial year and recent climate change in Melbourne, we can say that the evaporation is majorly affected by city's Minimum temperature, then relative humdity at 9 am, and specific month.

Appendix

Load the libraries and data

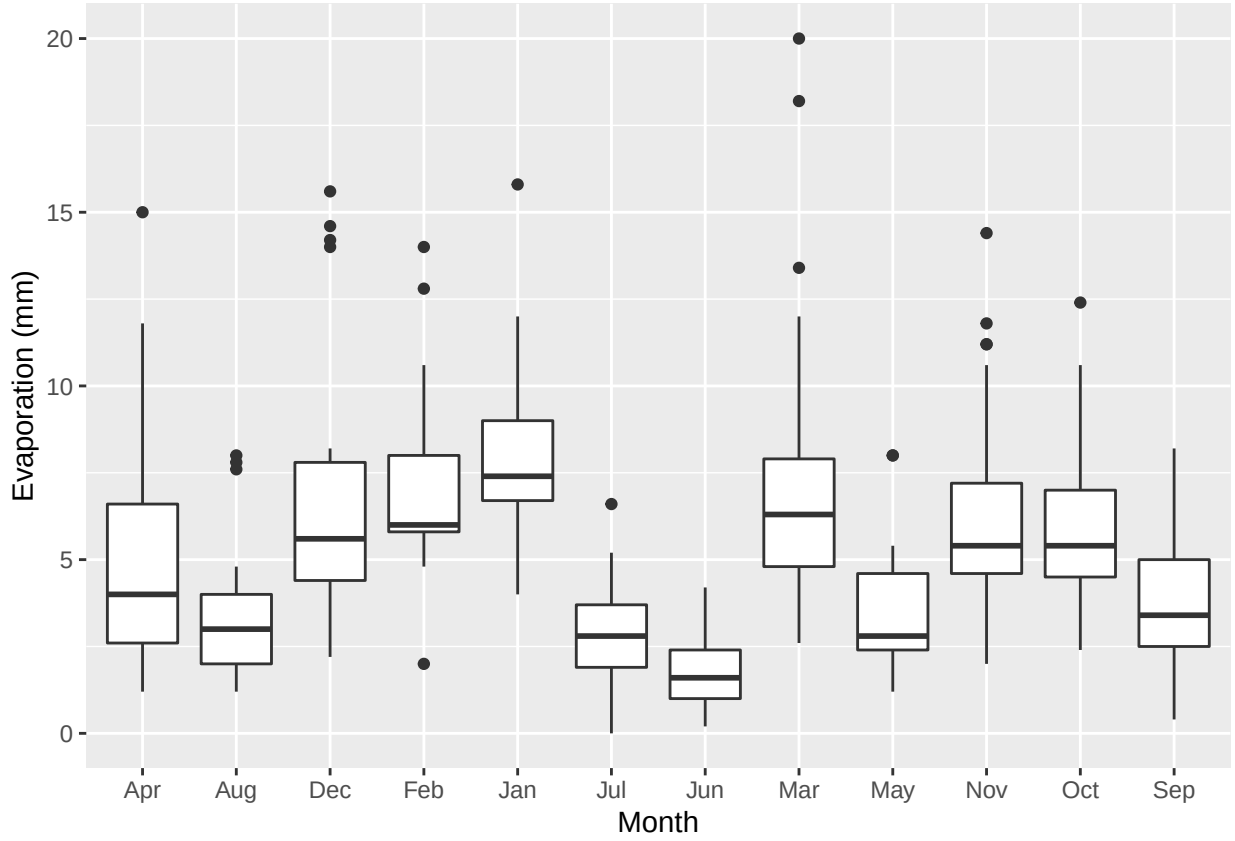
```
# load the libraries
pacman::p_load(tidyverse,dplyr,lubridate,inspectdf,tidyr,stringr,data.table,caret,tidymodels,skimr)
# load the data
data <- read_csv("C:\\Users\\Admin\\Downloads\\melbourne.csv")
```

```
## New names:
## Rows: 300 Columns: 22
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (5): Date, Direction of maximum wind gust, 9am wind direction, 9am win... dbl
## (16): ...1, Minimum temperature (Deg C), Maximum Temperature (Deg C), R... time
## (1): Time of maximum wind gust
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

Bivariate Summaries

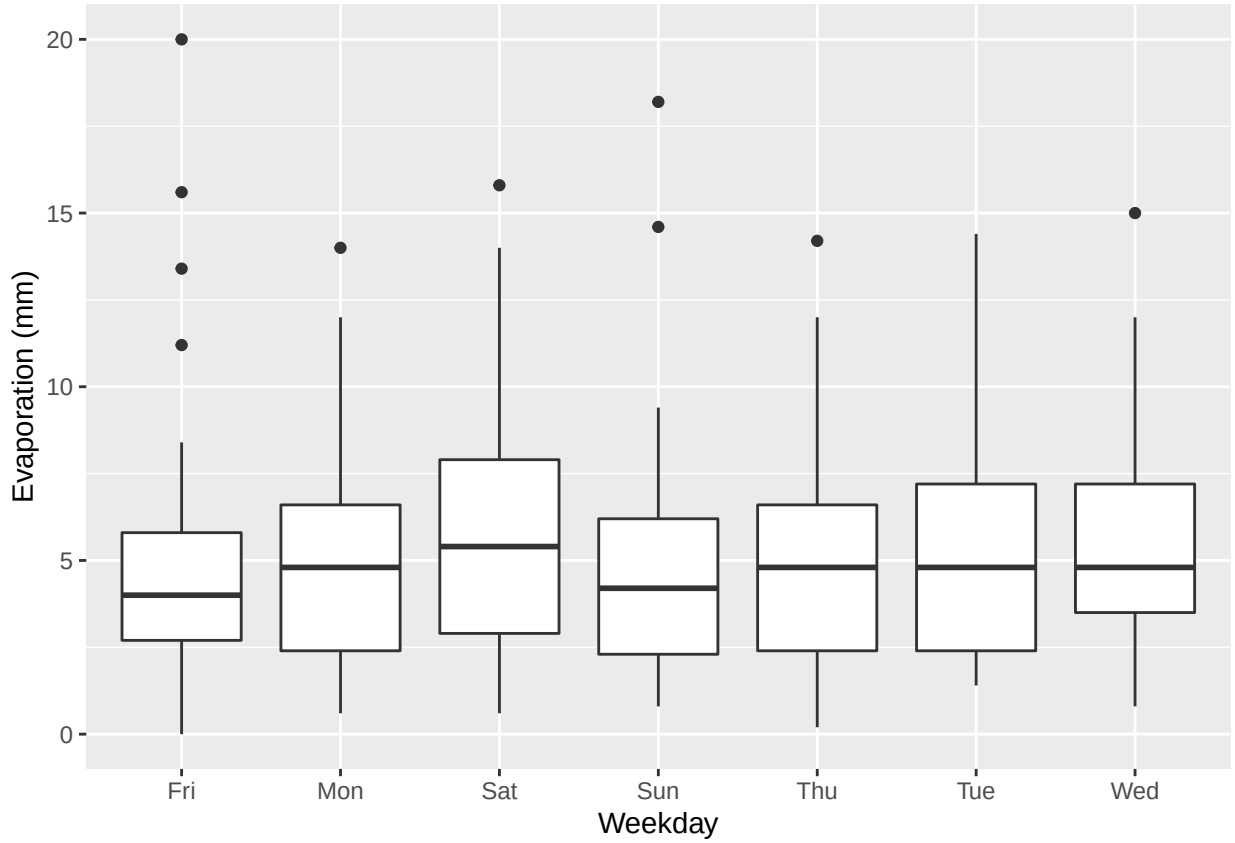
```
# converted into date-type.
data$Date <- as.Date(data$Date, format="%Y-%m-%d")
# create two variables named "Month" and "Weekday"
data <- data %>%
  mutate(Month = strftime(data$Date, format = "%b"))
data <- data %>%
  mutate(Weekday = strftime(data$Date, format = "%a"))
# convert both categorical variables into factor
data$Month <- factor(data$Month)
data$Weekday <- factor(data$Weekday)
# select the data as per the guidelines
data <- data %>%
  select(`Evaporation (mm)`,Month,Weekday,`Maximum Temperature (Deg C)`,`Minimum temperature (Deg C)`,`

# visualize the relationship between all variables with target variable.
ggplot(data, aes(x=Month,y=`Evaporation (mm)`)) + geom_boxplot()
```

```
## Warning: Removed 8 rows containing non-finite values (stat_boxplot).
```



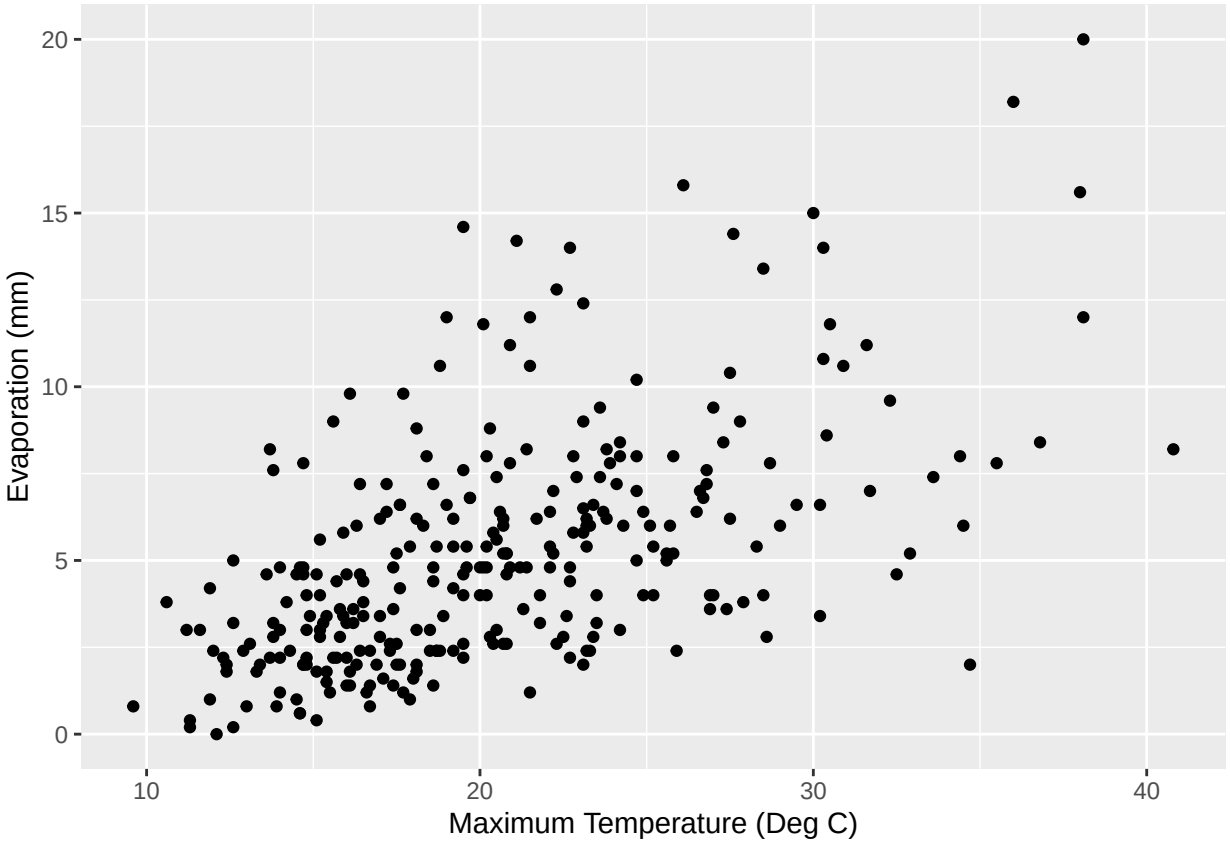

```
ggplot(data, aes(x=Weekday,y=`Evaporation (mm)`)) + geom_boxplot()
```

```
## Warning: Removed 8 rows containing non-finite values (stat_boxplot).
```



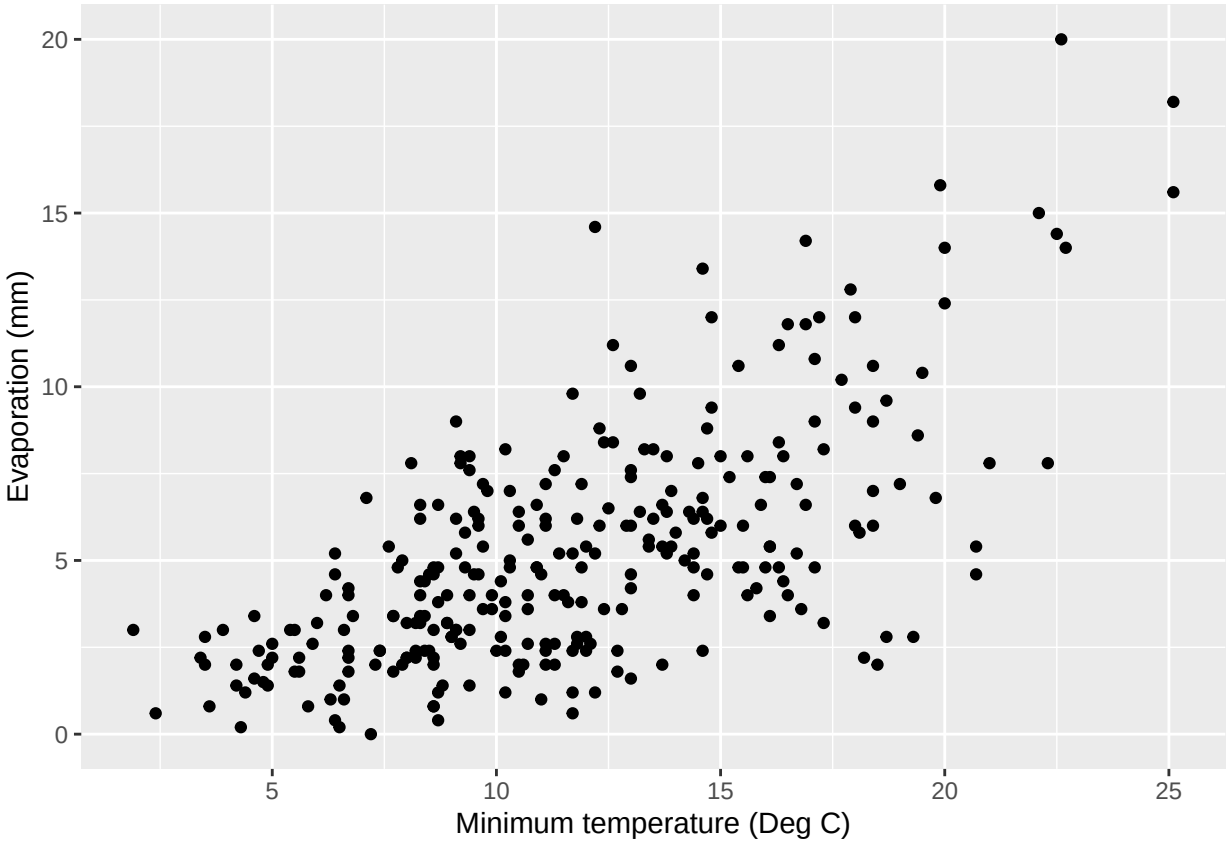

```
ggplot(data, aes(x=`Maximum Temperature (Deg C)`,y=`Evaporation (mm)`)) + geom_point()
```

```
## Warning: Removed 8 rows containing missing values (geom_point).
```



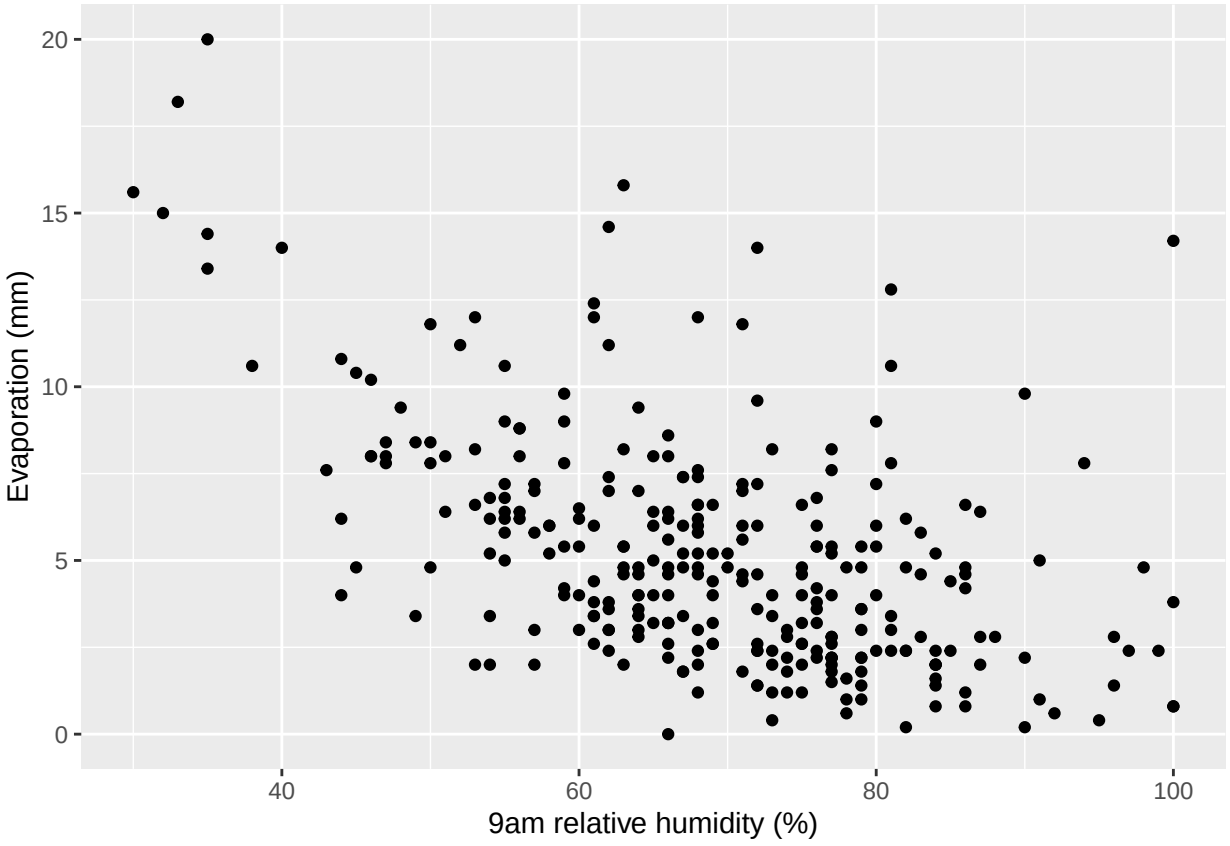

```
ggplot(data, aes(x=`Minimum temperature (Deg C)`,y=`Evaporation (mm)`)) + geom_point()
```

```
## Warning: Removed 8 rows containing missing values (geom_point).
```




```
ggplot(data, aes(x=`9am relative humidity (%)`,y=`Evaporation (mm)`)) + geom_point()
```

```
## Warning: Removed 8 rows containing missing values (geom_point).
```



Model Selection

```
# model with all the features.
model <- lm(`Evaporation (mm)` ~
              Month +
              Weekday +
              `Maximum Temperature (Deg C)` +
              `Minimum temperature (Deg C)` +
              `9am relative humidity (%)` +
              Month:`9am relative humidity (%)`,
            data=data)
# model description for numerical variables.
summary(model)
```

```
##
## Call:
## lm(formula = `Evaporation (mm)` ~ Month + Weekday + `Maximum Temperature (Deg C)` +
##     `Minimum temperature (Deg C)` + `9am relative humidity (%)` +
##     Month:`9am relative humidity (%)`, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8354 -1.0912 -0.0374  0.9608  9.7555
##
```

```
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           9.83232    2.60550   3.774 0.000199 ***
## MonthAug                            -10.66515    3.70259  -2.880 0.004302 **
## MonthDec                             -3.14599    2.99747  -1.050 0.294901
## MonthFeb                             -4.93048    3.75282  -1.314 0.190069
## MonthJan                             -2.21328    3.91299  -0.566 0.572137
## MonthJul                             -5.54858    3.62425  -1.531 0.126995
## MonthJun                            -10.83689    4.33645  -2.499 0.013071 *
## MonthMar                             2.65312    2.79810   0.948 0.343914
## MonthMay                             -6.10367    3.34185  -1.826 0.068931 .
## MonthNov                             -1.25597    2.91015  -0.432 0.666403
## MonthOct                             -8.27953    3.20519  -2.583 0.010337 *
## MonthSep                             -5.22342    3.76785  -1.386 0.166839
## WeekdayMon                            0.44173    0.49327   0.896 0.371343
## WeekdaySat                            0.99060    0.48352   2.049 0.041493 *
## WeekdaySun                            0.33711    0.47034   0.717 0.474182
## WeekdayThu                            0.05409    0.47301   0.114 0.909042
## WeekdayTue                            0.27889    0.46731   0.597 0.551152
## WeekdayWed                            0.15938    0.47429   0.336 0.737106
## `Maximum Temperature (Deg C)`         0.02002    0.03590   0.558 0.577543
## `Minimum temperature (Deg C)`         0.36221    0.04763   7.604 5.24e-13 ***
## `9am relative humidity (%)`          -0.14425    0.03369  -4.281 2.62e-05 ***
## MonthAug:`9am relative humidity (%)`  0.15687    0.05305   2.957 0.003392 **
## MonthDec:`9am relative humidity (%)`  0.05585    0.04368   1.279 0.202186
## MonthFeb:`9am relative humidity (%)`  0.07786    0.05693   1.368 0.172600
## MonthJan:`9am relative humidity (%)`  0.04133    0.05953   0.694 0.488207
## MonthJul:`9am relative humidity (%)`  0.07909    0.05098   1.551 0.122065
## MonthJun:`9am relative humidity (%)`  0.13490    0.05623   2.399 0.017139 *
## MonthMar:`9am relative humidity (%)` -0.02200    0.04135  -0.532 0.595048
## MonthMay:`9am relative humidity (%)`  0.07893    0.04719   1.673 0.095616 .
## MonthNov:`9am relative humidity (%)`  0.02879    0.04361   0.660 0.509781
## MonthOct:`9am relative humidity (%)`  0.14020    0.04848   2.892 0.004151 **
## MonthSep:`9am relative humidity (%)`  0.07930    0.05705   1.390 0.165737
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.115 on 260 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.6504, Adjusted R-squared:  0.6087
## F-statistic:  15.6 on 31 and 260 DF,  p-value: < 2.2e-16
```

```
# model description fro categorical variables.
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Evaporation (mm)
##                                    Df  Sum Sq Mean Sq F value    Pr(>F)
## Month                              11 1145.08  104.10 23.2606 < 2.2e-16 ***
## Weekday                             6   32.28    5.38  1.2021  0.305513
## `Maximum Temperature (Deg C)`       1  201.55  201.55 45.0357 1.206e-10 ***
## `Minimum temperature (Deg C)`       1  347.45  347.45 77.6377 < 2.2e-16 ***
## `9am relative humidity (%)`         1  300.74  300.74 67.1998 1.128e-14 ***
```

```
## Month:‘9am relative humidity (%)‘  11  137.24   12.48  2.7878  0.001909 **
## Residuals                          260 1163.58    4.48
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
# model after removing the variable having highest p-value not in defined significant level.
# remove Maximum Temperature (Deg C) variable.
model <- lm(`Evaporation (mm)` ~
              Month +
              Weekday +
              `Minimum temperature (Deg C)` +
              `9am relative humidity (%)` +
              Month:`9am relative humidity (%)`,
            data=data)
summary(model)
```

```
##
## Call:
## lm(formula = ‘Evaporation (mm)‘ ~ Month + Weekday + ‘Minimum temperature (Deg C)‘ +
##     ‘9am relative humidity (%)‘ + Month:‘9am relative humidity (%)‘,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9404 -1.0856 -0.0755  0.9788  9.7969
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          10.36798    2.41879   4.286 2.56e-05 ***
## MonthAug                            -10.86640    3.68010  -2.953  0.00344 **
## MonthDec                             -2.97519    2.97785  -0.999  0.31867
## MonthFeb                             -5.11470    3.73331  -1.370  0.17186
## MonthJan                             -2.63133    3.83544  -0.686  0.49329
## MonthJul                             -5.82805    3.58470  -1.626  0.10520
## MonthJun                            -10.97163    4.32399  -2.537  0.01175 *
## MonthMar                             2.63710    2.79425   0.944  0.34617
## MonthMay                             -6.26680    3.32462  -1.885  0.06055 .
## MonthNov                             -1.28934    2.90569  -0.444  0.65761
## MonthOct                             -8.24898    3.20049  -2.577  0.01050 *
## MonthSep                             -5.26635    3.76209  -1.400  0.16275
## WeekdayMon                            0.41640    0.49052   0.849  0.39672
## WeekdaySat                            0.97035    0.48152   2.015  0.04491 *
## WeekdaySun                            0.32752    0.46940   0.698  0.48596
## WeekdayThu                            0.03773    0.47147   0.080  0.93628
## WeekdayTue                            0.25268    0.46432   0.544  0.58678
## WeekdayWed                            0.12667    0.47002   0.269  0.78776
## ‘Minimum temperature (Deg C)‘         0.37008    0.04543   8.146 1.56e-14 ***
## ‘9am relative humidity (%)‘          -0.14670    0.03336  -4.397 1.60e-05 ***
## MonthAug:‘9am relative humidity (%)‘  0.15830    0.05292   2.991  0.00304 **
## MonthDec:‘9am relative humidity (%)‘  0.05388    0.04348   1.239  0.21640
## MonthFeb:‘9am relative humidity (%)‘  0.08106    0.05657   1.433  0.15309
## MonthJan:‘9am relative humidity (%)‘  0.04848    0.05806   0.835  0.40452
## MonthJul:‘9am relative humidity (%)‘  0.08160    0.05072   1.609  0.10886
## MonthJun:‘9am relative humidity (%)‘  0.13567    0.05614   2.417  0.01635 *
```

```
## MonthMar:‘9am relative humidity (%)‘  -0.02162    0.04128  -0.524  0.60099
## MonthMay:‘9am relative humidity (%)‘   0.08022    0.04708   1.704  0.08958 .
## MonthNov:‘9am relative humidity (%)‘   0.02885    0.04355   0.662  0.50826
## MonthOct:‘9am relative humidity (%)‘   0.13940    0.04839   2.881  0.00430 **
## MonthSep:‘9am relative humidity (%)‘   0.07870    0.05696   1.382  0.16830
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.113 on 261 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.6499, Adjusted R-squared:  0.6097
## F-statistic: 16.15 on 30 and 261 DF,  p-value: < 2.2e-16
```

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Evaporation (mm)
##                                   Df  Sum Sq Mean Sq  F value    Pr(>F)
## Month                             11 1145.08  104.10  23.3221 < 2.2e-16 ***
## Weekday                            6   32.28    5.38   1.2053  0.303834
## ‘Minimum temperature (Deg C)‘      1  505.05  505.05 113.1508 < 2.2e-16 ***
## ‘9am relative humidity (%)‘        1  342.69  342.69  76.7767 2.482e-16 ***
## Month:‘9am relative humidity (%)‘ 11  137.85   12.53   2.8076  0.001775 **
## Residuals                        261 1164.98    4.46
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
# model after removing the variable having highest p-value not in defined significant level.
# remove Weekday variable.
model <- lm(`Evaporation (mm)` ~
              Month +
              `Minimum temperature (Deg C)` +
              `9am relative humidity (%)` +
              Month:`9am relative humidity (%)`,
            data=data)
```

Model Interpretation

```
# Finalized model description for numerical variables.
summary(model)
```

```
##
## Call:
## lm(formula = ‘Evaporation (mm)‘ ~ Month + ‘Minimum temperature (Deg C)‘ +
##     ‘9am relative humidity (%)‘ + Month:‘9am relative humidity (%)‘,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2366 -1.1261 -0.0234  0.9798  9.6279
```

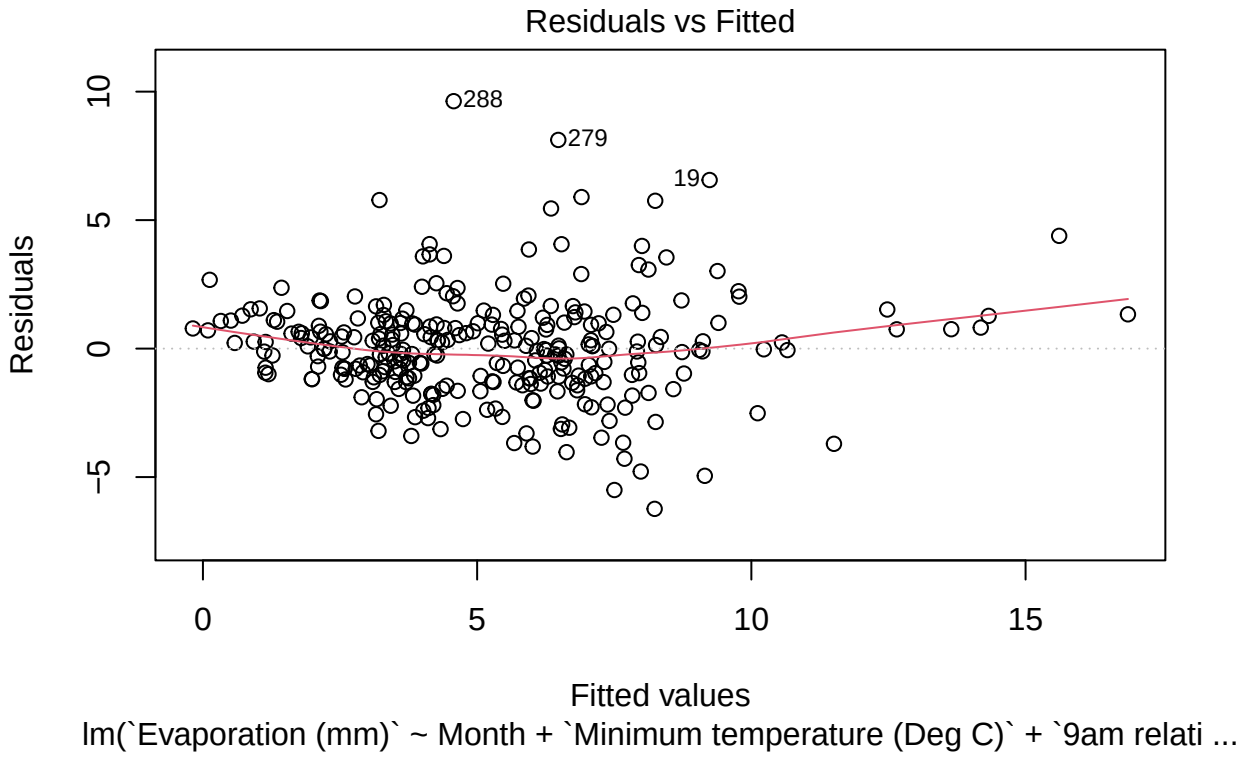
```
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          10.73020    2.36204   4.543 8.42e-06 ***
## MonthAug                            -11.77843    3.63466  -3.241  0.00134 **
## MonthDec                             -2.81612    2.96701  -0.949  0.34341
## MonthFeb                             -5.35322    3.72344  -1.438  0.15169
## MonthJan                             -2.17651    3.78975  -0.574  0.56624
## MonthJul                             -6.11033    3.57233  -1.710  0.08834 .
## MonthJun                            -11.67219    4.27738  -2.729  0.00678 **
## MonthMar                              2.26799    2.76406   0.821  0.41265
## MonthMay                             -6.32407    3.31136  -1.910  0.05723 .
## MonthNov                             -1.23184    2.89926  -0.425  0.67127
## MonthOct                             -7.98027    3.16532  -2.521  0.01228 *
## MonthSep                             -6.11998    3.71556  -1.647  0.10071
## `Minimum temperature (Deg C)`         0.37021    0.04472   8.278 6.06e-15 ***
## `9am relative humidity (%)`          -0.14776    0.03320  -4.450 1.26e-05 ***
## MonthAug:`9am relative humidity (%)`  0.17099    0.05229   3.270  0.00122 **
## MonthDec:`9am relative humidity (%)`  0.05178    0.04333   1.195  0.23316
## MonthFeb:`9am relative humidity (%)`  0.08479    0.05644   1.502  0.13418
## MonthJan:`9am relative humidity (%)`  0.04170    0.05732   0.728  0.46755
## MonthJul:`9am relative humidity (%)`  0.08589    0.05054   1.699  0.09044 .
## MonthJun:`9am relative humidity (%)`  0.14477    0.05559   2.604  0.00972 **
## MonthMar:`9am relative humidity (%)` -0.01659    0.04086  -0.406  0.68513
## MonthMay:`9am relative humidity (%)`  0.08027    0.04690   1.711  0.08819 .
## MonthNov:`9am relative humidity (%)`  0.02826    0.04347   0.650  0.51616
## MonthOct:`9am relative humidity (%)`  0.13512    0.04785   2.824  0.00510 **
## MonthSep:`9am relative humidity (%)`  0.09253    0.05624   1.645  0.10107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.111 on 267 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.6425, Adjusted R-squared:  0.6104
## F-statistic:    20 on 24 and 267 DF,  p-value: < 2.2e-16
```

```
# Finalized model description for categorical variables.
anova(model)
```

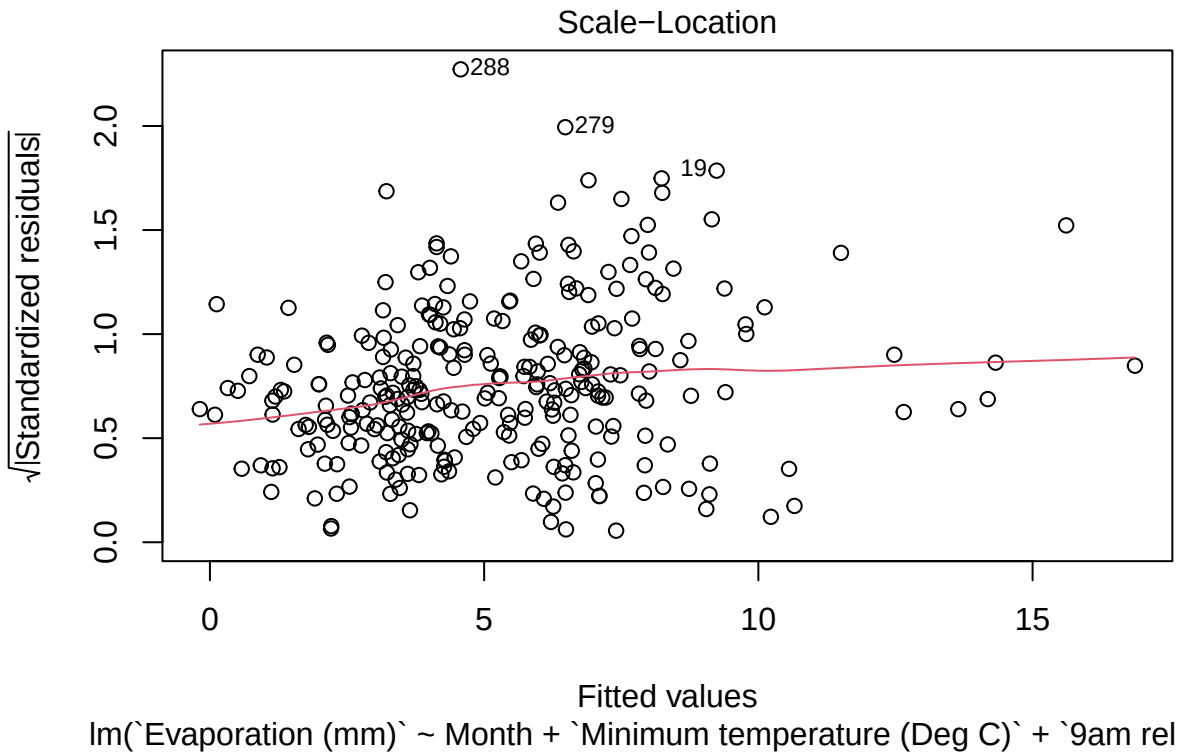
```
## Analysis of Variance Table
##
## Response: Evaporation (mm)
##                                   Df  Sum Sq Mean Sq  F value    Pr(>F)
## Month                             11 1145.08  104.10  23.3628 < 2.2e-16 ***
## `Minimum temperature (Deg C)`      1  512.32  512.32 114.9800 < 2.2e-16 ***
## `9am relative humidity (%)`        1  336.92  336.92  75.6141 3.578e-16 ***
## Month:`9am relative humidity (%)` 11  143.93   13.08   2.9366  0.001102 **
## Residuals                        267 1189.68    4.46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Diagnostics

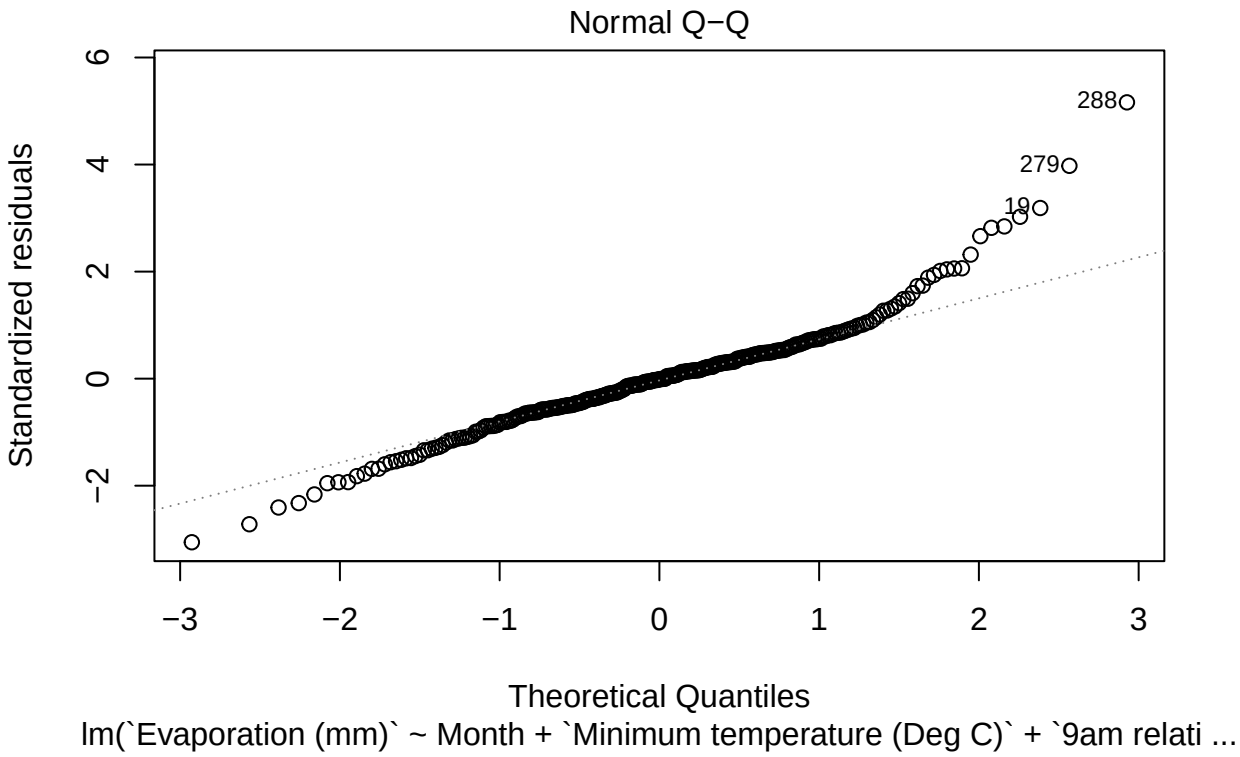
```
# check linear assusmptions.
# Linearity
plot(model, which=1)
```




```
# Homoscedasticity
plot(model, which=3)
```




```
# Normality
plot(model, which=2)
```



Prediction

```
# create a data-set for new data.
new_data= tibble(
  `Minimum temperature (Deg C)` = c(13.8,16.4,26.5,6.8),
  `9am relative humidity (%)` = c(74,57,35,76),
  Month = c("Feb","Dec","Jan","Jul")
)
print("Prediction Table:")
```

```
## [1] "Prediction Table:"
```

```
# Make predictions based on predictors.
tibble(Month = new_data$Month,
       `Min Temp(C)` = new_data$`Minimum temperature (Deg C)`,
       `Humidity at 9 am(%)` = new_data$`9am relative humidity (%)`,
       `Evaporation(mm)` =
         predict(model, new_data)) %>%
  knitr::kable(digits = 0, format.args = list(big.mark = ","))
```

Month	Min Temp(C)	Humidity at 9 am(%)	Evaporation(mm)
Feb	14	74	6
Dec	16	57	9
Jan	26	35	15
Jul	7	76	2

```
print("Prediction Table with 95% Confidence:")
```

```
## [1] "Prediction Table with 95% Confidence:"
```

```
# Add 95% confidence interval to make predictions.
# predict method is having by default level=0.95.
conf <- predict(model, new_data, interval = "confidence")
conf <- tibble(`Month` = new_data$Month,
               `Min Temp(C)` = new_data$`Minimum temperature (Deg C)`,
               `Humidity at 9 am(%)` = new_data$`9am relative humidity (%)`,
              `Lower bound for Evaporation` = conf[,2],
              `Expected Evaporation` = conf[,1],
              `Upper bound for Evaporationr` = conf[,3] )
conf %>% knitr::kable(digits = 1, format.args = list(big.mark = ","))
```

Month	Min Temp(C)	Humidity at 9 am(%)	Lower bound for Evaporation	Expected Evaporation	Upper bound for Evaporationr
Feb	13.8	74	4.7	5.8	7.0
Dec	16.4	57	7.4	8.5	9.6
Jan	26.5	35	11.6	14.7	17.7
Jul	6.8	76	1.5	2.4	3.3