CASE STUDY

Mark Werner

November, 2021

Contents

EXECUTIVE SUMMARY

In light of recent changes to Melbourne's climate, the Melbourne Water Corporation (MWC) has requested re-evaluation of previous estimates relating to water evaporation rates at its reservoirs and have asked for a new model to be built assessing the effects of Melbourne's day-to-day weather on evaporation.

A predictive model was built using Melbourne's weather observations, including evaporation, for the previous financial year. The findings of this study suggest a relationship between temporal and meteorological factors which have been found to have a significant impact on the amount of evaporation.

The model is able to predict , on a given day, the amount of evaporation based on three predictors: the month of the year, the percentage of relative humidity (measured at 9am on the given day), and the minimum temperature in degrees Celsius (i.e., the lowest temperature recorded in the preceding 24 hours to 9am of the given day).

Key Findings:

- There is a relationship between evaporation and month of the year where colder months correspond to lower evaporation and warmer months correspond to higher evaporation.
- A higher daily minimum temperature corresponds to a higher evaporation amount.
- The amount of evaporation decreases as the relative humidity increases.
- The is no relationship between evaporation amount and day of week
- There is a relationship between maximum daily temperature and evaporation amount however the variable for maximum temperature failed to pass the statistically significant threshold in the model building process and was therefore excluded from the final model. This is outlined in detail in the 'Discussion' section.

Note: all code used to create the model, as well as accompanying output and visualisations can be found in the 'Appendix' section.

METHODS

Cleaning

Data used to create the model was taken from the provided melbourne.csv file. This file contains 365 observations of 21 variables. A complete description of these variables can be found at

http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml.

The decision had been made that the following variables would be potentially useful influences on the amount of evaporation in a day:

- Month,
- Day of Week,
- Maximum temperature in degrees Celsius,
- Minimum temperature in degrees Celsius, and
- Relative humidity (as measured at 9am).

Once melbourne.csv was read using the R-Studio software package the above variables were selected (forming a new tibble called melbdata). These variables were recoded to allow for data extraction and further processing. In particular, the Date variable required recoding to be able to extract and create two new variables Month and Day of week. Once this was done, the Date variable was removed as it was superfluous. The variable 9am relative humidity (%) was recoded as an integer to assist with further processing. Each variable was checked for missing values. The response variable Evaporation (mm) was found to have 8 missing values, (*NA*'s). The corresponding 8 rows of data were omitted from the dataset.

The following dates corresponding to these omisssions were:

- 03/07/2018 to 08/07/2018 inclusive (i.e., 6 days in a row), and
- 27/12/2018 to 28/12/2018 inclusive (i.e., 2 days in a row).

Once these were removed, the data was checked again. The output of this can be found below in the section 'Summary Inspection'.

Univariate Analysis

Each variable underwent univariate analysis by producing appropriate plots as well as summaries to ascertain each variable's distribution (paying attention to location, shape, spread, and outliers). In each case skewness was also computed to help support the decision as to whether a transformation of a respective variable was required. Two variables, Evaporation (mm), the response variable, and Maximum Temperature (Deg C), a predictor variable, were transformed and renamed evap_transf and maxtemp_transf respectively.

As a final step, univariate analysis of the categorical variables as well as a summary of all variables within melbdata was done using the inspectdf function.

The following pages outline this process and provide full details of summaries in each case:

Univariate Summary: Evaporation

Univariate analysis of Evaporation (mm):

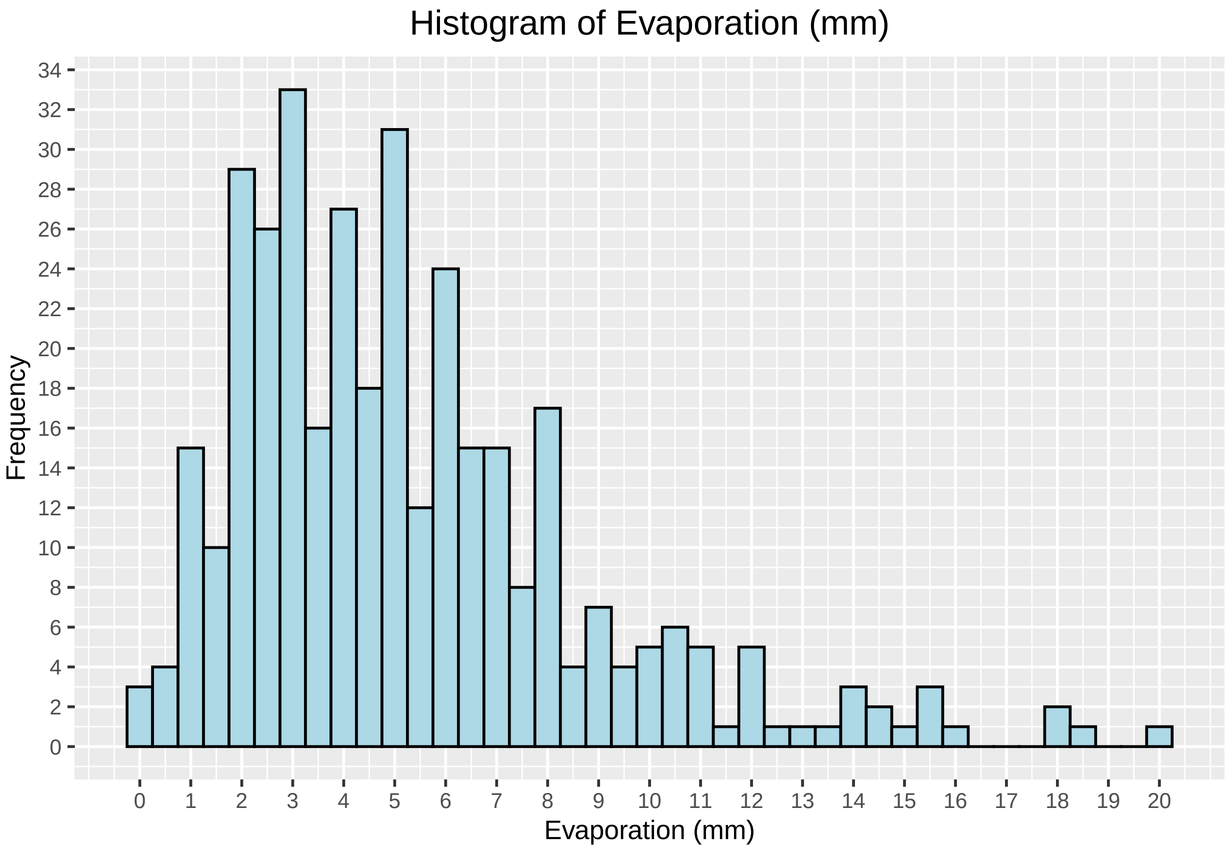


Figure 1: Plot of variable: Evaporation (mm)

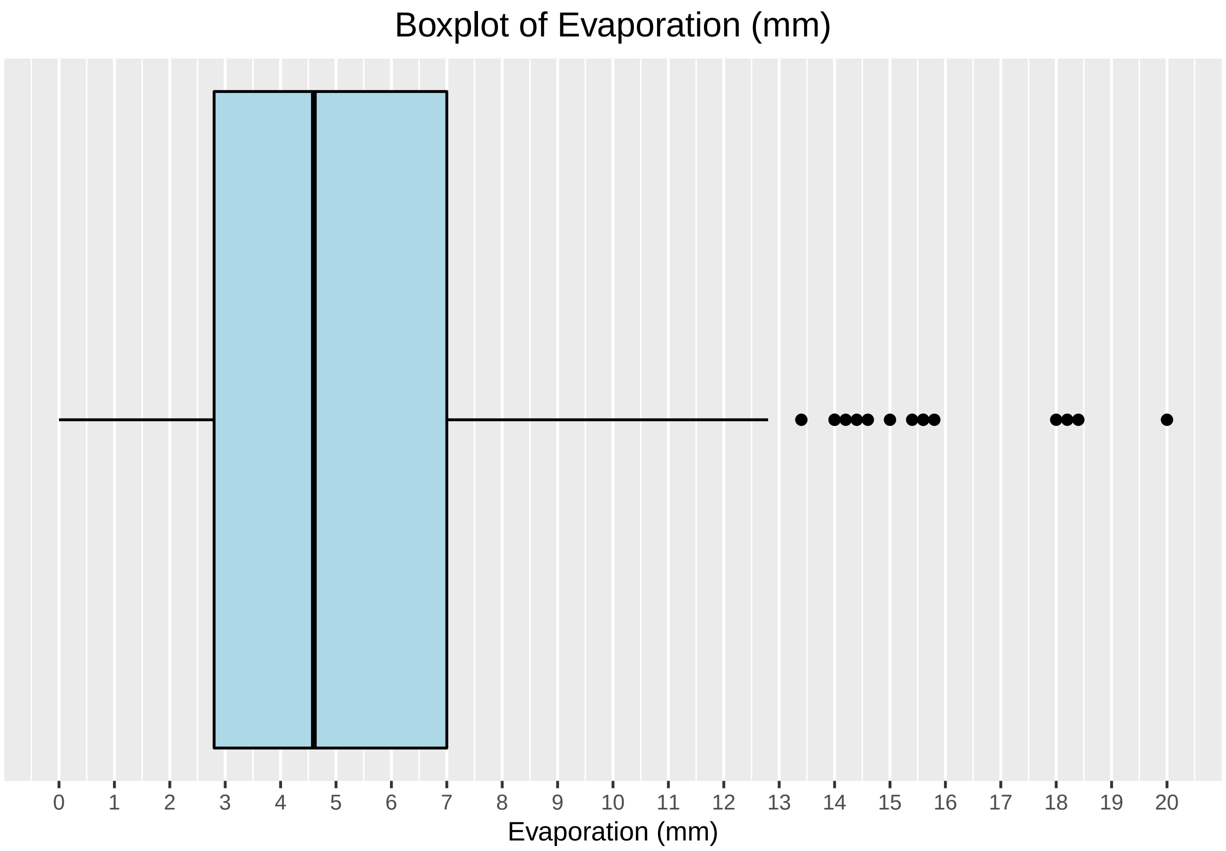


Figure 2: Plot of variable: Evaporation (mm)

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.800   4.600   5.312   7.000  20.000
```

```
## [1] 1.328865
```

```
## [1] 3.491281
```

The mean is 5.312 The median is 4.6 The shape is multimodal, right-skewed. Outliers can be seen beyond 13mm evaporation when observing the boxplot.

Assessing the spread, the standard deviation is 3.491281. The inter quartile range $IQR = Q_3 - Q_1 = 7 - 2.8 = 4.2$

The skewness of the distribution for this variable is 1.328865 whichi indicates it is highly right skewed.

As such, **this distribution will be transformed using a cube root transformation:**

Univariate Transformation: Evaporation

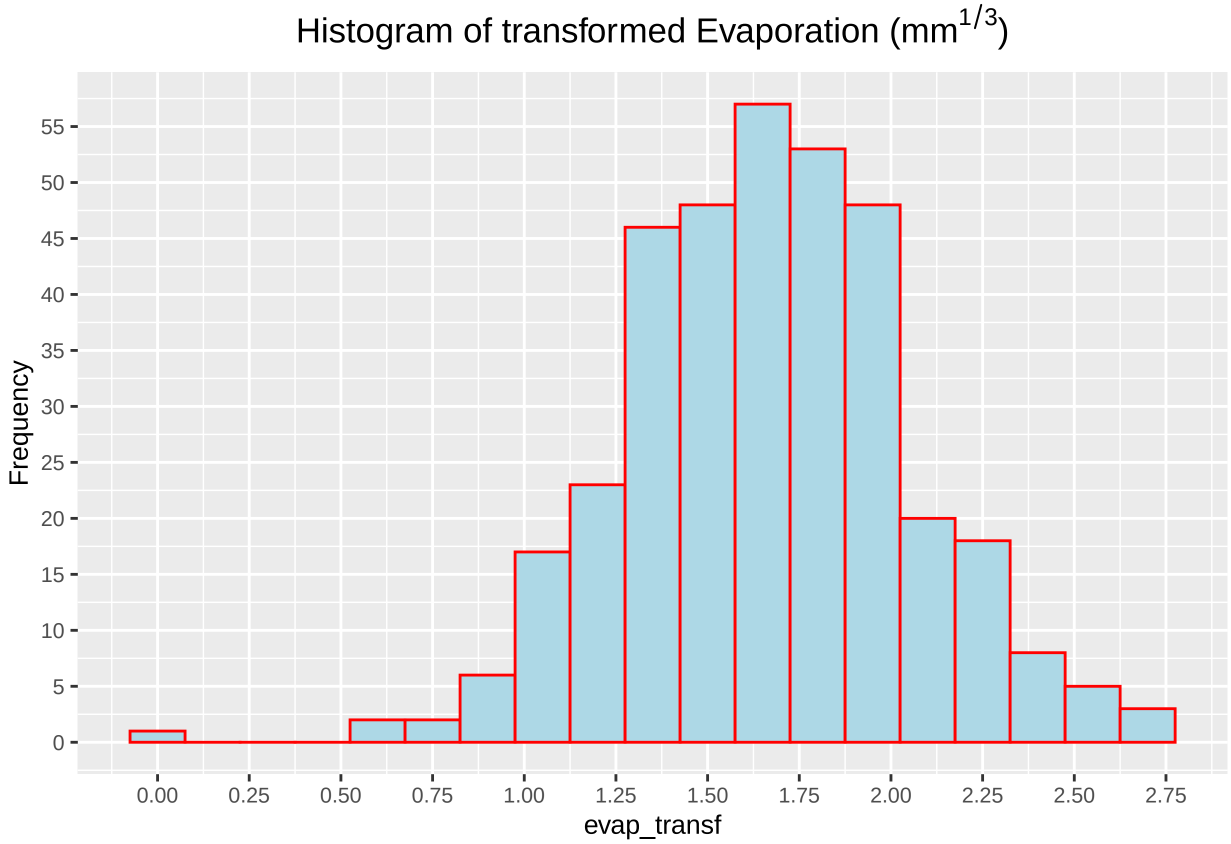


Figure 3: Plot of transformed variable Evaporation (mm)

```
## [1] -0.1130132
```

Testing the skewness of the transformed variable gives us -0.113 which indicates we have an acceptably symmetric distribution. **We will therefore use this transformed variable for our model.**

Univariate Summary: Minimum Temperature

Next we assess the variable “Minimum temperature (Deg C)”

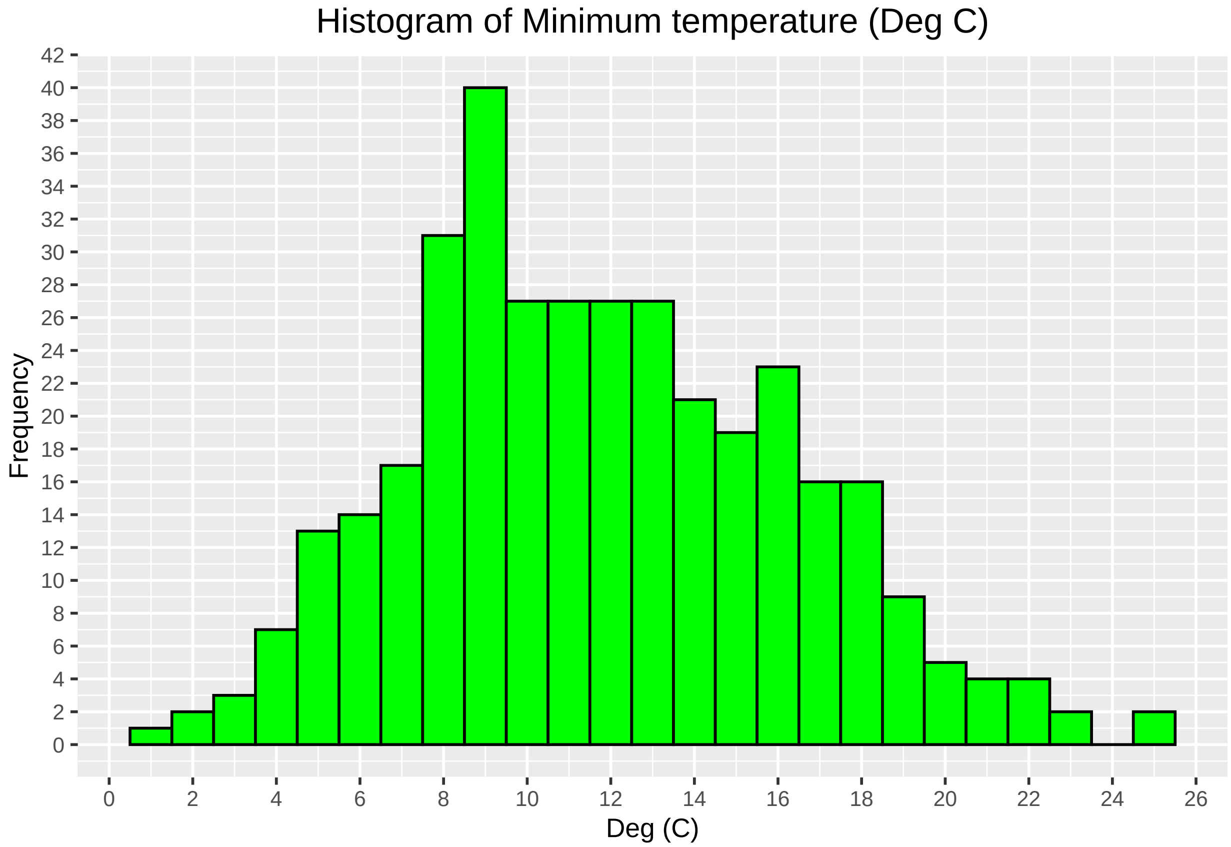


Figure 4: Plot of variable Minimum temperature (Deg C)

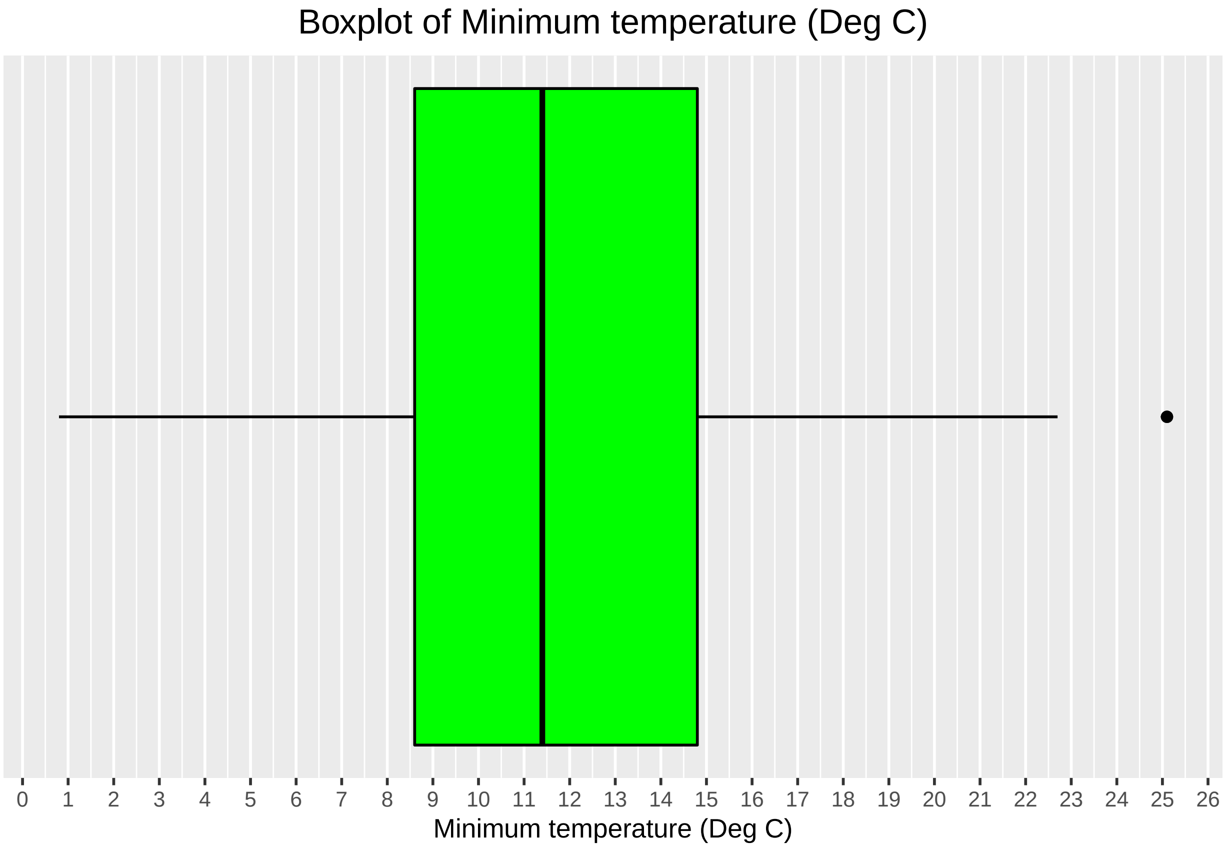


Figure 5: Plot of variable Minimum temperature (Deg C)

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.80    8.60   11.40   11.83   14.80   25.10
```

```
## [1] 0.3082482
```

```
## [1] 4.518612
```

The mean is 11.83 The median is 11.4 The shape is slightly right skewed. Outliers are visible in the boxplot beyond 25 degrees Celsius.

Assessing the spread, the standard deviation is 4.518612.
The inter quartile range is $IQR = Q_3 - Q_1 = 14.8 - 8.6 = 6.2$

The skewness of the distribution for this variable is 0.3082482 which indicates an acceptable level of symmetry. As such we will not perform any transformation on this variable.

Univariate Summary: Maximum Temperature

Next we assess the variable Maximum Temperature (Deg C):

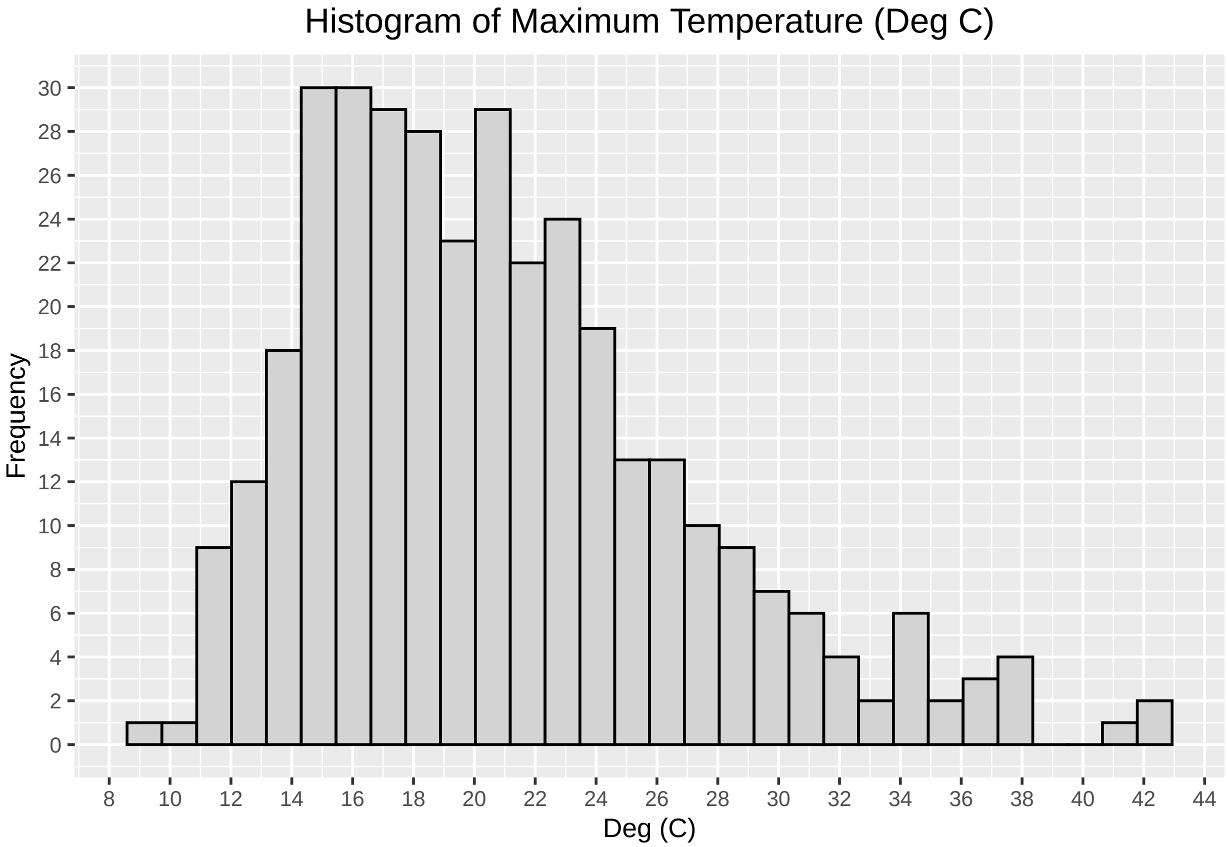


Figure 6: Plot of Maximum Temperature (Deg C)

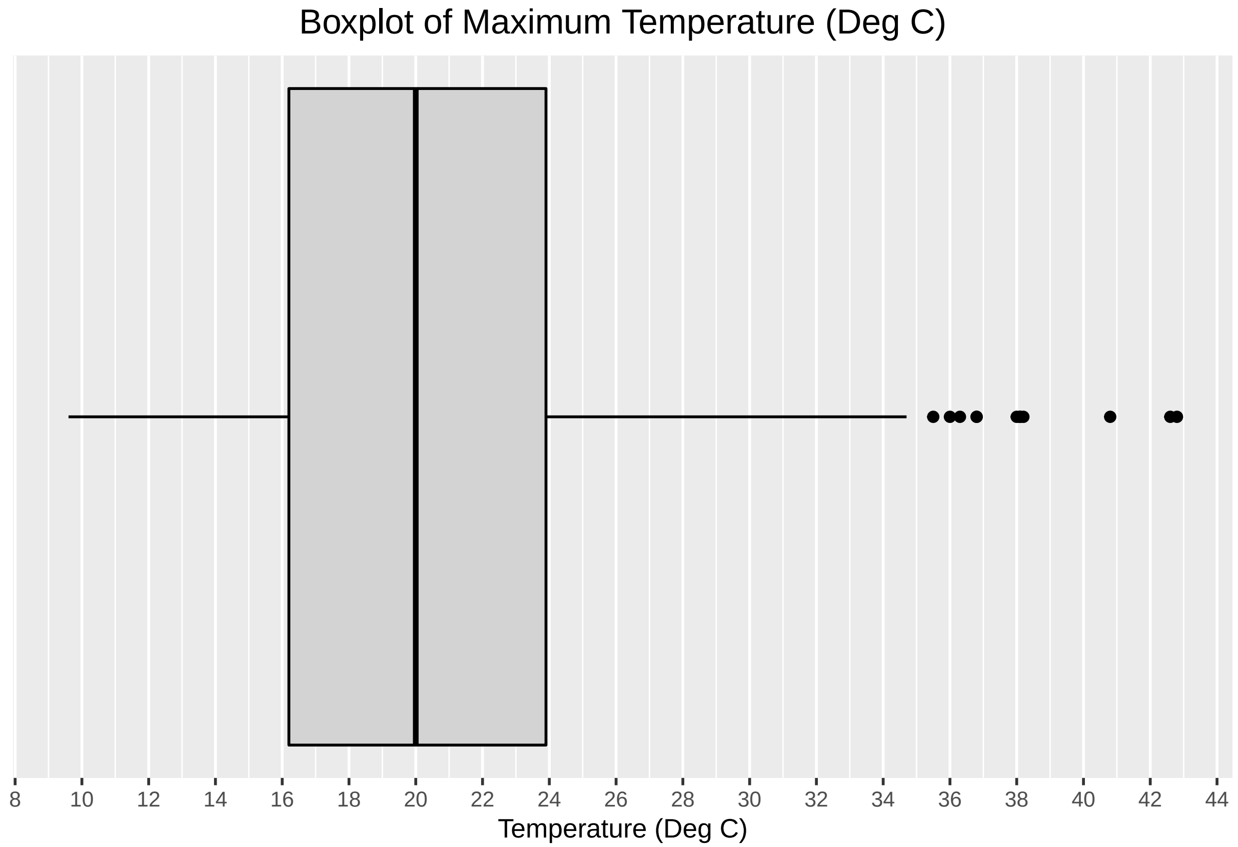


Figure 7: Plot of Maximum Temperature (Deg C)

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    9.60   16.20   20.00   20.87   23.90   42.80 
```

```
## [1] 0.9236407
```

```
## [1] 6.224088
```

The mean is 20.87 The median is 20 The shape is moderately right skewed with a skewness value of 0.9236407. Numerous outliers can be seen in the boxplot beyond 35 degrees Celsius.

Assessing the spread, the standard deviation is 6.224088. The inter quartile range is $IQR = Q_3 - Q_1 = 23.9 - 16.2 = 7.7$

This distribution is moderately skewed (= 0.9236407) so will perform a cube root transform (next page):

Univariate Transformation: Maximum Temperature

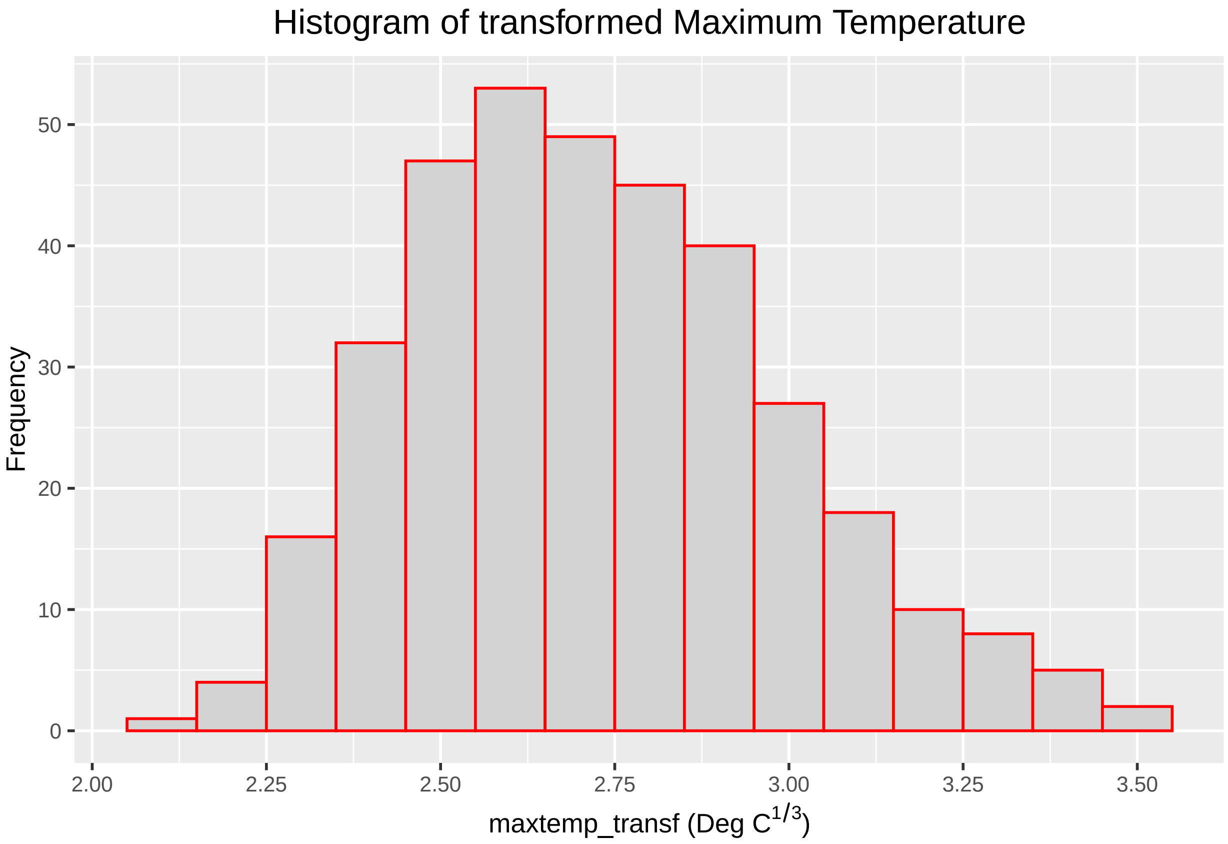


Figure 8: Plot of transformed variable Maximum Temperature.

```
## [1] 0.4437796
```

Testing the skewness of the transformed variable gives us 0.4437796 which indicates we have an acceptably symmetric distribution. We will therefore use this transformed variable for our model.

Univariate Summary: Relative Humidity

Next we assess the variable `9am relative humidity (%)`:

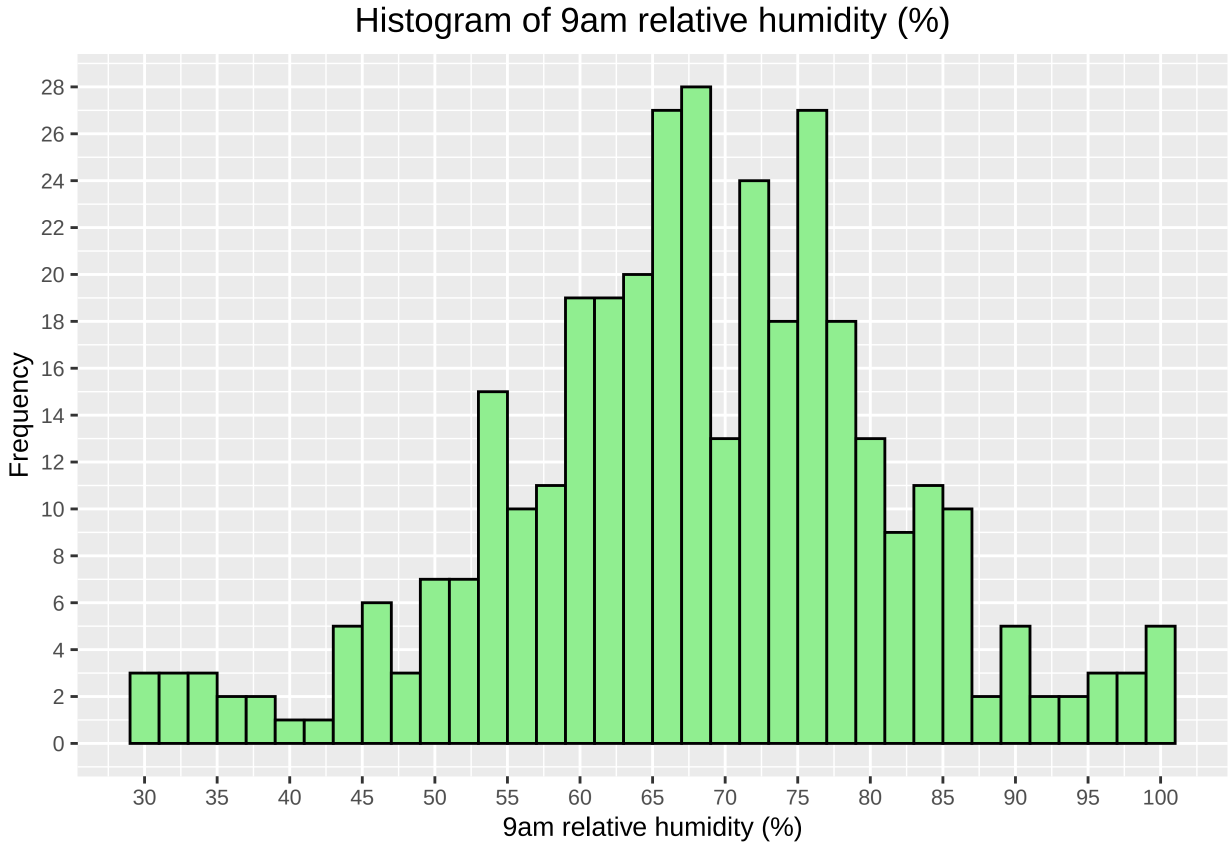


Figure 9: Plot of variable 9am relative humidity (%)

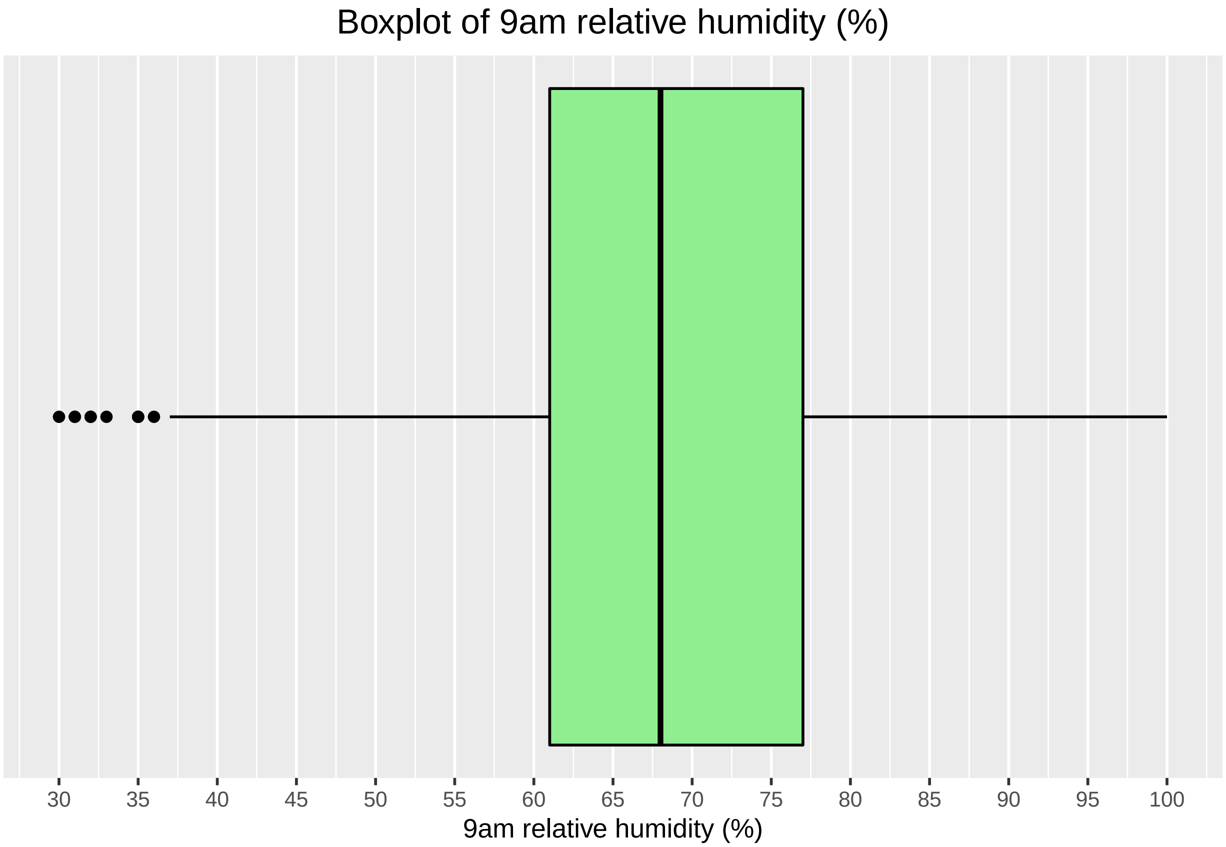


Figure 10: Plot of variable 9am relative humidity (%)

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30.00   61.00   68.00   68.36   77.00  100.00
```

```
## [1] -0.2736669
```

```
## [1] 13.58471
```

The mean is 68.36 The median is 68 The shape is multimodal. Outliers are visible in the boxplot beyond the lower whisker.

Assessing the spread, the standard deviation is 13.58471.
The inter quartile range is $IQR = Q_3 - Q_1 = 77 - 61 = 16$

The skewness of the distribution for this variable is -0.2736669 which indicates an acceptable level of symmetry. As such we will not perform any transformation on this variable.

Summary Inspection

A summary inspection of all the variables:

```
## # A tibble: 3 x 4
##   type             cnt  pcnt col_name
##   <chr>          <int> <dbl> <named list>
## 1 numeric            5  62.5 <chr [5]>
## 2 ordered factor     2  25   <chr [2]>
## 3 integer            1  12.5 <chr [1]>
```

```
## # A tibble: 6 x 10
##   col_name          min    q1 median  mean    q3    max     sd pcnt_na hist
##   <chr>           <dbl> <dbl>  <dbl> <dbl> <dbl>  <dbl>  <dbl>   <dbl> <named l>
## 1 Minimum temper~  0.8   8.6   11.4  11.8  14.8   25.1   4.52        0 <tibble ~
## 2 Maximum Temper~  9.6  16.2   20    20.9  23.9   42.8   6.22        0 <tibble ~
## 3 9am relative h~ 30    61     68    68.4  77    100    13.6         0 <tibble ~
## 4 Evaporation (m~  0     2.8    4.6   5.31  7     20     3.49        0 <tibble ~
## 5 evap_transf      0     1.41   1.66  1.66  1.91   2.71  0.384       0 <tibble ~
## 6 maxtemp_transf   2.13  2.53   2.71  2.73  2.88   3.50  0.263       0 <tibble ~
```

```
## # A tibble: 2 x 5
##   col_name      cnt common common_pcnt levels
##   <chr>       <int> <chr>        <dbl> <named list>
## 1 Day of Week     7 Monday       14.6  <tibble [7 x 3]>
## 2 Month          12 Aug           8.68 <tibble [12 x 3]>
```

Bivariate Analysis

Once the univariate analysis was complete and the appropriate transformations were made, plots were produced to explore the relationships between the transformed response variable `evap_transf` and each potential predictor.

Summary of Bivariate Analysis:

- `evap_transf` & `Month` Side-by-side boxplot (*Refer to Figure 11*)
- `evap_transf` & `Day of Week` Side-by-side boxplot (*Refer to Figure 12*)
- `evap_transf` & `9am relative humidity (%)` Scatterplot (*Refer to Figure 13*)
- `evap_transf` & `Minimum temperature (Deg C)` Scatterplot (*Refer to Figure 14*)
- `evap_transf` & `maxtemp_transf` Scatterplot (*Refer to Figure 15*)

Detailed interpretations of these are outlined alongside each plot.

Producing plots to explore the relationship between the transformed response variable and each predictor:

Bivariate Summary: Month

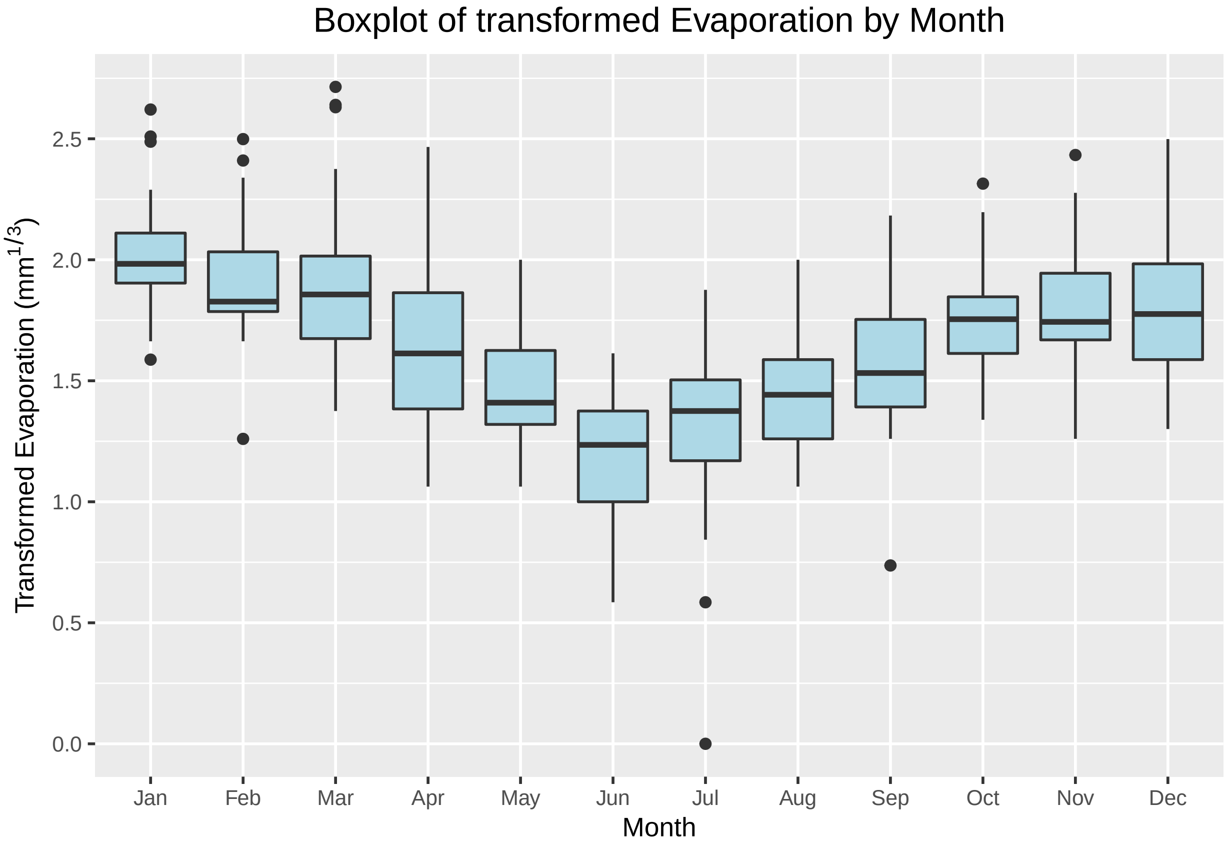


Figure 11: Side-by-side Box plot of Evaporation by Month

Here we see a definite cyclical pattern where the amount of evaporation decreases over the colder months and increases during the hotter months. This indicates there exists a relationship between the predictor variable `Month` and the response variable 'evap_transf' (in units of $mm^{1/3}$).

Bivariate Summary: Day of Week

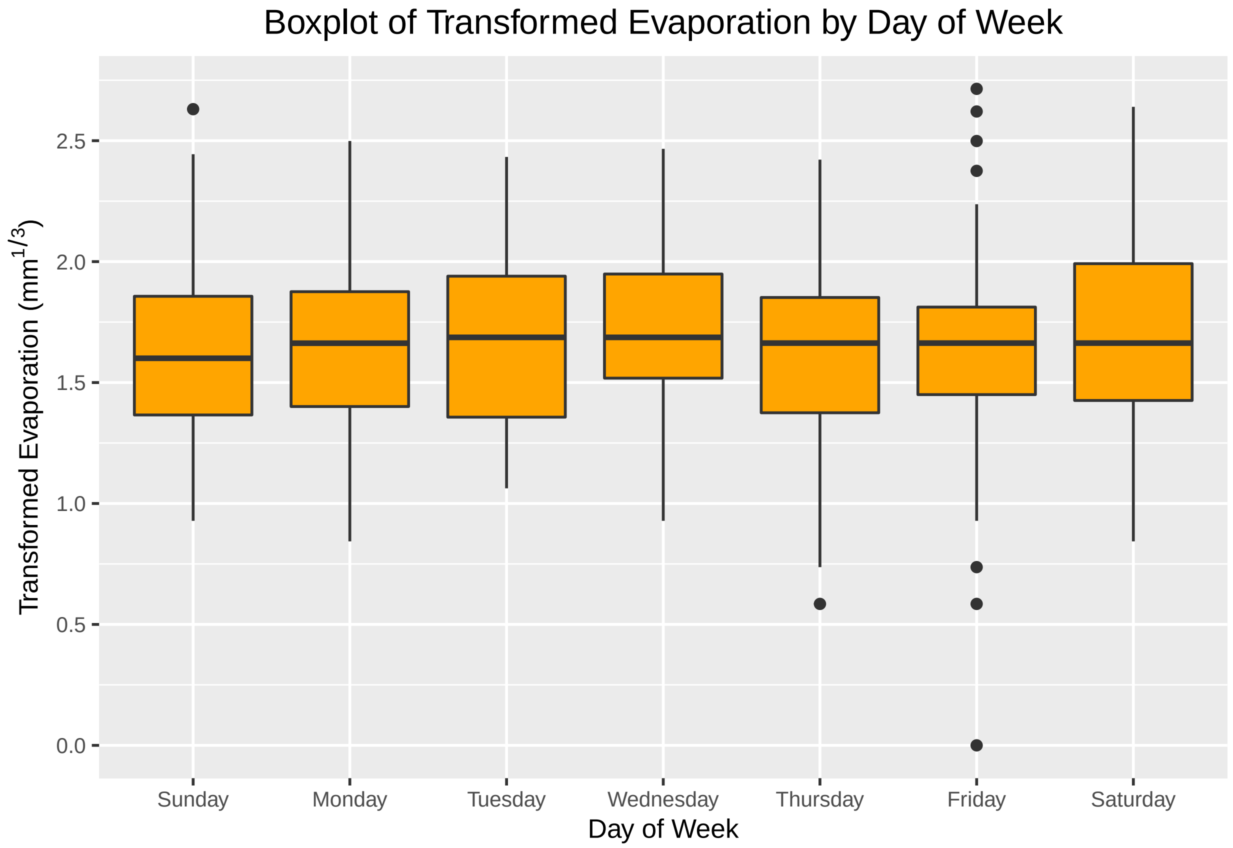


Figure 12: Side-by-side Box plot of Evaporation by Day of Week

From observation of the boxplot we see no discernible relationship between the predictor variable `Day of Week` and the transformed response variable `evap_transf` (expressed in units of $mm^{1/3}$) where the median value for each day sits between $1.5mm^{1/3}$ and $1.75mm^{1/3}$.

Bivariate Summary: Minimum Temperature

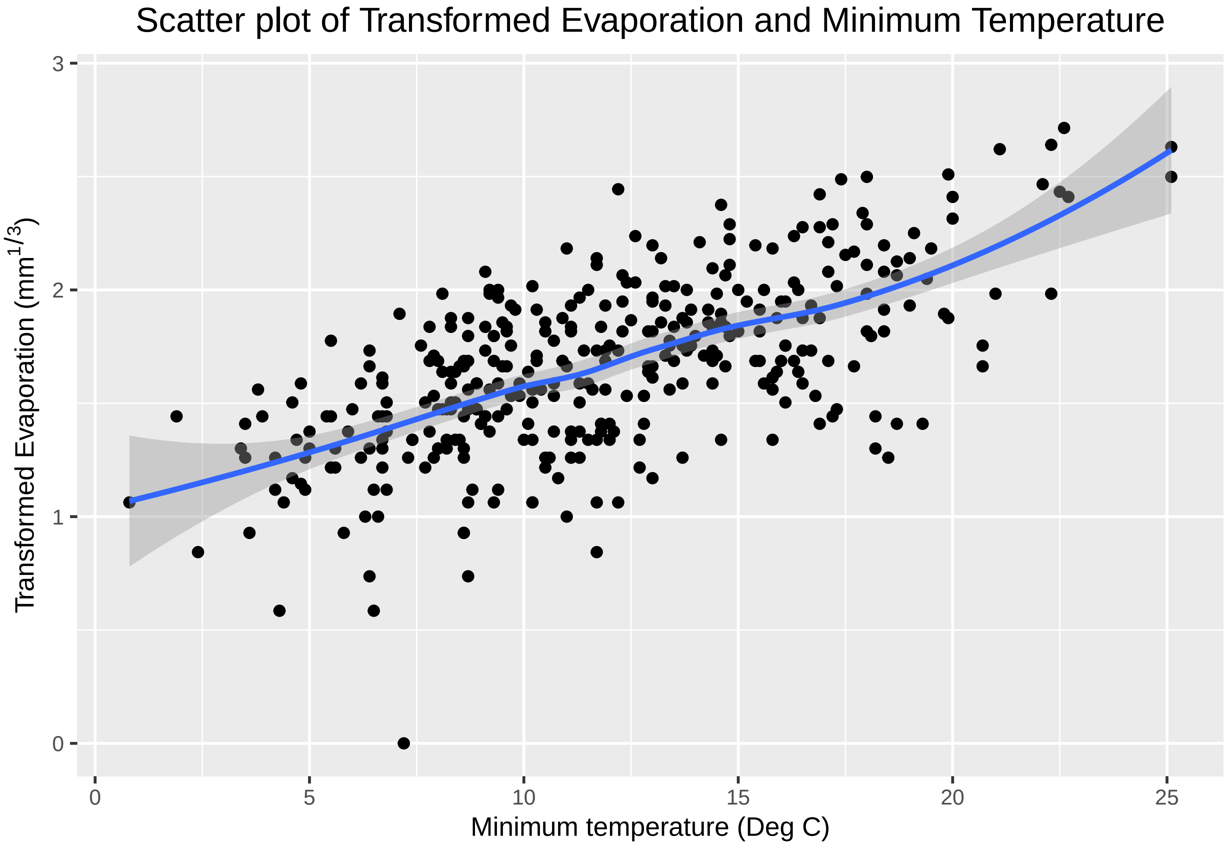


Figure 13: Scatter plot of transformed Evaporation and Minimum temperature (Deg C)

```
## [1] 0.6412193
```

Here we see a moderate positive relationship between the transformed response variable `evap_transf` (expressed in units of $mm^{1/3}$) and the predictor variable `Minimum temperature (Deg C)`. We can see from the output that this moderate positive relationship carries a correlation coefficient value of 0.6412193.

Bivariate Summary: Maxiumum Temperature

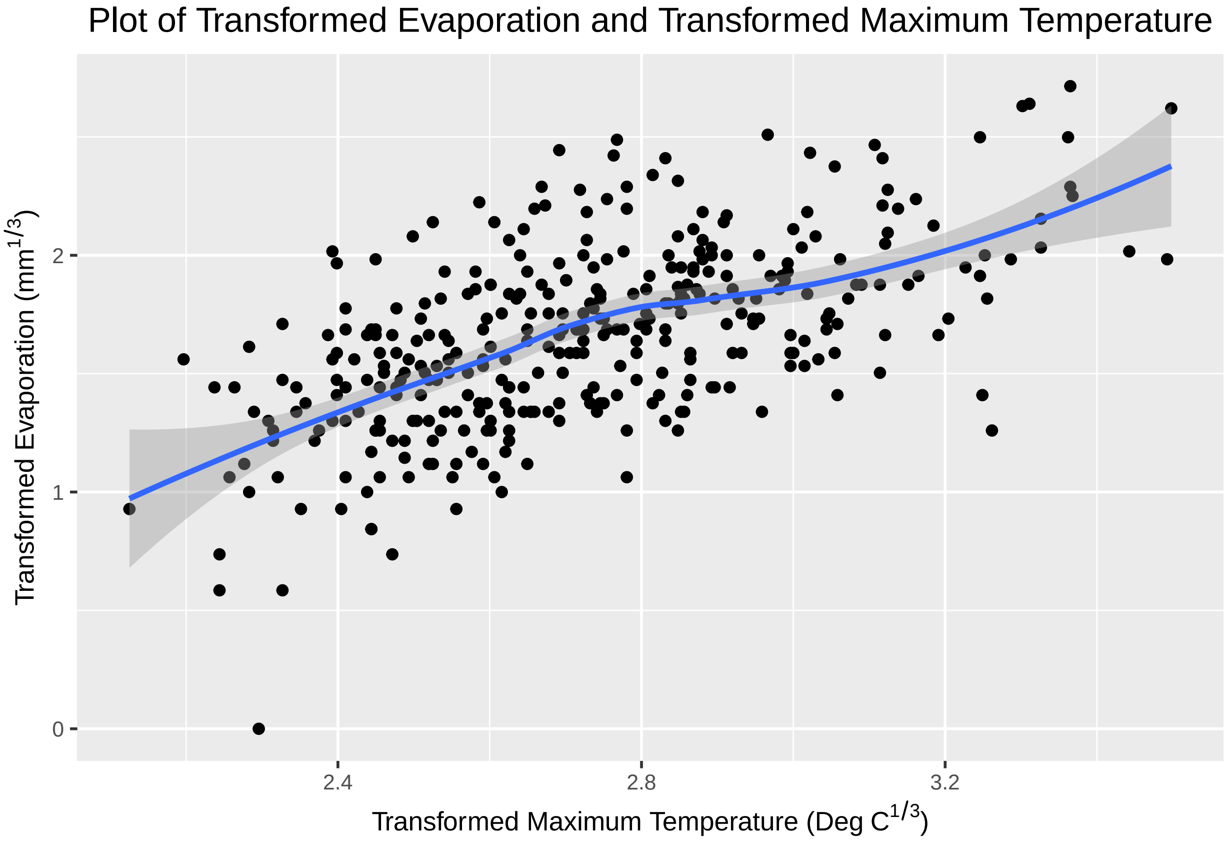


Figure 14: Scatter plot of transformed Evaporation and transformed Maximum Temperature (Deg C)

```
## [1] 0.5957085
```

Here we see a moderate positive relationship between the transformed response variable `evap_transf` (expressed in units of $mm^{1/3}$) and the transformed predictor variable `maxtemp_transf` (expressed in units of Deg $C^{1/3}$). We can see from the output that this moderate positive relationship carries a correlation coefficient value of 0.5957085.

Bivariate Summary: Relative Humidity

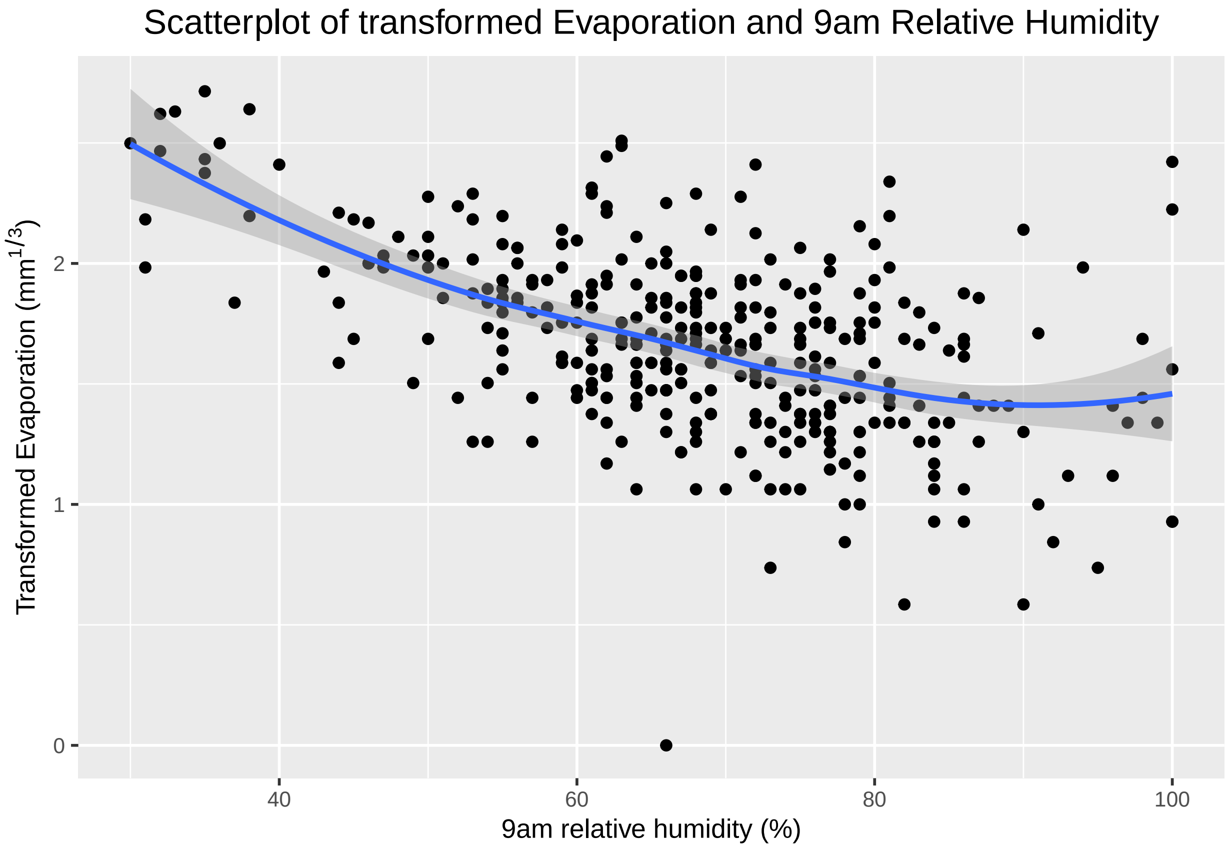


Figure 15: Scatterplot of transformed Evaporation and 9am Relative Humidity

```
## [1] -0.5168967
```

Here we see a moderate negative relationship between the transformed response variable `evap_transf` (expressed in units of $mm^{1/3}$) and the predictor variable `9am relative humidity (%)`. We can see from the output that this moderate negative relationship carries a correlation coefficient value of -0.5168967.

Trivariate Analysis

This is added as an extension to assist the reader by providing enhanced visualisation to better understand the intrinsic relationship between respective quantitative variables and month of the year. In each case we see a clear relationship between the amount of evaporation and time of year (*Refer to Figures 16, 17, and 18 over the following pages*).

Visualisations

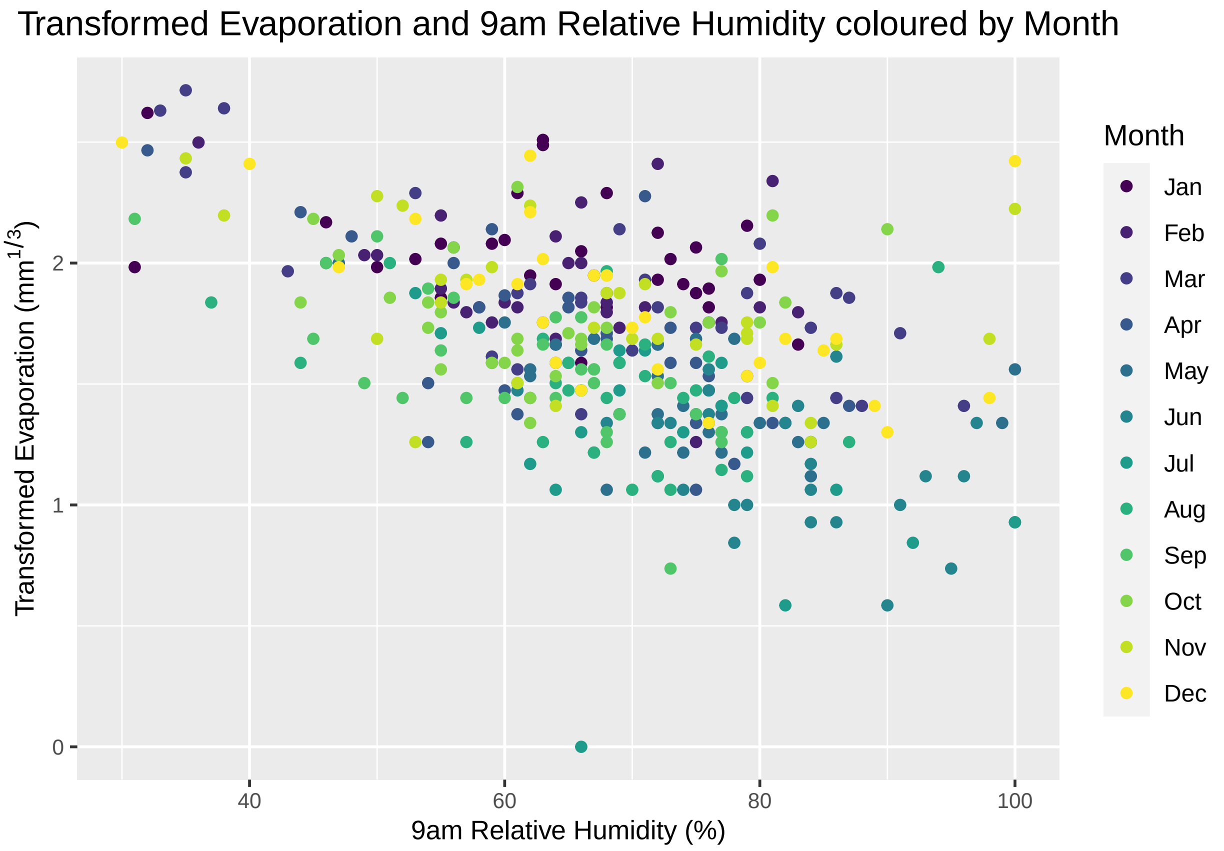


Figure 16: Scatter plot of transformed Evaporation and 9am Relative Humidity coloured by Month.

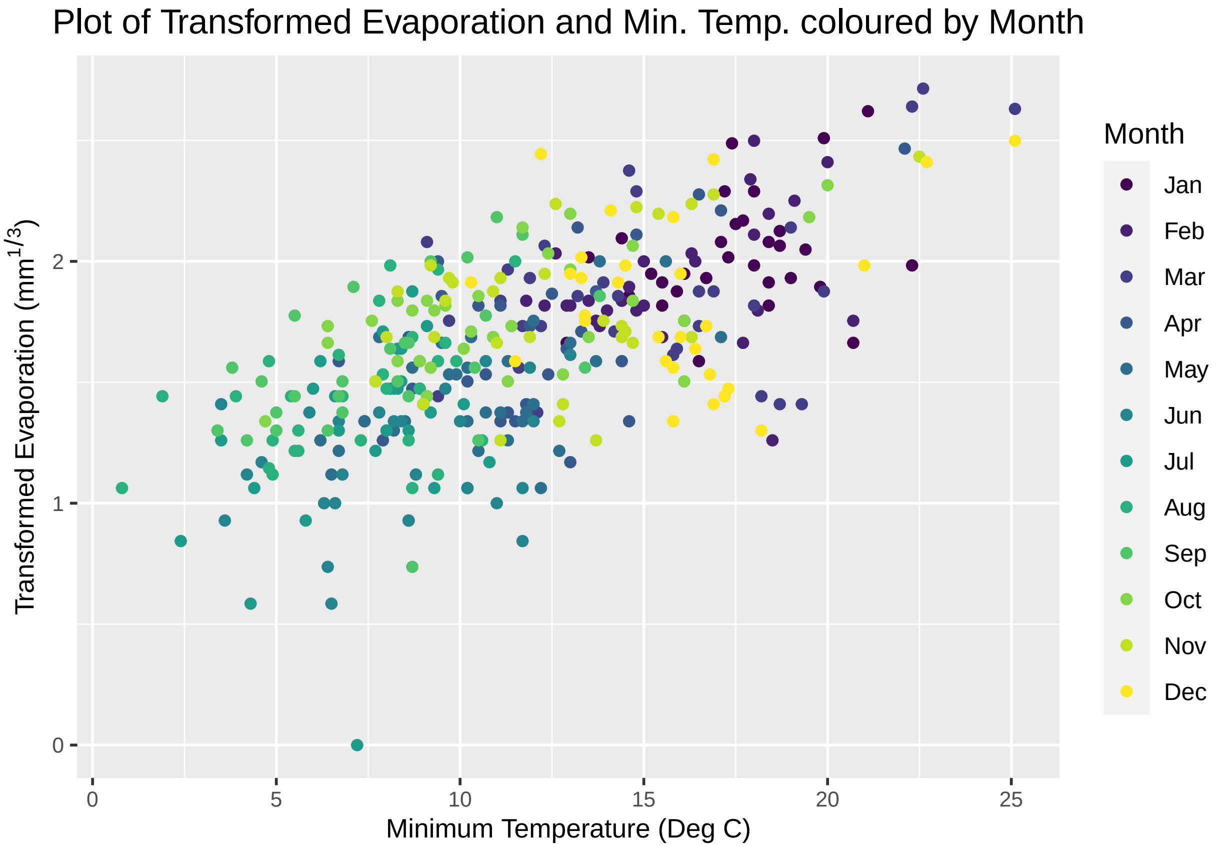


Figure 17: Scatter plot of transformed Evaporation and Minimum temperature (Deg C) coloured by Month.

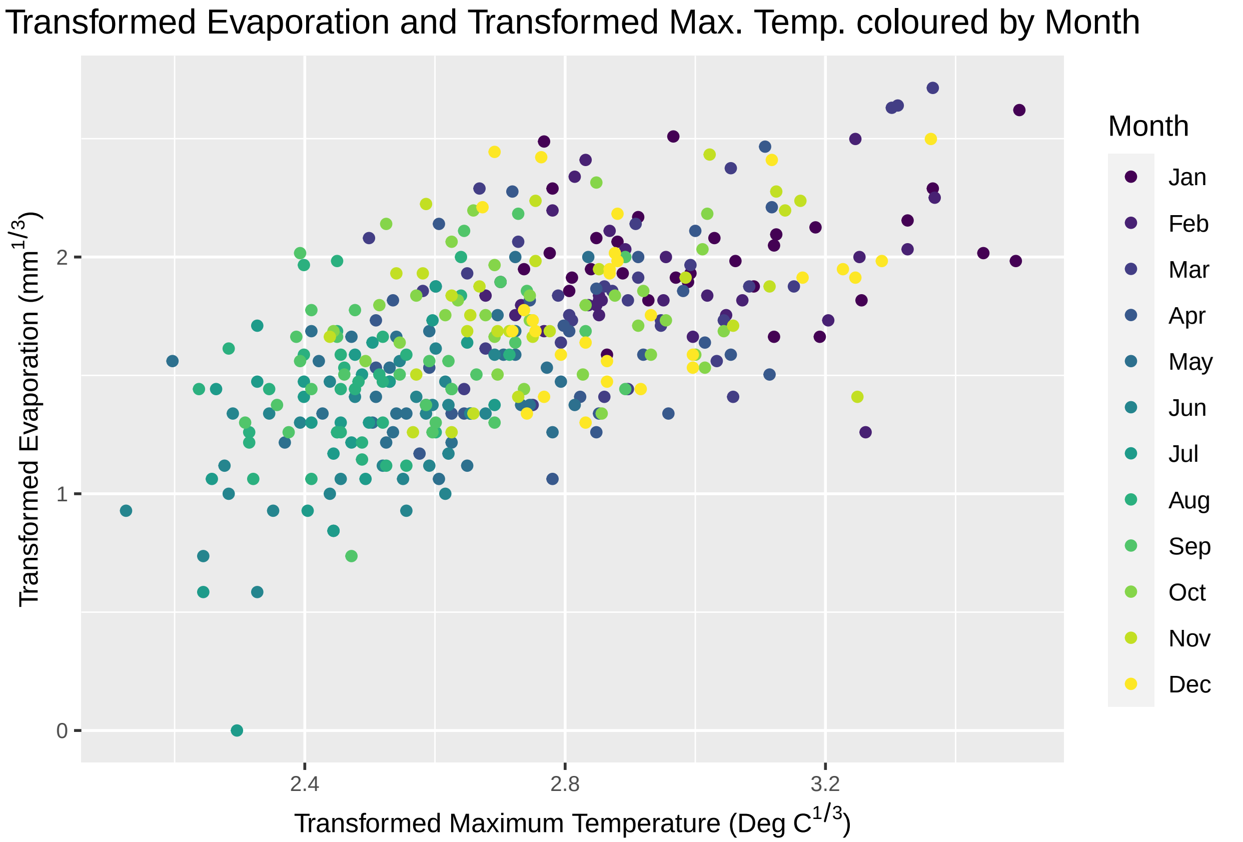


Figure 18: Scatter plot of transformed Evaporation and transformed Maximum Temperature coloured by Month.

Model Selection

A base linear model was built in R to predict evaporation (in mm) on a given day in Melbourne using all of the predictors listed on page 4. An interaction term between `Month` and `9am relative humidity (%)` was also included in this model to determine whether humidity has a different effect in different months.

Using the five predictors we built a model of the following form (with the inclusion of the interaction term):

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i1} x_{i2} + \epsilon_i$$

The categorical variables `Month` and `Day of Week` were recoded as unordered factors to assist with their interpretation in the summary output of the model before undertaking a dimensionality reduction process. This iterative process involved fitting a linear model containing all the possible predictors, then determining the p-values for inclusion of each predictor. P-values for quantitative variables were determined using the linear model summary. P-values for categorical values (as well as the interaction term which contained a categorical variable) were determined using an ANOVA. The predictor with the highest p-value was removed (unless all remaining predictors were significant at the 5% level). The model was then updated to include only the remaining predictors and the process was repeated until only significant predictors remained.

The First Iteration

```
##
## Call:
## lm(formula = evap_transf ~ Month + ‘9am relative humidity (%)‘ +
##     ‘Day of Week‘ + ‘Minimum temperature (Deg C)‘ + maxtemp_transf +
##     Month:‘9am relative humidity (%)‘, data = melbdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.28691 -0.14224  0.02032  0.13069  0.95086
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          1.8132492  0.3410269   5.317 1.97e-07 ***
## MonthFeb                             0.1529285  0.3697612   0.414   0.6794
## MonthMar                             0.3603272  0.2911988   1.237   0.2168
## MonthApr                             0.2746630  0.3435437   0.799   0.4246
## MonthMay                            -0.1641296  0.3699075  -0.444   0.6576
## MonthJun                            -0.1216493  0.4391415  -0.277   0.7819
## MonthJul                            -0.0103887  0.3957198  -0.026   0.9791
## MonthAug                            -0.3723295  0.3559141  -1.046   0.2963
## MonthSep                             0.2353231  0.3492938   0.674   0.5010
## MonthOct                            -0.5295953  0.3447521  -1.536   0.1255
## MonthNov                            -0.0561845  0.3083123  -0.182   0.8555
## MonthDec                             0.1255404  0.3094386   0.406   0.6852
## ‘9am relative humidity (%)‘         -0.0075528  0.0035997  -2.098   0.0367 *
## ‘Day of Week‘Monday                 -0.0063929  0.0478733  -0.134   0.8939
## ‘Day of Week‘Tuesday                -0.0006349  0.0483030  -0.013   0.9895
## ‘Day of Week‘Wednesday               0.0085864  0.0482822   0.178   0.8590
## ‘Day of Week‘Thursday               -0.0659488  0.0482436  -1.367   0.1726
## ‘Day of Week‘Friday                 -0.0689326  0.0490681  -1.405   0.1610
## ‘Day of Week‘Saturday                0.0638668  0.0478728   1.334   0.1831
## ‘Minimum temperature (Deg C)‘         0.0289214  0.0049274   5.870 1.08e-08 ***
## maxtemp_transf                        0.0663854  0.0827591   0.802   0.4231    
## MonthFeb:‘9am relative humidity (%)‘ -0.0032041  0.0056420  -0.568   0.5705    
## MonthMar:‘9am relative humidity (%)‘ -0.0060684  0.0043800  -1.385   0.1669    
## MonthApr:‘9am relative humidity (%)‘ -0.0066876  0.0052157  -1.282   0.2007    
## MonthMay:‘9am relative humidity (%)‘ -0.0012971  0.0052903  -0.245   0.8065    
## MonthJun:‘9am relative humidity (%)‘ -0.0035085  0.0058404  -0.601   0.5484    
## MonthJul:‘9am relative humidity (%)‘ -0.0046843  0.0056807  -0.825   0.4102    
## MonthAug:‘9am relative humidity (%)‘  0.0025800  0.0052459   0.492   0.6232    
## MonthSep:‘9am relative humidity (%)‘ -0.0064009  0.0054522  -1.174   0.2413    
## MonthOct:‘9am relative humidity (%)‘  0.0074700  0.0052534   1.422   0.1560    
## MonthNov:‘9am relative humidity (%)‘  0.0003071  0.0046171   0.067   0.9470    
## MonthDec:‘9am relative humidity (%)‘ -0.0031906  0.0045813  -0.696   0.4866    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.2407 on 325 degrees of freedom
## Multiple R-squared:  0.641,  Adjusted R-squared:  0.6068
## F-statistic: 18.72 on 31 and 325 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Response: evap_transf
##                                    Df  Sum Sq Mean Sq  F value    Pr(>F)    
## Month                              11 21.8352  1.9850  34.2594 < 2.2e-16 ***
## ‘9am relative humidity (%)‘         1  7.3142  7.3142 126.2347 < 2.2e-16 ***
## ‘Day of Week‘                       6  0.7297  0.1216   2.0991   0.05299 .  
## ‘Minimum temperature (Deg C)‘       1  2.8753  2.8753  49.6239 1.117e-11 ***
## maxtemp_transf                      1  0.0378  0.0378   0.6523   0.41990    
## Month:‘9am relative humidity (%)‘  11  0.8357  0.0760   1.3112   0.21648    
## Residuals                         325 18.8308  0.0579                       
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

From the summary output we can see that the predictor with the highest p-value is maxtemp_transf = 0.4231. This was removed from the model and the process was repeated with the remaining predictors.

The Second Iteration

```
##
## Call:
## lm(formula = evap_transf ~ Month + ‘9am relative humidity (%)‘ +
##     ‘Day of Week‘ + ‘Minimum temperature (Deg C)‘ + Month:‘9am relative humidity (%)‘,
##     data = melbdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.29891 -0.14570  0.02408  0.13498  0.95273
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          1.9999362  0.2491324   8.028 1.82e-14 ***
## MonthFeb                             0.1487838  0.3695229   0.403   0.6875
## MonthMar                             0.3641490  0.2910006   1.251   0.2117
## MonthApr                             0.2812358  0.3432581   0.819   0.4132
## MonthMay                            -0.1808390  0.3691185  -0.490   0.6245
## MonthJun                            -0.1253151  0.4388775  -0.286   0.7754
## MonthJul                            -0.0479624  0.3927228  -0.122   0.9029
## MonthAug                            -0.3876643  0.3552059  -1.091   0.2759
## MonthSep                             0.2358028  0.3491022   0.675   0.4999
## MonthOct                            -0.5199331  0.3443531  -1.510   0.1320
## MonthNov                            -0.0588121  0.3081263  -0.191   0.8487
## MonthDec                             0.1448894  0.3083283   0.470   0.6387
## ‘9am relative humidity (%)‘         -0.0076487  0.0035958  -2.127   0.0342 *
## ‘Day of Week‘Monday                 -0.0070863  0.0478393  -0.148   0.8823
## ‘Day of Week‘Tuesday                -0.0023285  0.0482304  -0.048   0.9615
## ‘Day of Week‘Wednesday               0.0050482  0.0480540   0.105   0.9164
## ‘Day of Week‘Thursday               -0.0660807  0.0482169  -1.370   0.1715
## ‘Day of Week‘Friday                 -0.0670842  0.0489872  -1.369   0.1718
## ‘Day of Week‘Saturday                0.0622499  0.0478041   1.302   0.1938
## ‘Minimum temperature (Deg C)‘        0.0301695  0.0046727   6.457 3.89e-10 ***
## MonthFeb:‘9am relative humidity (%)‘ -0.0031949  0.0056389  -0.567   0.5714
## MonthMar:‘9am relative humidity (%)‘ -0.0062142  0.0043739  -1.421   0.1563
## MonthApr:‘9am relative humidity (%)‘ -0.0069258  0.0052044  -1.331   0.1842
## MonthMay:‘9am relative humidity (%)‘ -0.0013321  0.0052872  -0.252   0.8012
## MonthJun:‘9am relative humidity (%)‘ -0.0037552  0.0058291  -0.644   0.5199
## MonthJul:‘9am relative humidity (%)‘ -0.0045210  0.0056740  -0.797   0.4261
## MonthAug:‘9am relative humidity (%)‘  0.0024469  0.0052404   0.467   0.6409
## MonthSep:‘9am relative humidity (%)‘ -0.0067133  0.0054353  -1.235   0.2177
## MonthOct:‘9am relative humidity (%)‘  0.0071626  0.0052366   1.368   0.1723
## MonthNov:‘9am relative humidity (%)‘  0.0001931  0.0046124   0.042   0.9666
## MonthDec:‘9am relative humidity (%)‘ -0.0035489  0.0045570  -0.779   0.4367
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.2406 on 326 degrees of freedom
## Multiple R-squared:  0.6403, Adjusted R-squared:  0.6072
## F-statistic: 19.35 on 30 and 326 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
```

```
## Response: evap_transf
##                                    Df  Sum Sq Mean Sq  F value    Pr(>F)    
## Month                              11 21.8352  1.9850  34.2969 < 2.2e-16 ***
## ‘9am relative humidity (%)‘         1  7.3142  7.3142 126.3729 < 2.2e-16 ***
## ‘Day of Week‘                       6  0.7297  0.1216   2.1014   0.05272 .  
## ‘Minimum temperature (Deg C)‘       1  2.8753  2.8753  49.6783 1.085e-11 ***
## Month:‘9am relative humidity (%)‘  11  0.8362  0.0760   1.3135   0.21521    
## Residuals                         326 18.8681  0.0579                       
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

From the ANOVA output we can see that the interaction term `Month`:`9am relative humidity (%)` has the highest p-value = 0.21521. This interaction term was removed and the process was repeated using only the remaining predictors.

The Third Iteration

```
##
## Call:
## lm(formula = evap_transf ~ Month + ‘9am relative humidity (%)‘ +
##     ‘Day of Week‘ + ‘Minimum temperature (Deg C)‘, data = melbdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.28320 -0.14270  0.02131  0.14336  0.90367
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   2.106554   0.124156  16.967  < 2e-16 ***
## MonthFeb                     -0.052831   0.063513  -0.832 0.406103
## MonthMar                     -0.041575   0.062294  -0.667 0.504967
## MonthApr                     -0.161018   0.066792  -2.411 0.016455 *
## MonthMay                     -0.244112   0.068757  -3.550 0.000439 ***
## MonthJun                     -0.371970   0.074874  -4.968 1.08e-06 ***
## MonthJul                     -0.333899   0.079346  -4.208 3.31e-05 ***
## MonthAug                     -0.188046   0.076985  -2.443 0.015094 *
## MonthSep                     -0.169777   0.075506  -2.249 0.025188 *
## MonthOct                     -0.046124   0.067815  -0.680 0.496884
## MonthNov                     -0.028459   0.065840  -0.432 0.665842
## MonthDec                     -0.085103   0.063132  -1.348 0.178553
## ‘9am relative humidity (%)‘   -0.009952   0.001032  -9.642  < 2e-16 ***
## ‘Day of Week‘Monday          -0.007462   0.047659  -0.157 0.875676
## ‘Day of Week‘Tuesday          0.014088   0.048009   0.293 0.769366
## ‘Day of Week‘Wednesday        0.015062   0.047926   0.314 0.753506
## ‘Day of Week‘Thursday        -0.050059   0.047997  -1.043 0.297712
## ‘Day of Week‘Friday          -0.063396   0.048221  -1.315 0.189513
## ‘Day of Week‘Saturday         0.069005   0.047735   1.446 0.149223
## ‘Minimum temperature (Deg C)‘  0.032056   0.004571   7.012 1.29e-11 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.2418 on 337 degrees of freedom
## Multiple R-squared:  0.6244, Adjusted R-squared:  0.6032
## F-statistic: 29.48 on 19 and 337 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Response: evap_transf
##                               Df  Sum Sq Mean Sq  F value    Pr(>F)
## Month                         11 21.8352  1.9850  33.9495 < 2.2e-16 ***
## ‘9am relative humidity (%)‘     1  7.3142  7.3142 125.0930 < 2.2e-16 ***
## ‘Day of Week‘                  6  0.7297  0.1216   2.0801   0.05505 .
## ‘Minimum temperature (Deg C)‘   1  2.8753  2.8753  49.1751 1.287e-11 ***
## Residuals                    337 19.7043  0.0585
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The ANOVA output shows the categorical predictor `Day of Week` has a p-value lying outside the 5% significance level = 0.05505. This was removed and the process was repeated using only the remaining predictors.

The Fourth Iteration

```
## 
## Call:
## lm(formula = evap_transf ~ Month + ‘9am relative humidity (%)‘ + 
##     ‘Minimum temperature (Deg C)‘, data = melbdata)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.34284 -0.14329  0.01179  0.14189  0.84510 
## 
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                    2.083995   0.121611  17.137  < 2e-16 ***
## MonthFeb                      -0.051789   0.063863  -0.811 0.417964    
## MonthMar                      -0.040210   0.062598  -0.642 0.521075    
## MonthApr                      -0.157985   0.067030  -2.357 0.018988 *  
## MonthMay                      -0.244943   0.069000  -3.550 0.000439 ***
## MonthJun                      -0.367604   0.075087  -4.896 1.51e-06 ***
## MonthJul                      -0.329578   0.079440  -4.149 4.22e-05 ***
## MonthAug                      -0.186548   0.077153  -2.418 0.016130 *  
## MonthSep                      -0.162282   0.075743  -2.143 0.032854 *  
## MonthOct                      -0.041868   0.068024  -0.615 0.538637    
## MonthNov                      -0.029887   0.066147  -0.452 0.651681    
## MonthDec                      -0.078758   0.063405  -1.242 0.215032    
## ‘9am relative humidity (%)‘   -0.009784   0.001036  -9.445  < 2e-16 ***
## ‘Minimum temperature (Deg C)‘  0.032522   0.004553   7.143 5.49e-12 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 0.2432 on 343 degrees of freedom
## Multiple R-squared:  0.6132, Adjusted R-squared:  0.5985 
## F-statistic: 41.83 on 13 and 343 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
## 
## Response: evap_transf
##                                Df  Sum Sq Mean Sq F value    Pr(>F)    
## Month                          11 21.8352  1.9850  33.556 < 2.2e-16 ***
## ‘9am relative humidity (%)‘     1  7.3142  7.3142 123.641 < 2.2e-16 ***
## ‘Minimum temperature (Deg C)‘   1  3.0187  3.0187  51.029  5.49e-12 ***
## Residuals                     343 20.2906  0.0592                      
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

At the completion of this iteration, all remaining predictors were significant at the 5% level.

Significant Terms

Following the dimensionality reduction algorithm outlined above, the remaining predictor variables were:

- The categorical variable Month, with a p-value $< 2.2 \times 10^{-16}$ (taken from the ANOVA output), where we also note that **January is set as the reference category in the summary output.**

- The quantitative variable `9am relative humidity (%)`, with a p-value $< 2 \times 10^{-16}$ (taken from the linear model summary output).

- The quantitative variable `Minimum temperature (Deg C)`, with a p-value of 5.49×10^{-12} (taken from the linear model summary output).

These three predictor terms all exhibited a moderate to strong correlation to the response variable as seen in the bivariate analysis. However the variable `maxtemp_transf` failed to exhibit significance at the 5% level (and therefore was not used in the final model) despite a moderate correlation with the response variable. This is due to **colinearity** which is explored and discussed in detail in the 'Results' section.

Model Diagnostics

In this section we tested model assumptions. Namely:

- Linearity (*Figure 19*)
- Normality of ϵ_i's (*Figure 20*)
- Constant variance of ϵ_i's (*Figure 21*).
- Independence of the error terms.

These are outlined over the following pages:

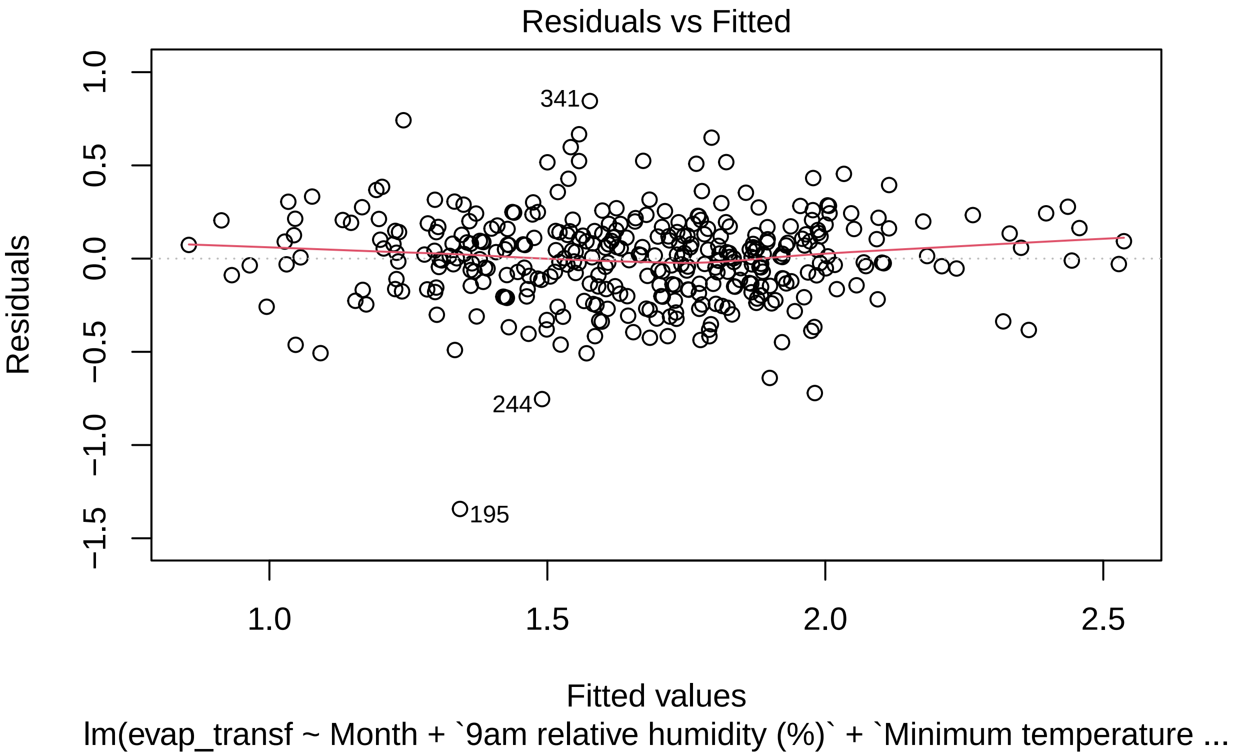


Figure 19: Plot of Residuals vs Fitted to check for Linearity

Linearity - checking that a straight line was the best way to model the relationship between the variables.

We see no discernible curvature in the points as we go from left to right. The red line is overall quite straight. We also notice no trend in the residuals which is a good indication that **the assumption of linearity is satisfied.**

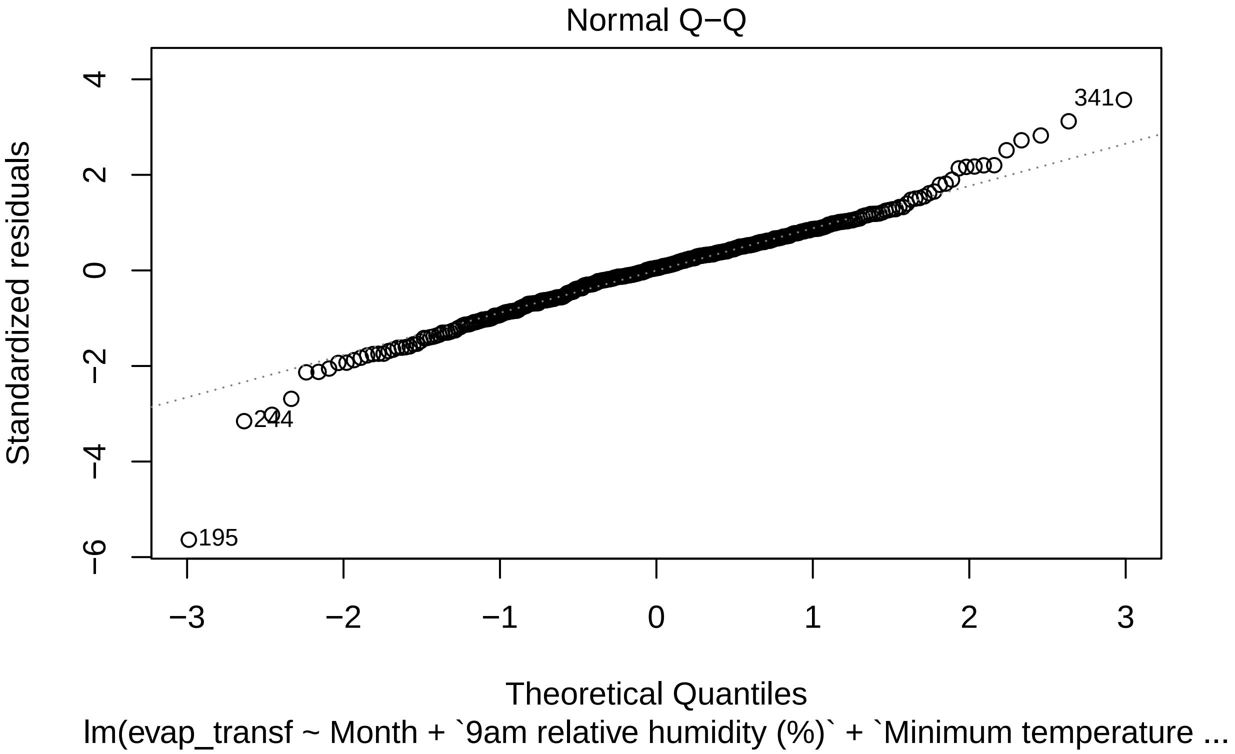


Figure 20: Q-Q Plot of Residuals to check for Normality

Normality of ϵ_i's

Normally distributed residuals lie along the dotted line. In both instances we see points beginning to drift away from the line beyond $x = -2$ and $x = 2$. Since these points drift away beyond $x = -2$ and $x = 2$ we can assume that about 95% of the residuals are normally distributed as this represents two standard deviations from the mean. As such, the **normality of the error terms is upheld.**

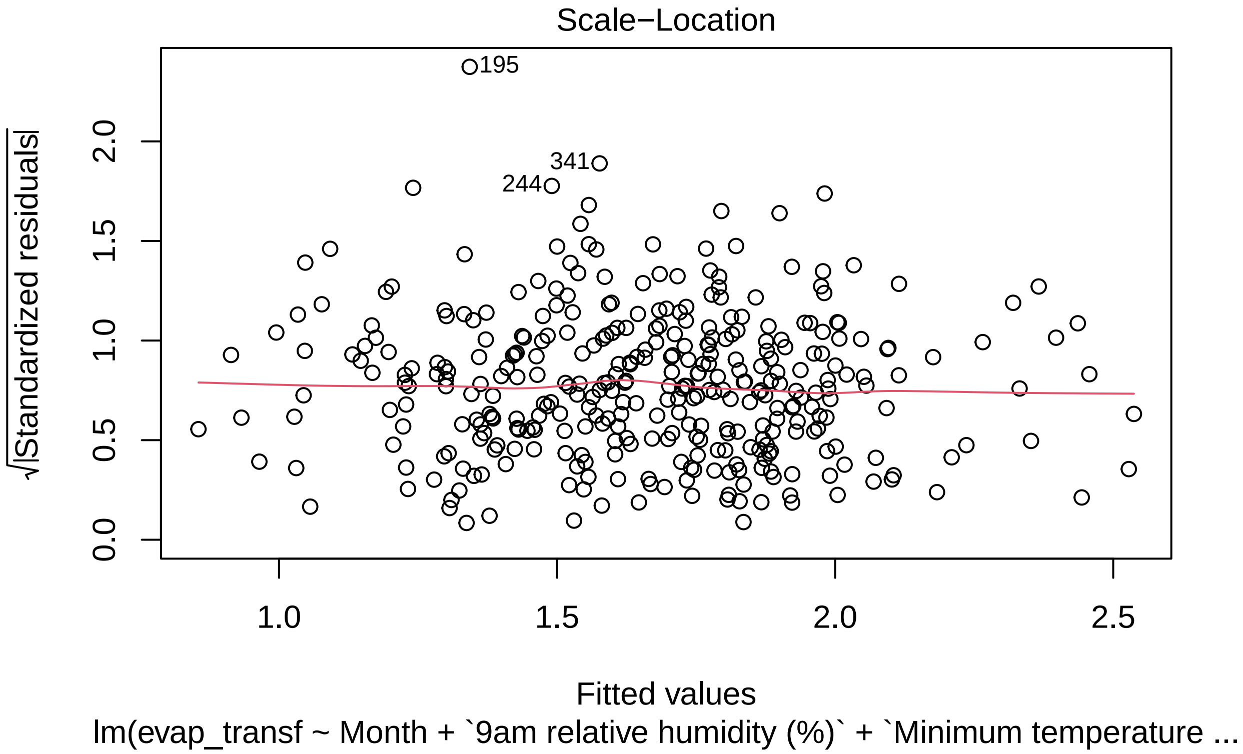


Figure 21: Plot of the square root of the Standardised Residuals to test for Homoscedascity

Constant variance of ϵ_i's (i.e., homoscedascity).

We see a relatively even spread of the points with slight condensing in the middle however no apparent trends as we move from left to right. Also, the red line is rather straight and flat. Therefore, **the constant variance assumption is upheld.**

Independence of the error terms.

If we consider the fact we have a time series of meteorological values as daily observations it is clear that there exists a dependence between observations. The annual periodicity evident in seasonal temperatures is a function of the Earth/Sun system producing warmer and cooler months of the year in a cyclical fashion. Generally, observations made in a particular month will carry similar average conditions (e.g. temperatures) as other observations made around the same time. Furthermore, with respect to the relationship between the variables used in our model, the rate at which water evaporates from the surface of a lake is directly influenced by temperature as well as relative humidity. Higher temperatures correlate to higher evaporation rates because as temperature increases, the amount of energy necessary for evaporation decreases. As such, the loss of water by evaporation in warm weather is greater than in cool weather. Since the relative humidity is simply a measure of the water vapor content of the air, lower relative humidity means drier air, and thus, a higher evaporation rate. The more humid the air, the closer the air is to saturation, and less evaporation can occur. With respect to temperature, warm air can also hold higher concentrations of water vapour, so essentially there is more room for more water vapour to be stored in warmer air compared to colder air (Integrate Program, 2021).

Because of these fundamental dependencies between temperature, relative humidity, and amount of water evaporation, coupled with annual cyclical temperature variations, **the independence of our respective error terms cannot be upheld.**

RESULTS

As outlined in the Methods section above, the predictor variables used in the model are:

- The categorical variable Month.
- The quantitative variable `9am relative humidity (%)`.
- The quantitative variable `Minimum temperature (Deg C)`.

Therefore our model takes the form:

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$

Where

- y_i is the i^{th} transformed response value for evaporation (with transformed units of $(mm^{1/3})$),
- β_0 is the intercept = 2.083995,
- β_1 is the coefficient for the i^{th} month (x_{i1}),
- $\beta_2 = -0.009784$ is the coefficient for the i^{th} value of relative humidity (x_{i2}), and
- $\beta_3 = 0.032522$ is the coefficient for the i^{th} value of minimum temperature (x_{i3}).

Putting it together we get:

evap_transf = 2.083995 + β_1(`Month`) -0.009784(`9am relative humidity`) + 0.032522(`Minimum temperature (Deg C)`) +ϵ_i

The following two plots show the relationship between each quantitative predictor and the response variable for our model:

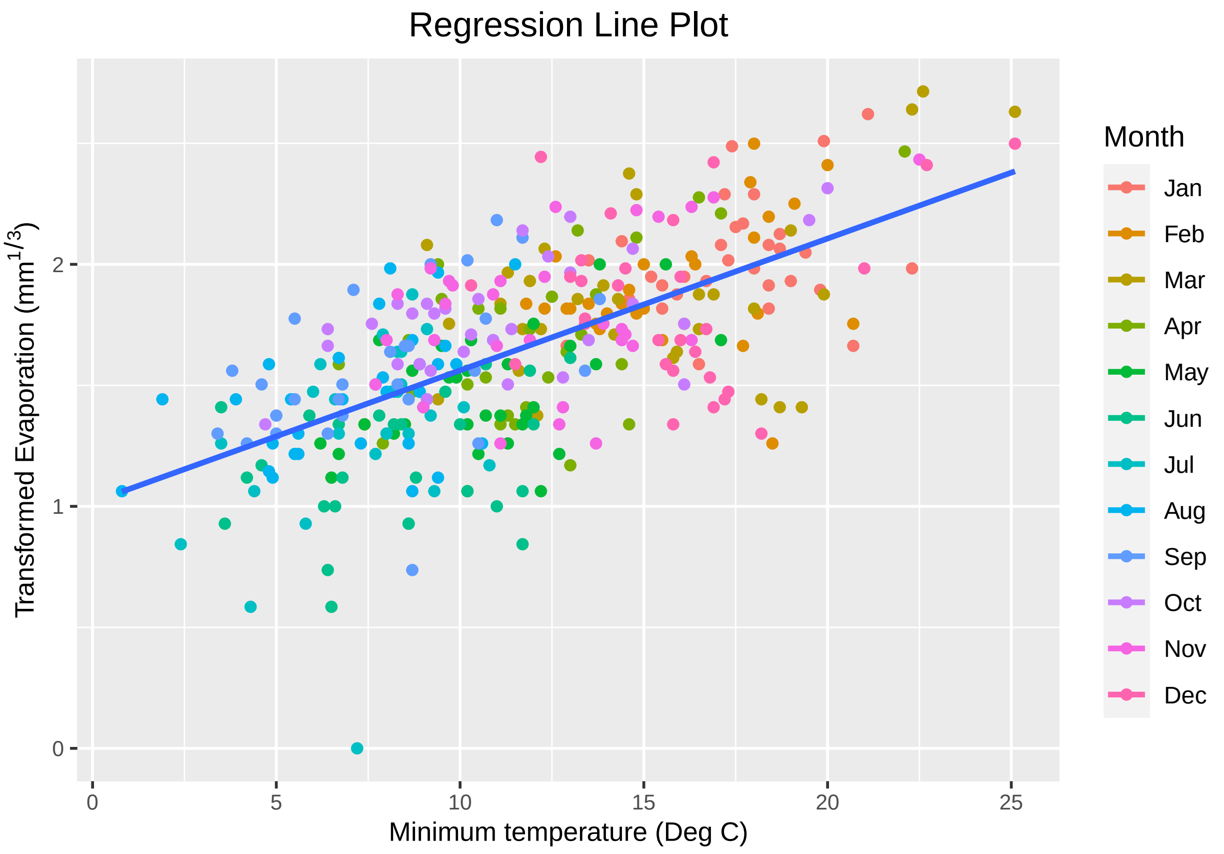


Figure 22: Plot of Minimum Temperature vs Transformed Evaporation

Here we see the positive linear relationship between the predictor variable `Minimum temperature (Deg C)` and the transformed response variable.

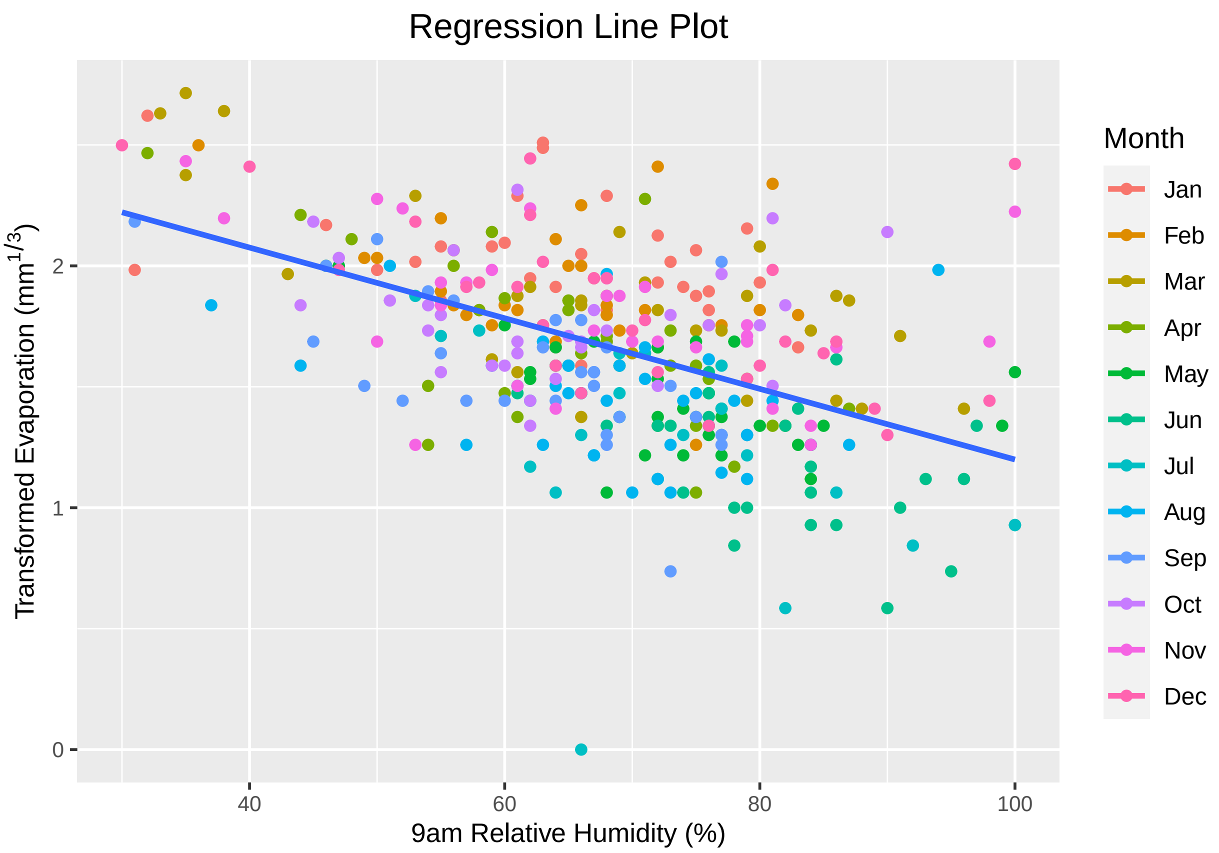


Figure 23: Plot of Relative Humidity vs Transformed Evaporation

Here we can see the negative linear relationship between the predictor variable `9am Relative Humidity` and the transformed response variable.

Interpretation of Intercept and Slope Coefficients

The following examples aim to provide easy to understand interpretations of the intercept and coefficients relating to each predictor:

- If we set the Month = January (i.e., $\beta_1 = 0$), set relative humidity to zero, and the minimum temperature to zero then our model becomes:

`evap_transf` = 2.083995 + 0 – (0.009784 × 0) + (0.032522 × 0)

Which means our x_{i1}, x_{i2} and x_{i3} terms drop off leaving only the **intercept** term. Therefore, our model becomes:

`evap_transf` = 2.083995

Remembering this is the transformed value, so raising 2.083995 to the power of 3 gives us 9.050864. Therefore, our model predicts that with these (unlikely) conditions in January we still see approximately 9.05mm of evaporation.

- If, for example, we pick Month = February (i.e., $\beta_1 = -0.051789$), set relative humidity to zero, and the minimum temperature to zero then our model becomes:

`evap_transf` = 2.083995 - 0.051789 – (0.009784 × 0) + (0.032522 × 0)

Which means our x_{i2} and x_{i3} terms drop off leaving only the **Month coefficient** and **intercept** term. Therefore, our model becomes:

`evap_transf` = 2.083995 - 0.051789 = 2.032206

Raising 2.032206 to the power of 3 gives us 8.392729. Therefore, our model predicts that with these (also unlikely) conditions in February we still see approximately 8.39mm of evaporation.

- Finally, if we set the Month = January (i.e., $\beta_1 = 0$), the relative humidity at 100%, and the minimum temperature to 10 Deg C we get:

`evap_transf` = 2.083995 – (0.009784 × 100) + (0.032522 × 10) = 1.430815

Which means we have non-zero values for our x_{i2} and x_{i3} **slope coefficient terms**. Raising 1.430815 to the power of 3 gives us an estimated evaporation of 2.92mm for days in January when the relative humidity at 9am is 100% and the minimum temperature (in the 24 hours to 9am) is 10 Deg C.

In general we can say the following with respect to the coefficients for `Minimum Temperature (Deg C)` and `9am relative humidity (%)`:

- The coefficient for `Minimum Temperature (Deg C)`= 0.032522 is *positive* which means that for every unit increase in the variable there is an increase in the amount of evaporation.
- The coefficient for `9am relative humidity (%)`= -0.009784 is *negative* which means that for every unit increase in the variable there is an decrease in the amount of evaporation (i.e. an inversely proportional relationship).

These relationships are evident in the following bivariate plots (which are reproduced for clarity):

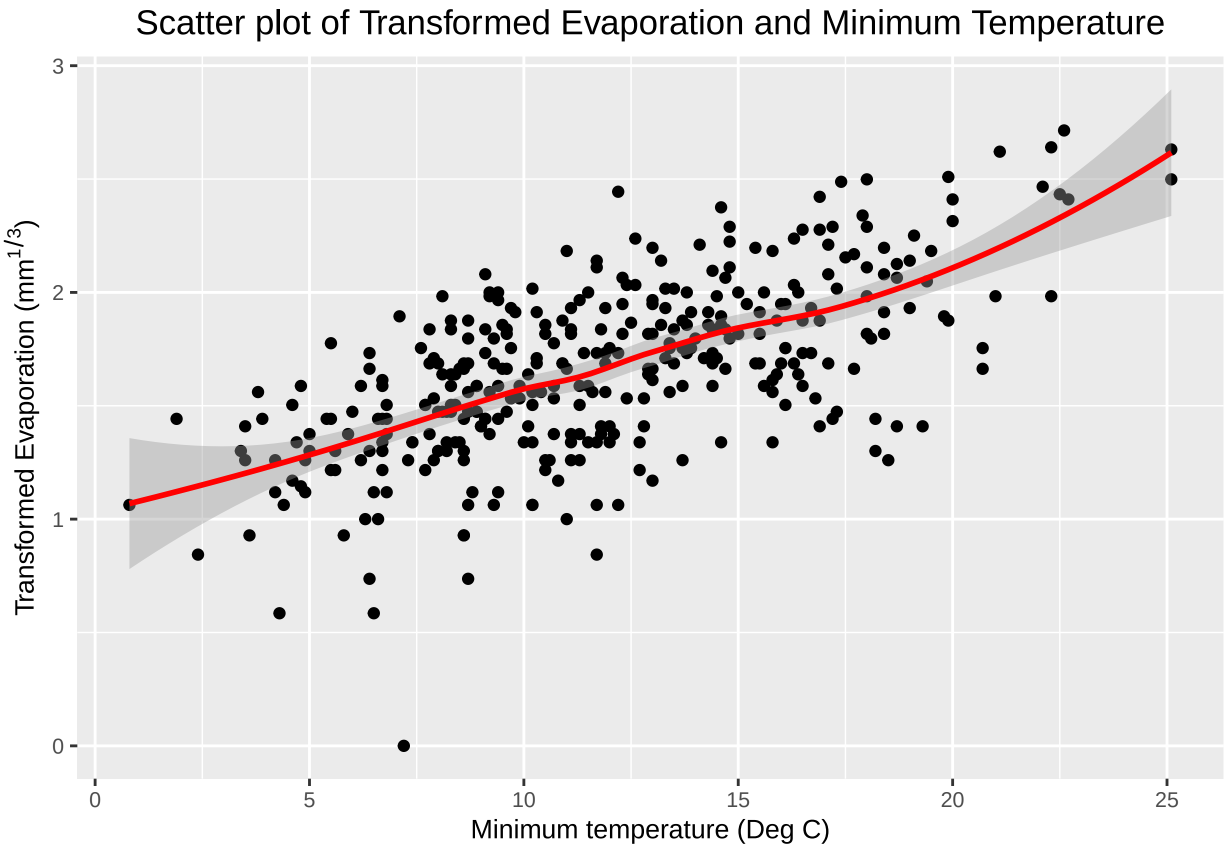


Figure 24: Scatter plot of transformed Evaporation and Minimum temperature (Deg C)

Note the positive relationship between minimum temperature and evaporation amount.

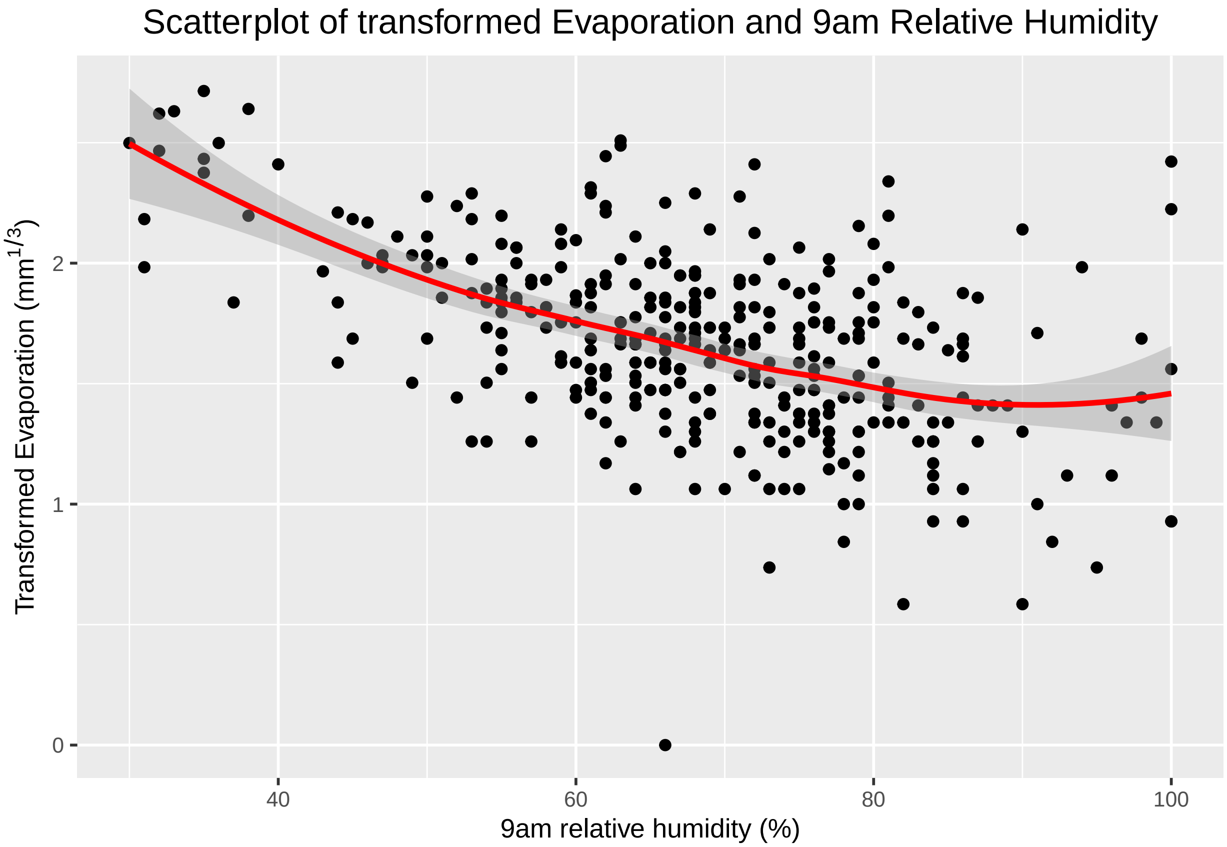
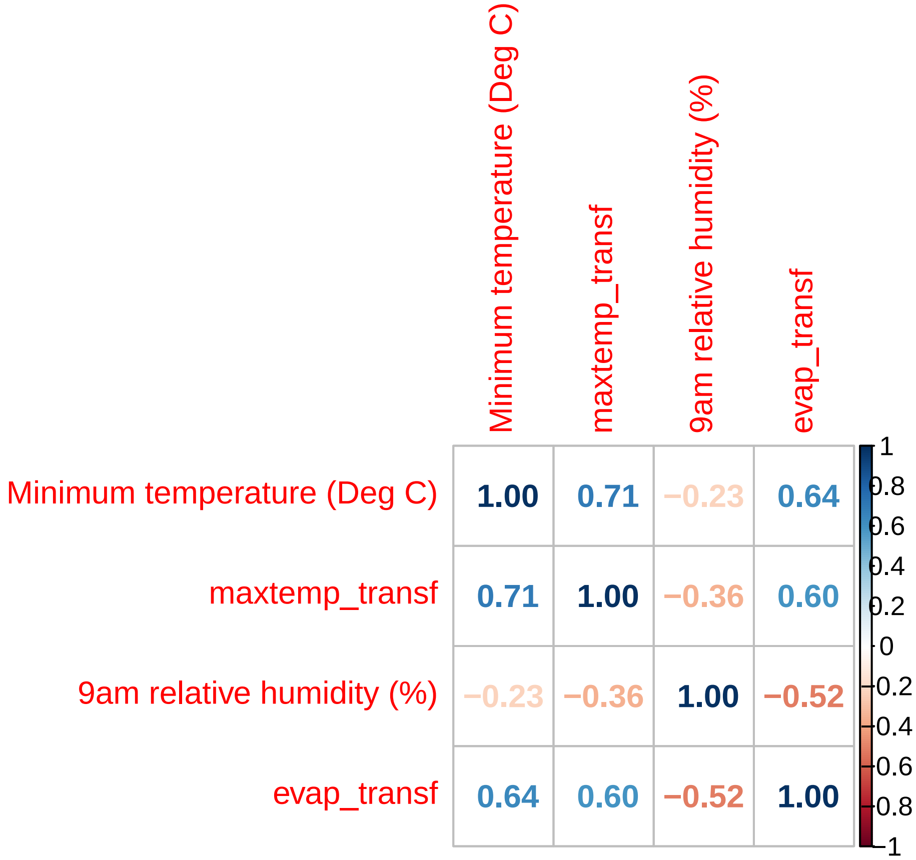


Figure 25: Scatterplot of transformed Evaporation and 9am Relative Humidity

Note the negative relationship between 9am relative humidity and evaporation amount.

Further Comments on Results

The variable `maxtemp_transf` showed a moderate correlation with the response variable however was not significant at the 5% level. It was therefore excluded from the model. The assumption here is that this is due to colinearity. This is verified with the following linear correlation matrix:



	Minimum temperature (Deg C)	maxtemp_transf	9am relative humidity (%)	evap_transf
Minimum temperature (Deg C)	1.00	0.71	−0.23	0.64
maxtemp_transf	0.71	1.00	−0.36	0.60
9am relative humidity (%)	−0.23	−0.36	1.00	−0.52
evap_transf	0.64	0.60	−0.52	1.00

Here we note a high positive correlation (= 0.71) between `Minimum temperature (Deg C)` and `maxtemp_transf` (our transformed maximum temperature variable).

It is worth noting that if we remove the variable `Minimum temperature (Deg C)` from the model selection algorithm altogether and execute the dimensional reduction algorithm as above, then `maxtemp_transf` becomes highly significant and ends up as a predictor in the final model. This is further evidence which points to the existence of colinearity between `Minimum temperature (Deg C)` and `maxtemp_transf`, where the linear regression performed by R has placed weighted significance on `Minimum temperature (Deg C)` in this case.

DISCUSSION

In response to the request from Melbourne Water Company (MWC), the model was used to predict the amount of evaporation in millimeters for some specific extreme scenarios. The details are:

- February 29, 2020, with a minimum temperature of 13.8 degrees and a maximum temperature of 23.2 degrees, and 74% humidity at 9am.
- December 25, 2020, with a minimum temperature of 16.4 degrees and a maximum temperature of 31.9 degrees, and 57% humidity at 9am.
- January 13, 2020, with a minimum temperature of 26.5 degrees and a maximum temperature of 44.3 degrees, and 35% humidity at 9am.
- July 6, 2020, with a minimum temperature of 6.8 degrees and a maximum temperature of 10.6 degrees, and 76% humidity at 9am.

Prediction

The following table lists the required attributes from the specific scenarios above relevant to the predictors in our model and provides calculated prediction intervals in each case:

Table 1: 95% prediction intervals for expected evaporation under specific conditions for specific days

Date	Min Temp. (Deg C)	9am Rel. Humidity (%)	Fit (mm)	Lower Prediction Interval (mm)	Upper Prediction Interval (mm)
February	13.8	74	5.4	2.0	11.3
December	16.4	57	7.8	3.3	15.0
January	26.5	35	17.6	9.4	29.7
July	6.8	76	1.9	0.4	5.1

Based on our model, we can say the following with respect to the prediction intervals generated for the scenarios above (making note of the fact that a specific date falls within a particular month which our model uses as the categorical predictor):

- On February 29, 2020, if this day has a minimum temperature of 13.8 degrees, and has 74% relative humidity at 9am, we are 95% confident the amount of evaporation will fall between 2mm and 11.3mm.
- On December 25, 2020, if this day has a minimum temperature of 16.4 degrees, and has 57% relative humidity at 9am, we are 95% confident the amount of evaporation will fall between 3.3mm and 15mm.
- On January 13, 2020, if this day has a minimum temperature of 26.5 degrees, and has 35% relative humidity at 9am, we are 95% confident the amount of evaporation will fall between 9.4mm and 29.7mm.
- On July 6, 2020, if this day has a minimum temperature of 6.8 degrees, and has 76% relative humidity at 9am, we are 95% confident the amount of evaporation will fall between 0.4mm and 5.1mm.

Prediction Comparison

From these predictions we can see that the day with the highest 9am relative humidity also has the lowest minimum temperature and thus produces the lowest predicted amount of evaporation. Conversely, we see the day with the highest minimum temperature having the lowest 9am relative humidity and thus the greatest predicted amount of evaporation. These contrasting results agree with our understanding of the hydrological process of evaporation as outlined on page 36. Further evidence which supports this understanding is provided by the fact that we also see greater amounts of evaporation in warmer months. Additionally, these contrasting examples also highlight the corresponding relationships between the quantitative predictor coefficients with the response variable in our model, where there is a proportional relationship between minimum temperature and evaporation, and an inversely proportional relationship between relative humidity and evaporation.

Prediction: 10mm Evaporation

If there is more than 10mm of evaporation at the reservoir, the MWC takes temporary measures to ensure continuous supply of water. As such, confidence intervals were calculated for the specific days above. This allows us to state, with 95% confidence, whether this will occur for days with the same conditions:

Table 2: 95% confidence intervals for expected evaporation under specific conditions on specific dates

Date	Min Temp. (Deg C)	9am Rel. Humidity (%)	Fit (mm)	Lower Confidence Interval (mm)	Upper Confidence Interval (mm)
February	13.8	74	5.4	4.6	6.3
December	16.4	57	7.8	6.7	8.9
January	26.5	35	17.6	15.2	20.3
July	6.8	76	1.9	1.5	2.3

Using our model, and from the confidence intervals in Table 2, we can say:

- On January 13, 2020 (and other days in January with the same conditions), if the minimum temperature is 26.5 degrees, and the relative humidity measured at 9am is 35%, the amount of evaporation will fall into the range of 15.2mm to 20.3mm 95% of the time. Therefore, we can say (with 95% confidence) that days in January with these conditions **require** temporary measures.
- On February 29, 2020, (and other days in February with the same conditions), if the minimum temperature is 13.8 degrees, and the relative humidity measured at 9am is 74%, the amount of evaporation will fall into the range of 4.6mm to 6.3mm 95% of the time. Therefore, we can say (with 95% confidence) that days in February with these conditions **do not require** temporary measures.
- On July 6, 2020, (and other days in July with the same conditions) if the minimum temperature is 6.8 degrees, and the relative humidity measured at 9am is 76%, the amount of evaporation will fall into the range of 1.5mm to 2.3mm 95% of the time. Therefore, we can say (with 95% confidence) that days in July with these conditions **do not require** temporary measures.
- On December 25, 2020, (and other days in December with the same conditions) if the minimum temperature is 16.4 degrees, and the relative humidity measured at 9am is 57%, the amount of evaporation will fall into the range of 6.7mm to 8.9mm 95% of the time. Therefore, we can say (with 95% confidence) that days in December with these conditions **do not require** temporary measures.

CONCLUSION

This study presents an improved prediction model allowing the Melbourne Water Corporation (MWC) to accurately predict daily evaporation amounts (in mm) at the Cardinia Reservior based on three predictors: the month of the year, the percentage of relative humidity (meaasured at 9am on the given day), and the minimum temperature in degrees Celsius (i.e., the lowest temperature recorded in the preceding 24 hours to 9am of the given day).

In formulating this model, it was determined that colder months correspond to lower evaporation amounts and warmer months correspond to higher evaporation. Furthermore, higher daily minimum temperatures correspond to higher evaporation amounts however the amount of evaporation decreases as the relative humidity increases. These observations (as well as the resulting model behaviour) agree with the hydrological physical processes of water evaporation and the dependent relationship evaporation has on meteorological and climatic conditions.

Another outcome of the analysis found that there is no relationship between evaporation amount and day of week. The result seems obvious, nevertheless, the statistical evidence supports this.

A relationship between maximum daily temperature and evaporation amount was discovered however the variable for maximum temperature failed to pass the statistically significant threshold in the model building process and was therefore excluded from the final model.

It is recommended that future iterations of this model take this into account and also, due to the inherent relationship between water evaporation and temperature, further research into the viability of using this variable as a predictor is suggested.

APPENDIX

Setup Code

```
install.packages("tidyverse")
```

```
library(readr)
setwd("C:/Users/Mark/Documents/Assessment 3")
melbourne <- read_csv("melbourne.csv")
```

Cleaning Code

```
library(dplyr)
library(lubridate)
#Selecting the columns I need from the dataset and recoding variables:
melbdata <- select(melbourne, "Date" : "Maximum Temperature (Deg C)",
"9am relative humidity (%)", "Evaporation (mm)") # selecting the columns.
melbdata$Date <- ymd(melbdata$Date) # recoding as date.
melbdata$`9am relative humidity (%)` <-
as.integer(melbdata$`9am relative humidity (%)`) # recoding to integer.
melbdata <- melbdata %>% mutate("Month" =
month(Date, label = TRUE, abbr = TRUE)) #adding a column for "Month".
melbdata <- melbdata %>% mutate("Day of Week" =
wday(Date, label = TRUE, abbr = FALSE)) # adding a column for "Day of Week".
melbdata <- select(melbdata, -`Date`) # removing the original date column.
inspectdf::inspect_na(melbdata) #looking for missing values.
```

```
melbdata<- na.omit(melbdata) # We shall remove these 8 rows of missing data for
#the purose of building the model, then perform a quick check:
inspectdf::inspect_na(melbdata) #looking for missing values
```

Univariate Analysis Code

```
library(ggplot2)
melbdata %>% ggplot(aes(`Evaporation (mm)`)) +
geom_histogram(fill = "light blue", col = "black", binwidth = 0.5) +
scale_y_continuous(breaks = round(seq(min(0), max(34), by = 2),2)) +
scale_x_continuous(breaks = round(seq(min(0), max(20), by = 1),2)) +
labs(title = "Histogram of Evaporation (mm)", y = "Frequency") +
theme(plot.title = element_text(hjust = 0.5),
axis.text = element_text(size = 8), axis.title = element_text(size = 10))
```

```
melbdata %>% ggplot(aes(`Evaporation (mm)`)) + geom_boxplot(fill = "light blue",
col = "black") + scale_x_continuous(breaks = round(seq(min(0), max(20), by = 1),2)) +
scale_y_continuous(breaks = NULL) + labs(title = "Boxplot of Evaporation (mm)") +
theme(plot.title = element_text(hjust = 0.5),
axis.text = element_text(size = 8), axis.title = element_text(size = 10))
```

```r
summary(melbdata$`Evaporation (mm)`)
library("moments")
skewness(melbdata$`Evaporation (mm)`, na.rm = TRUE)
sd(melbdata$`Evaporation (mm)`, na.rm = TRUE)
```

```r
# transformation of variable
melbdata <- melbdata %>% mutate(evap_transf = (`Evaporation (mm)`)^(1/3))
melbdata %>% ggplot(aes(evap_transf))+geom_histogram(fill = "light blue",
col = "red", binwidth = 0.15)+ scale_y_continuous(breaks = round(seq(min(0),
max(60), by = 5),2)) + scale_x_continuous(breaks = round(seq(min(0), max(3), by =
0.25),2)) + labs(title = bquote("Histogram of transformed Evaporation ("*mm^{1/3}*")"),
y = "Frequency") + theme(plot.title = element_text(hjust = 0.5),
axis.text = element_text(size = 8), axis.title = element_text(size = 10))
skewness(melbdata$evap_transf, na.rm = TRUE)
```

```r
melbdata %>% ggplot(aes(`9am relative humidity (%)`)) + geom_histogram(fill =
"light green", col = "black", binwidth = 2) + scale_y_continuous(breaks =
round(seq(min(0), max(40), by = 2),2)) + scale_x_continuous(breaks =
round(seq(min(30), max(100), by = 5),2)) + labs(title =
"Histogram of 9am relative humidity (%)", y = "Frequency") + theme(plot.title =
element_text(hjust = 0.5), axis.text = element_text(size = 8), axis.title =
element_text(size = 10))
```

```r
melbdata %>% ggplot(aes(`9am relative humidity (%)`)) + geom_boxplot(fill =
"light green", col = "black") + scale_x_continuous(breaks = round(seq(min(20),
max(100), by = 5),2)) + scale_y_continuous(breaks = NULL) + labs(title =
"Boxplot of 9am relative humidity (%)") + theme(plot.title = element_text(hjust = 0.5),
axis.text = element_text(size = 8), axis.title = element_text(size = 10))
```

```r
summary(melbdata$`9am relative humidity (%)`)
skewness(melbdata$`9am relative humidity (%)`)
sd(melbdata$`9am relative humidity (%)`)
```

```r
melbdata %>% ggplot(aes(`Minimum temperature (Deg C)`)) + geom_histogram(fill = "green",
col = "black", binwidth = 1) + scale_y_continuous(breaks = round(seq(min(0), max(42),
by = 2),2)) + scale_x_continuous(breaks = round(seq(min(0), max(26), by = 2),2)) +
labs(title = "Histogram of Minimum temperature (Deg C)", x = "Deg (C)",
y = "Frequency") + theme(plot.title = element_text(hjust = 0.5),
axis.text = element_text(size = 8), axis.title = element_text(size = 10))
```

```r
melbdata %>% ggplot(aes(`Minimum temperature (Deg C)`)) + geom_boxplot(fill = "green",
col = "black") + scale_x_continuous(breaks = round(seq(min(0), max(26), by = 1),2)) +
scale_y_continuous(breaks = NULL) + labs(title =
"Boxplot of Minimum temperature (Deg C)") + theme(plot.title =
element_text(hjust = 0.5), axis.text = element_text(size = 8),
axis.title = element_text(size = 10))
```

```r
summary(melbdata$`Minimum temperature (Deg C)`)
skewness(melbdata$`Minimum temperature (Deg C)`)
sd(melbdata$`Minimum temperature (Deg C)`)
```

```r
melbdata %>% ggplot(aes(`Maximum Temperature (Deg C)`)) + geom_histogram(fill =
"light grey", col = "black") + scale_y_continuous(breaks = round(seq(min(0),
max(34), by = 2),2)) + scale_x_continuous(breaks = round(seq(min(6), max(44),
by = 2),2)) + labs(title = "Histogram of Maximum Temperature (Deg C)", x =
"Deg (C)", y = "Frequency") + theme(plot.title = element_text(hjust = 0.5),
axis.text = element_text(size = 8), axis.title = element_text(size = 10))
```

```r
melbdata %>% ggplot(aes(`Maximum Temperature (Deg C)`)) + geom_boxplot(fill =
"light grey", col = "black") + scale_x_continuous(breaks = round(seq(min(0),
max(44), by = 2),2)) + scale_y_continuous(breaks = NULL) + labs(title =
"Boxplot of Maximum Temperature (Deg C)", x = "Temperature (Deg C)") +
theme(plot.title = element_text(hjust = 0.5),
axis.text = element_text(size = 8), axis.title = element_text(size = 10))
```

```r
summary(melbdata$`Maximum Temperature (Deg C)`)
skewness(melbdata$`Maximum Temperature (Deg C)`)
sd(melbdata$`Maximum Temperature (Deg C)`)
```

```r
# transformation of variable
melbdata <- melbdata %>% mutate(maxtemp_transf = (`Maximum Temperature (Deg C)`)^(1/3))
melbdata %>% ggplot(aes(maxtemp_transf))+geom_histogram(fill = "light grey", col =
"red", binwidth = 0.1)+ scale_y_continuous(breaks = round(seq(min(0), max(100),
by = 10),2)) + scale_x_continuous(breaks = round(seq(min(2), max(4), by = 0.25),2)) +
labs(title = "Histogram of transformed Maximum Temperature", x =
bquote("maxtemp_transf (Deg "*C^{1/3}*")"), y = "Frequency") +
theme(plot.title = element_text(hjust = 0.5),
axis.text = element_text(size = 8), axis.title = element_text(size = 10))
skewness(melbdata$maxtemp_transf, na.rm = TRUE)
```

```r
inspectdf::inspect_types(melbdata)
inspectdf::inspect_num(melbdata)
inspectdf::inspect_cat(melbdata)
```

Bivariate Analysis Code

```r
melbdata %>% group_by(Month) %>% ggplot(aes(x = Month, y = `evap_transf`)) +
geom_boxplot(fill = "light blue") + labs(title =
"Boxplot of transformed Evaporation by Month", y =
bquote("Transformed Evaporation ("*mm^{1/3}*")")) +
scale_y_continuous(breaks = round(seq(min(0), max(3), by = 0.5),2)) +
theme(plot.title = element_text(hjust = 0.5), axis.text = element_text(size = 8),
axis.title = element_text(size = 10))
```

```r
melbdata %>% group_by(`Day of Week`) %>% ggplot(aes(x = `Day of Week`, y =
`evap_transf`)) + geom_boxplot(fill = "Orange") + labs(title =
"Boxplot of Transformed Evaporation by Day of Week", y =
bquote("Transformed Evaporation ("*mm^{1/3}*")")) + scale_y_continuous(breaks =
round(seq(min(0), max(3), by = 0.5),2)) +
theme(plot.title = element_text(hjust = 0.5), axis.text = element_text(size = 8),
axis.title = element_text(size = 10))
```

```
ggplot(melbdata, aes(x = `9am relative humidity (%)`, y = `evap_transf`)) +
geom_point() + geom_smooth() + labs(title =
"Scatterplot of transformed Evaporation and 9am Relative Humidity",
y = bquote("Transformed Evaporation ("*mm^{1/3}*")")) +
theme(plot.title = element_text(hjust = 0.5), axis.text = element_text(size = 8),
axis.title = element_text(size = 10))

cor(melbdata$evap_transf,melbdata$`9am relative humidity (%)`, use = "complete.obs")
```

```
ggplot(melbdata, aes(x = `Minimum temperature (Deg C)`, y = `evap_transf`)) +
geom_point() + geom_smooth()+ labs(title =
"Scatter plot of Transformed Evaporation and Minimum Temperature",
y = bquote("Transformed Evaporation ("*mm^{1/3}*")")) +
theme(plot.title = element_text(hjust = 0.5), axis.text = element_text(size = 8),
axis.title = element_text(size = 10))

cor(melbdata$evap_transf, melbdata$`Minimum temperature (Deg C)`)
```

```
ggplot(melbdata, aes(x = `maxtemp_transf`, y = `evap_transf`)) + geom_point() +
geom_smooth()+ labs(title =
"Plot of Transformed Evaporation and Transformed Maximum Temperature",
x = bquote("Transformed Maximum Temperature ("*'Deg' ~C^{1/3}*")"),
y = bquote("Transformed Evaporation ("*mm^{1/3}*")")) +
theme(plot.title = element_text(hjust = 0.5), axis.text = element_text(size = 8),
axis.title = element_text(size = 10))

cor(melbdata$evap_transf,melbdata$maxtemp_transf)
```

Trivariate Analysis Code

```
ggplot(melbdata, aes(x = `9am relative humidity (%)`, y = `evap_transf`, col = `Month`)) +
geom_point() + labs(title =
"Transformed Evaporation and 9am Relative Humidity coloured by Month", x =
"9am Relative Humidity (%)", y = bquote("Transformed Evaporation ("*mm^{1/3}*")")) +
theme(plot.title = element_text(hjust = 0.5), axis.text = element_text(size = 8),
axis.title = element_text(size = 10))
```

```
ggplot(melbdata, aes(x = `Minimum temperature (Deg C)`, y = `evap_transf`, col =
`Month`)) + geom_point() +
labs(title = "Plot of Transformed Evaporation and Min. Temp. coloured by Month",
x = "Minimum Temperature (Deg C)", y =
bquote("Transformed Evaporation ("*mm^{1/3}*")")) +
theme(plot.title = element_text(hjust = 0.5), axis.text = element_text(size = 8),
axis.title = element_text(size = 10))
```

```
ggplot(melbdata, aes(x = `maxtemp_transf`, y = `evap_transf`, col = `Month`)) +
geom_point() + labs(title =
"Transformed Evaporation and Transformed Max. Temp. coloured by Month",
x = bquote("Transformed Maximum Temperature ("*'Deg' ~C^{1/3}*")"),
y = bquote("Transformed Evaporation ("*mm^{1/3}*")")) +
theme(plot.title = element_text(hjust = 0.5), axis.text = element_text(size = 8),
axis.title = element_text(size = 10))
```

Model Selection Code

```
# making sure to recode categorical variables as unordered factors

melbdata <-  melbdata %>% filter(Month %in% c("Jan", "Feb", "Mar", "Apr", "May",
"Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"))
melbdata <- melbdata %>% mutate(Month = factor(Month, ordered = FALSE))
melbdata <-  melbdata %>% filter(`Day of Week` %in% c("Sunday", "Monday",
"Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))
melbdata <- melbdata %>% mutate(`Day of Week` = factor(`Day of Week`, ordered = FALSE))
```

```
# 1st iteration
melbdata_lm <- lm(evap_transf ~ `Month` + `9am relative humidity (%)` + `Day of Week` +
`Minimum temperature (Deg C)` + maxtemp_transf + Month:`9am relative humidity (%)`,
data = melbdata)

summary(melbdata_lm)
anova(melbdata_lm)
```

```
# 2nd iteration
melbdata_lm <- lm(evap_transf ~ `Month` + `9am relative humidity (%)` +
`Day of Week` + `Minimum temperature (Deg C)` + Month:`9am relative humidity (%)`,
data = melbdata)

summary(melbdata_lm)
anova(melbdata_lm)
```

```
# 3rd iteration
melbdata_lm <- lm(evap_transf ~ `Month` + `9am relative humidity (%)` +
`Day of Week` + `Minimum temperature (Deg C)`, data = melbdata)
summary(melbdata_lm)
anova(melbdata_lm)
```

```
# 4th iteration
melbdata_lm <- lm(evap_transf ~ `Month` + `9am relative humidity (%)` +
`Minimum temperature (Deg C)`, data = melbdata)
summary(melbdata_lm)
anova(melbdata_lm)
```

Model Diagnostics Code

```
plot(melbdata_lm, which = 1)
plot(melbdata_lm, which = 2)
```

```
plot(melbdata_lm, which = 3)
```

```
melbdata %>% ggplot(aes(x = Date, y = `Minimum temperature (Deg C)`)) +
geom_line() + geom_smooth()
```

```
melbdata %>% ggplot(aes(x = Date, y = `evap_transf`)) + geom_line() + geom_smooth()
```

```
melbdata %>% ggplot(aes(x = Date, y = `9am relative humidity (%)`)) + geom_line() +
geom_smooth()
```

```
melbdata %>% ggplot(aes(x = Month, y = `Minimum temperature (Deg C)`)) + geom_point()
```

```
melbdata %>% ggplot(aes(x = Month, y = `evap_transf`)) + geom_point()
```

Results Code

```
ggplot(melbdata, aes(x = `Minimum temperature (Deg C)`, `evap_transf`,
col = Month))+ labs(title = "Regression Line Plot", x="Minimum temperature (Deg C)",
y = bquote("Transformed Evaporation ("*mm^{1/3}*")")) +
theme(plot.title = element_text(hjust = 0.5), axis.text = element_text(size = 8),
axis.title = element_text(size = 10)) + geom_point()+geom_smooth(aes(group = 1),
method = lm, se = FALSE)
```

```
ggplot(melbdata, aes(x = `9am relative humidity (%)`, `evap_transf`,
col = Month))+ labs(title = "Regression Line Plot", x="9am Relative Humidity (%)",
y = bquote("Transformed Evaporation ("*mm^{1/3}*")")) +
theme(plot.title = element_text(hjust = 0.5), axis.text = element_text(size = 8),
axis.title = element_text(size = 10)) + geom_point()+geom_smooth(aes(group = 1),
method = lm, se = FALSE)
```

```
library(corrplot)
melbdata_df <- melbdata %>% select(`Minimum temperature (Deg C)`,
`maxtemp_transf`,`9am relative humidity (%)`,`evap_transf`)
melbdata_df <- as.data.frame(melbdata_df)
mdf<- cor(melbdata_df, use = "complete.obs")
corrplot(mdf, method = "number")
```

```
ggplot(melbdata, aes(x = `Minimum temperature (Deg C)`, y = `evap_transf`)) +
geom_point() + geom_smooth(col = "red")+ labs(title =
"Scatter plot of Transformed Evaporation and Minimum Temperature",
y = bquote("Transformed Evaporation ("*mm^{1/3}*")")) +
theme(plot.title = element_text(hjust = 0.5), axis.text = element_text(size = 8),
axis.title = element_text(size = 10))
```

```
ggplot(melbdata, aes(x = `9am relative humidity (%)`, y = `evap_transf`)) +
geom_point() + geom_smooth(col = "red") + labs(title =
"Scatterplot of transformed Evaporation and 9am Relative Humidity",
y = bquote("Transformed Evaporation ("*mm^{1/3}*")")) +
theme(plot.title = element_text(hjust = 0.5), axis.text = element_text(size = 8),
axis.title = element_text(size = 10))
```

Discussion (Prediction) Code

```
#The following code produces predictions for the amount of evaporation, in mm,
#for days with specific characteristics and stores these values respectively:

melb_model_1 <- tibble(`Month` = "Feb", `Minimum temperature (Deg C)` = 13.8,
`9am relative humidity (%)` = 74)

melb_model_2 <- tibble(`Month` = "Dec", `Minimum temperature (Deg C)` = 16.4,
`9am relative humidity (%)` = 57)

melb_model_3 <- tibble(`Month` = "Jan", `Minimum temperature (Deg C)` = 26.5,
`9am relative humidity (%)` = 35)

melb_model_4 <- tibble(`Month` = "Jul", `Minimum temperature (Deg C)` = 6.8,
`9am relative humidity (%)` = 76)
```

```
#Now to produce 95% prediction intervals and display in a table:

melb_model_1_pi <- c(predict(melbdata_lm, newdata = melb_model_1, interval =
"prediction", level = 0.95))^3
melb_model_2_pi <- c(predict(melbdata_lm, newdata = melb_model_2, interval =
"prediction", level = 0.95))^3
melb_model_3_pi <- c(predict(melbdata_lm, newdata = melb_model_3, interval =
"prediction", level = 0.95))^3
melb_model_4_pi <- c(predict(melbdata_lm, newdata = melb_model_4, interval =
"prediction", level = 0.95))^3
library(knitr)

Date <- c('February','December','January','July')
MinTemp <- c(13.8,16.4,26.5,6.8)
RelHum_9am <- c(74,57,35,76)
pred_table1 <- data.frame(Date, MinTemp,RelHum_9am)
names(pred_table1)[names(pred_table1) == 'Date'] <- 'Date'
names(pred_table1)[names(pred_table1) == 'MinTemp'] <- 'Min Temp. (Deg C)'
names(pred_table1)[names(pred_table1) == 'RelHum_9am'] <- '9am Rel. Humidity (%)'


table2 <- data.frame(round(rbind(melb_model_1_pi,melb_model_2_pi,melb_model_3_pi,
melb_model_4_pi), digits = 1))
names(table2)[names(table2) == 'X1'] <- 'Fit (mm)'
names(table2)[names(table2) == 'X2'] <- 'Lower Prediction Interval (mm)'
names(table2)[names(table2) == 'X3'] <- 'Upper Prediction Interval (mm)'
rownames(table2) <-  NULL
```

```
table_1 <- cbind(pred_table1,table2) #joining the two tables together

kable(table_1, caption = "95% prediction intervals for expected
evaporation under specific conditions for specific days")
```

```
# Now to produce 95% confidence intervals and display in a table:

melb_model_1_ci <- c(predict(melbdata_lm, newdata = melb_model_1, interval =
"confidence", level = 0.95))^3
melb_model_2_ci <- c(predict(melbdata_lm, newdata = melb_model_2, interval =
"confidence", level = 0.95))^3
melb_model_3_ci <- c(predict(melbdata_lm, newdata = melb_model_3, interval =
"confidence", level = 0.95))^3
melb_model_4_ci <- c(predict(melbdata_lm, newdata = melb_model_4, interval =
"confidence", level = 0.95))^3
library(knitr)
# creating a basic table with information for each given date/weather scenario
Date <- c('February','December','January','July')
MinTemp <- c(13.8,16.4,26.5,6.8)
RelHum_9am <- c(74,57,35,76)
pred_table <- data.frame(Date, MinTemp,RelHum_9am)
names(pred_table)[names(pred_table) == 'Date'] <- 'Date'
names(pred_table)[names(pred_table) == 'MinTemp'] <- 'Min Temp. (Deg C)'
names(pred_table)[names(pred_table) == 'RelHum_9am'] <- '9am Rel. Humidity (%)'


table1 <- data.frame(round(rbind(melb_model_1_ci,melb_model_2_ci,melb_model_3_ci,
melb_model_4_ci), digits = 1))
names(table1)[names(table1) == 'X1'] <- 'Fit (mm)'
names(table1)[names(table1) == 'X2'] <- 'Lower Confidence Interval (mm)'
names(table1)[names(table1) == 'X3'] <- 'Upper Confidence Interval (mm)'
rownames(table1) <-  NULL

table <- cbind(pred_table,table1) # joining the two tables together

kable(table, caption = "95% confidence intervals for expected
evaporation under specific conditions on specific dates")
```

References

Integrate Program, 2021, Evaporation and Climate. [online] Available at: < https://serc.carleton.edu/integrate/teaching_materials/food_supply/student_materials/905 > [Accessed 3 December 2021].