MATHS 7107 Data Taming Assignment 4 Report

Ky Phong Mai

2023-03-29

Contents

1 Executive Summary

Melbourne Water Corporation (MWC), which manages the supply of water in Melbourne, has identified reliability issue with existing estimates of evaporation at their reservoirs due to recent changes in Melbourne's climate. Having good prediction of evaporation assists MWC in managing the city's water supply at Cardinia Reservoir.

Based the data provided for Melbourne's weather observations, for the financial year period from July 1st 2018 to June 30th 2019, MWC wishes to develop a new model to predict evaporation. First, the report outlines the bivariate summaries of temporal and meteorological factors that may have significant impact on evaporation. From that, the best model are selected by identifying 4 significant factors including Month, Minimum Temperature (Deg C), 9am Relative Humidity (%) and the interaction between Month and Relative Humidity.

The model is then interpreted in the result section to provide an overview of how the model operates in relation to the terms. In the discussion section, the expected evaporation along with the predicted range of the forecasts for some extreme scenarios is presented for comparison. Using the results from the model, MWC can decide whether they need to take temporary measures to ensure a continuous supply of water. The scenario on January 13,2020 with very high minimum temperature of 26.5 Degree Celsius and low humidity of 35% at 9am will likely (with 95% confidence) have evaporation amount greater than 10 mm. For scenarios like this, transferring water from Silvan Reservoir upstream might be necessary.

2 Methods

In this report, we examine the Melbourne's weather data for the financial year from 01/07/2018 to 30/06/2019. The dataset is stored in **"melbourne.csv"** file which contains daily weather observations including evaporation. The bivariate analysis is performed and a new linear regression model is proposed to predict evaporation using statistical software R Studio.

The dataset is first cleaned and some important features are selected before performing analysis (Table 1)

Table 1: Weather data with relevant factors (10 random rows)

Date	Month	Weekday	MinTemp	MaxTemp	Humidity	Evaporation
2018-11-07	Nov	Wed	11.1	17.2	55	7.2
2019-03-21	Mar	Thu	18.2	24.3	86	3.0
2018-07-25	Jul	Wed	8.3	15.7	69	4.4
2018-07-11	Jul	Wed	6.6	11.6	62	3.0
2019-01-07	Jan	Mon	17.1	23.1	55	9.0
2019-03-27	Mar	Wed	11.1	21.7	66	6.2
2018-12-08	Dec	Sat	22.7	30.3	40	14.0
2018-07-03	Jul	Tue	4.1	14.3	69	NA
2018-10-16	Oct	Tue	20.0	23.1	61	12.4
2019-01-25	Jan	Fri	21.1	42.8	32	18.0

2.1 Bivariate summaries

In this first part of the analysis, we are interested in finding potential influences of the following variables on the amount of evaporation in a day.

- Month (**Month**)
- Day of the Week (**Weekday**)
- Maximum temperature in degrees Celsius (**MaxTemp**)
- Minimum temperature in degrees Celsius (**MinTemp**)
- Relative humidity, as measured at 9am (**Humidity**)

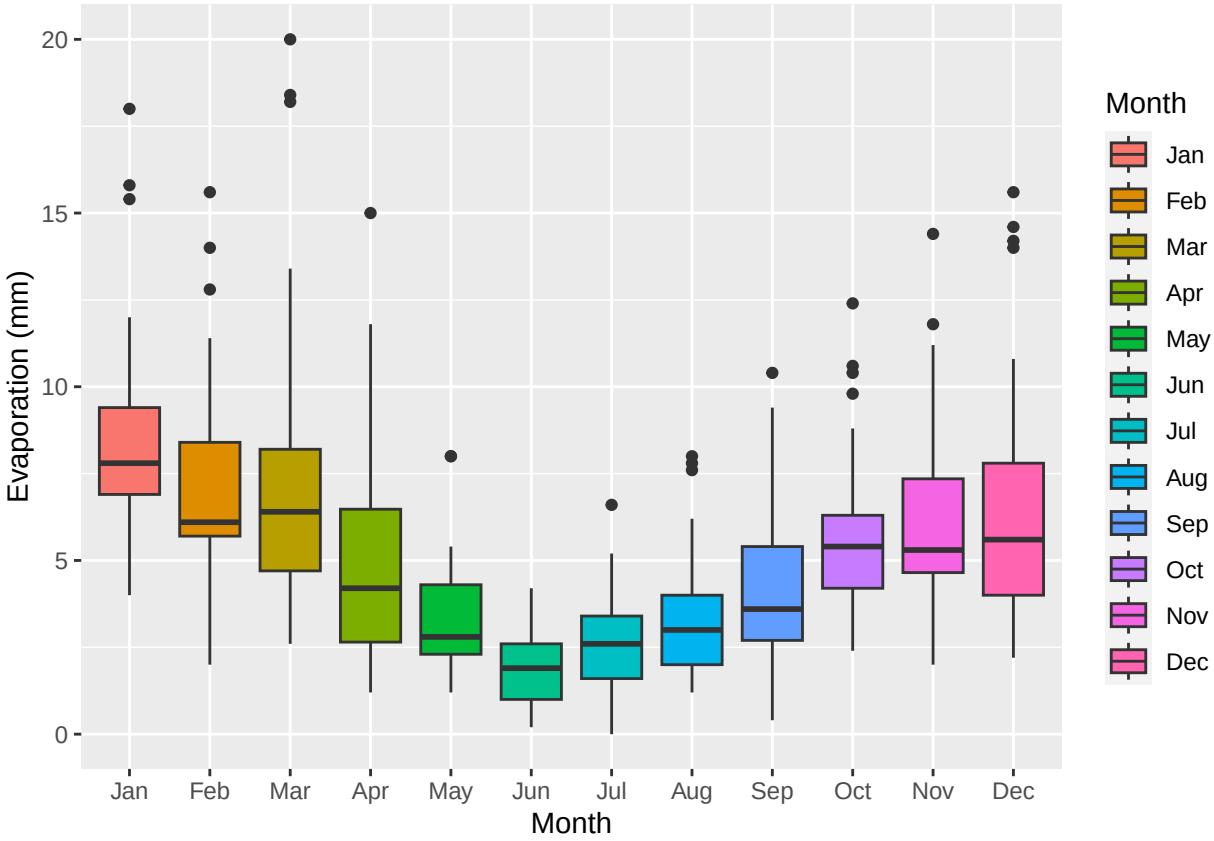

Figure 1: Side-by-side boxplot of amount of evaporation for each month

As can be seen from Figure 1, there are some substantial differences in the amount of evaporation for each month. Evaporation seems to be highest at the start of the year (January) and becomes lower towards the middle of the year before increasing again at the end of the year. January has the highest median evaporation while June has the lowest median evaporation. There seems to be a strong relationship between evaporation and month. Day with the highest evaporation was recorded in March and day with the lowest was recorded in July. Outliers were observed across most of the months except for July.

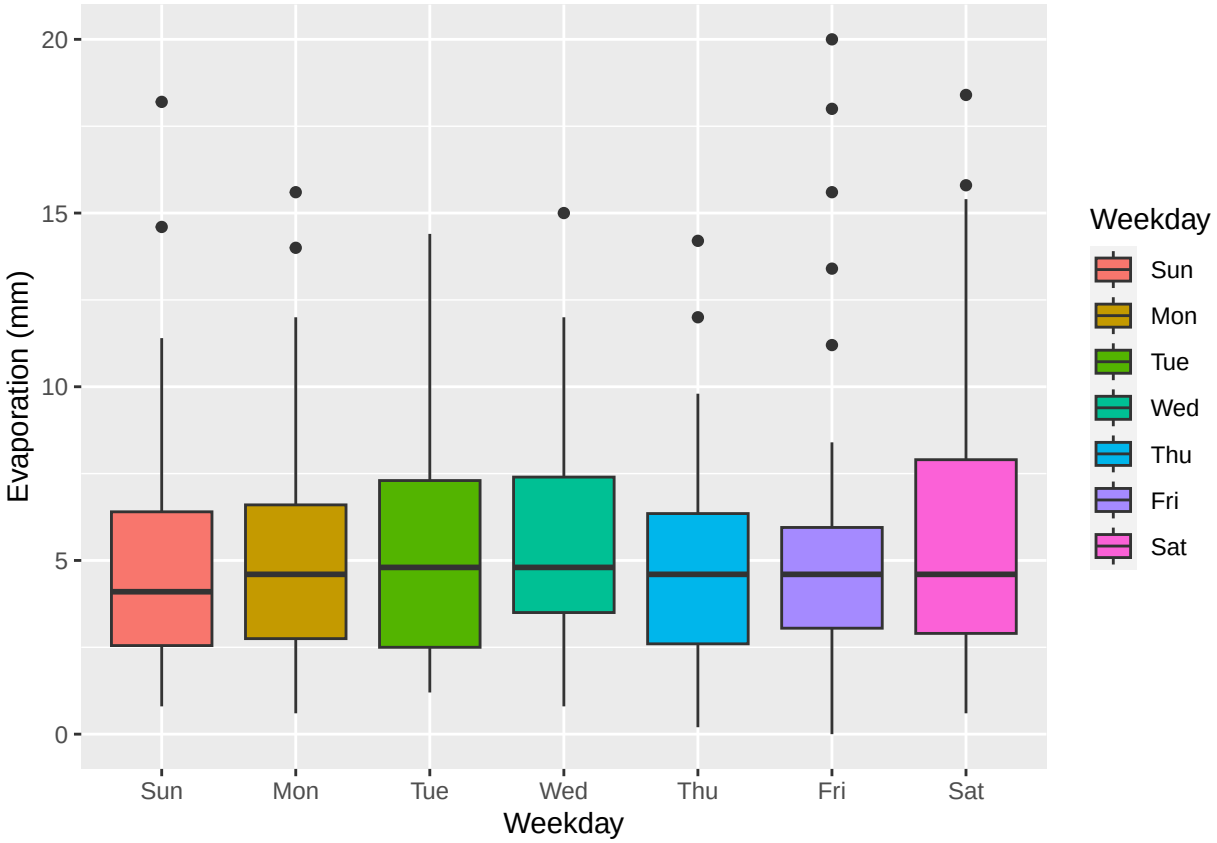

Figure 2: Side-by-side boxplot of amount of evaporation for each Day of the Week

From Figure 2, there seems to be no substantial difference in the amount of evaporation across each day of the week. The median amount of evaporation is more or less similar for any day of the week. There are outliers in most day of the week except for Tuesday. The spread of amount of evaporation seems smallest on Friday and largest on Saturday and Tuesday. No strong relationship observed.

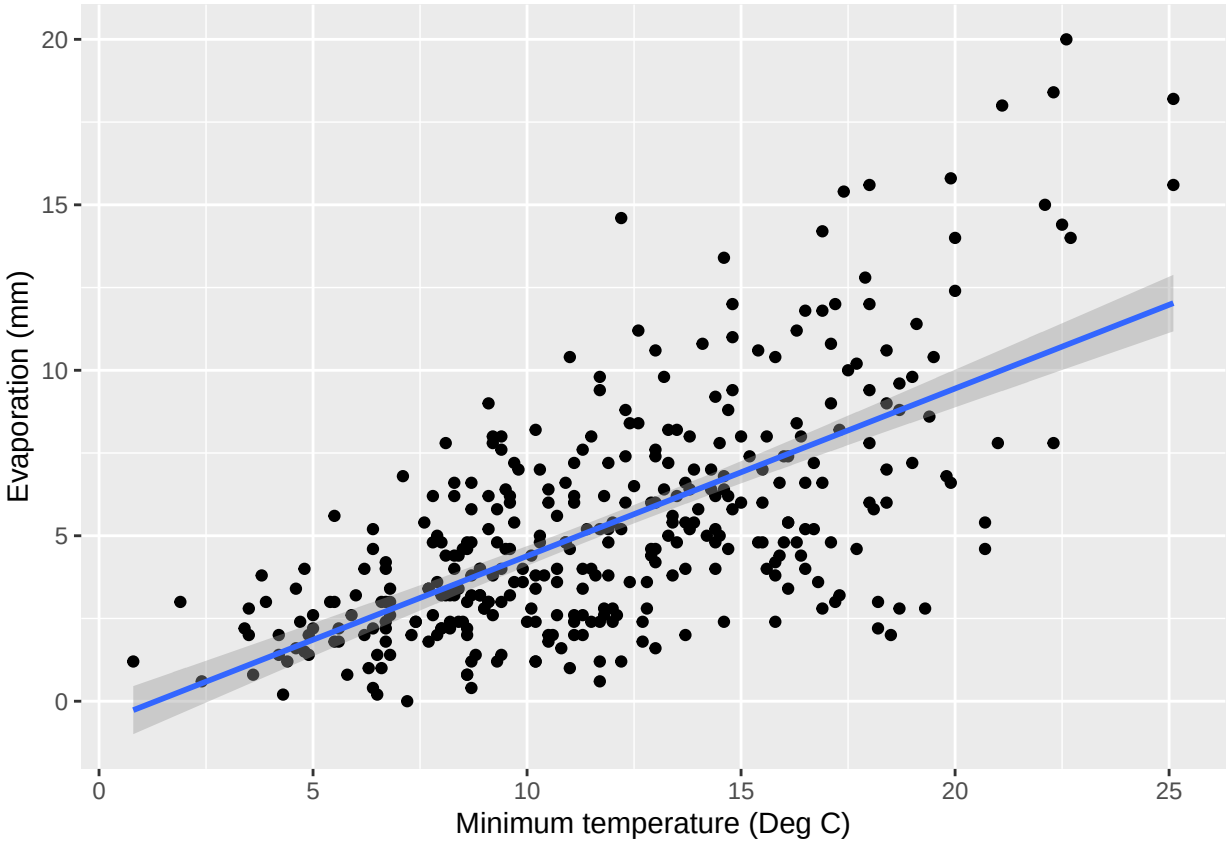

Figure 3: Scatter plot of amount of evaporation against minimum temperature

Figure 3 shows a positive, moderate linear relationship between the amount of evaporation and the minimum temperature recorded in degree Celsius

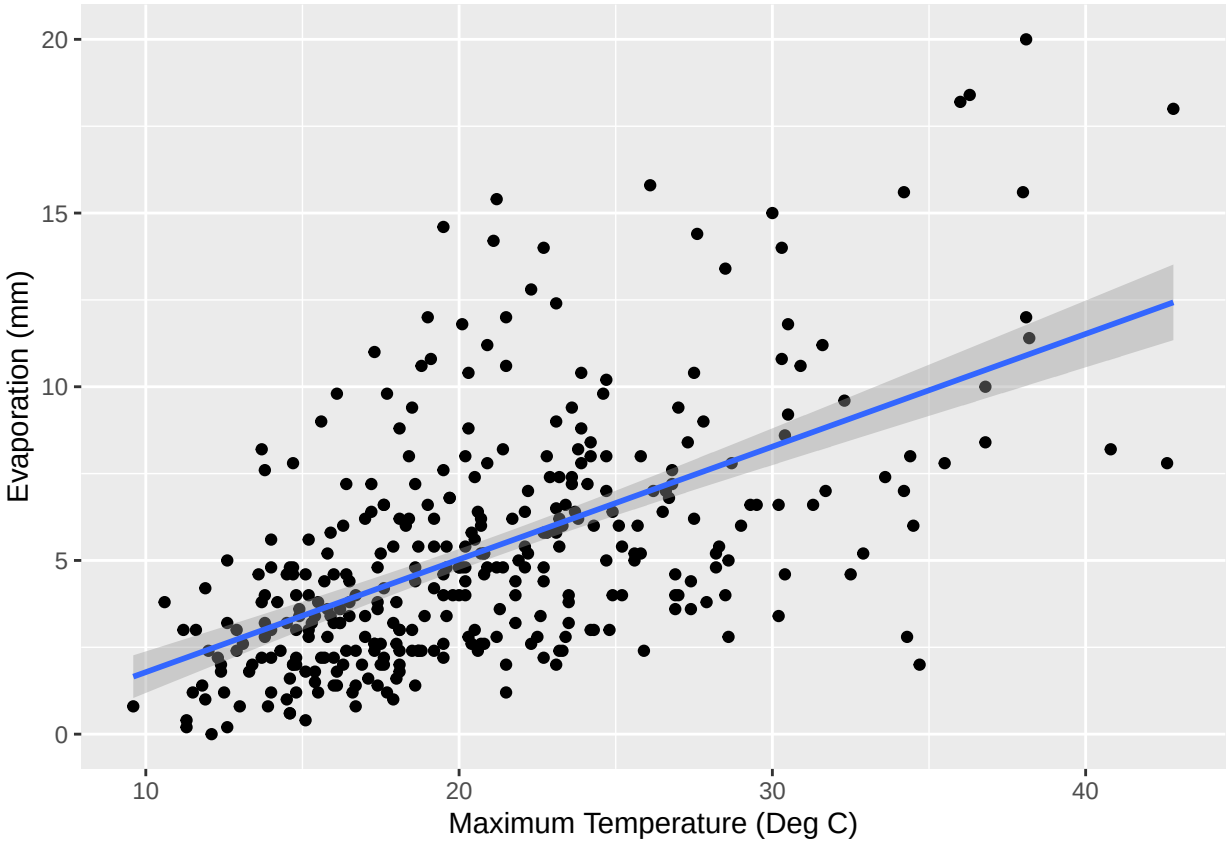

Figure 4: Scatter plot of amount of evaporation against maximum temperature

Figure 4 shows a positive, moderate linear relationship between the amount of evaporation and the maximum temperature recorded in degree Celsius

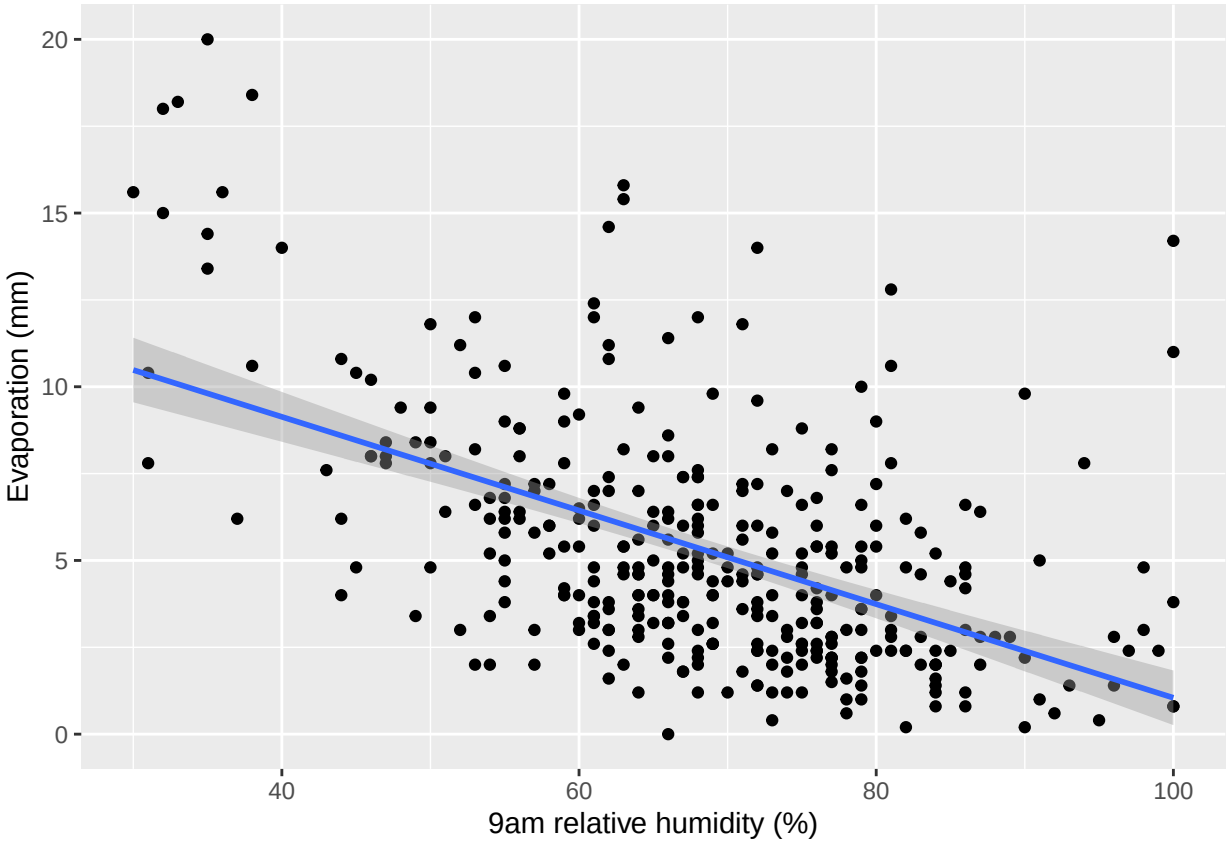

Figure 5: Scatter plot of amount of evaporation against relative humidity measured at 9am

Figure 5 shows a negative, moderate linear relationship between the amount of evaporation and the minimum temperature recorded

2.2 Model selection

This part of the report shows the steps taken to build the model using backwards selection method. We start with a full model that contains all the predictors in *Bivariate Summary* section, remove the highest p-value (if not significant). The steps are repeated until all the predictors have significant effect (p-value <0.5). Note that we also add interaction term between Month and Relative humidity in the model.

2.2.1 First model with all predictors

The following predictors are used:

- Month
- Day of the Week
- Maximum temperature in degrees Celsius
- Minimum temperature in degrees Celsius
- Relative humidity, as measured at 9am
- Interaction term between Month and Relative humidity

```
##
## Call:
## lm(formula = Evaporation ~ Month + Weekday + MinTemp + MaxTemp +
##     Humidity + Month:Humidity, data = mwc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5166 -1.1713 -0.0523  1.0677 11.0447
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         8.313165   2.375833   3.499 0.000532 ***
## MonthFeb            1.122982   3.341422   0.336 0.737028
## MonthMar            5.340251   2.630467   2.030 0.043155 *
## MonthApr            1.729320   3.102811   0.557 0.577679
## MonthMay           -4.255253   3.347211  -1.271 0.204537
## MonthJun           -7.914716   3.972809  -1.992 0.047183 *
## MonthJul           -4.930279   3.580302  -1.377 0.169442
## MonthAug           -6.310577   3.222937  -1.958 0.051083 .
## MonthSep           -0.544108   3.157664  -0.172 0.863298
## MonthOct           -6.307800   3.112895  -2.026 0.043546 *
## MonthNov           -1.080420   2.787061  -0.388 0.698525
## MonthDec            0.667154   2.793904   0.239 0.811420
## WeekdayMon         -0.272388   0.432537  -0.630 0.529304
## WeekdayTue         -0.083051   0.436596  -0.190 0.849252
## WeekdayWed         -0.078214   0.436180  -0.179 0.857801
## WeekdayThu         -0.536148   0.435847  -1.230 0.219539
## WeekdayFri         -0.408977   0.443221  -0.923 0.356828
## WeekdaySat          0.499638   0.432760   1.155 0.249127
## MinTemp             0.357912   0.044596   8.026 1.86e-14 ***
## MaxTemp             0.017765   0.030507   0.582 0.560738
## Humidity           -0.098209   0.032565  -3.016 0.002765 **
## MonthFeb:Humidity  -0.026262   0.050976  -0.515 0.606776
## MonthMar:Humidity  -0.080822   0.039559  -2.043 0.041850 *
## MonthApr:Humidity  -0.043164   0.047080  -0.917 0.359914
```

```
## MonthMay:Humidity  0.034968   0.047799   0.732 0.464966
## MonthJun:Humidity  0.078436   0.052691   1.489 0.137560
## MonthJul:Humidity  0.049674   0.051370   0.967 0.334276
## MonthAug:Humidity  0.079397   0.047371   1.676 0.094686 .
## MonthSep:Humidity -0.006753   0.049154  -0.137 0.890813
## MonthOct:Humidity  0.092502   0.047400   1.952 0.051853 .
## MonthNov:Humidity  0.015097   0.041694   0.362 0.717527
## MonthDec:Humidity -0.018916   0.041366  -0.457 0.647783
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.175 on 325 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.6458, Adjusted R-squared:  0.612
## F-statistic: 19.12 on 31 and 325 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Response: Evaporation
##                  Df  Sum Sq Mean Sq  F value    Pr(>F)
## Month            11 1478.85  134.44  28.4288 < 2.2e-16 ***
## Weekday           6   50.51    8.42   1.7801 0.1025018
## MinTemp           1  588.63  588.63 124.4719 < 2.2e-16 ***
## MaxTemp           1   74.85   74.85  15.8275  8.56e-05 ***
## Humidity          1  448.57  448.57  94.8548 < 2.2e-16 ***
## Month:Humidity   11  160.95   14.63   3.0941 0.0005645 ***
## Residuals       325 1536.94    4.73
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 2: Summary Statistics Table for first model

term	estimate	std.error	statistic	p.value
MinTemp	0.358	0.045	8.026	0.000
Humidity	-0.098	0.033	-3.016	0.003
MaxTemp	0.018	0.031	0.582	0.561

Table 3: Anova table for first model

term	df	sumsq	meansq	statistic	p.value
Month	11	1478.848	134.441	28.429	0.000
Month:Humidity	11	160.954	14.632	3.094	0.001
Weekday	6	50.508	8.418	1.780	0.103

From Table 2 and Table 3, it can be shown that MaxTemp has the highest p-value of 0.561, hence it will be removed in the next model.

2.2.2 Second model with Maximum Temperature removed

The following predictors are used:

- Month
- Day of the Week
- Minimum temperature in degrees Celsius
- Relative humidity, as measured at 9am
- Interaction term between Month and Relative humidity

```
##
## Call:
## lm(formula = Evaporation ~ Month + Weekday + MinTemp + Humidity +
##     Month:Humidity, data = mwc)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.619 -1.194 -0.085  1.098 11.063
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        8.754075   2.249680   3.891 0.000121 ***
## MonthFeb           1.070387   3.336814   0.321 0.748582
## MonthMar           5.349365   2.627753   2.036 0.042587 *
## MonthApr           1.736304   3.099641   0.560 0.575753
## MonthMay          -4.410731   3.333162  -1.323 0.186667
## MonthJun          -8.038559   3.963090  -2.028 0.043337 *
## MonthJul          -5.201366   3.546311  -1.467 0.143422
## MonthAug          -6.473398   3.207531  -2.018 0.044390 *
## MonthSep          -0.610357   3.152414  -0.194 0.846597
## MonthOct          -6.286771   3.109529  -2.022 0.044016 *
## MonthNov          -1.139353   2.782399  -0.409 0.682452
## MonthDec           0.781062   2.784222   0.281 0.779248
## WeekdayMon        -0.277981   0.431992  -0.643 0.520361
## WeekdayTue        -0.096705   0.435524  -0.222 0.824420
## WeekdayWed        -0.101325   0.433930  -0.234 0.815516
## WeekdayThu        -0.537121   0.435402  -1.234 0.218233
## WeekdayFri        -0.397814   0.442357  -0.899 0.369154
## WeekdaySat         0.485861   0.431674   1.126 0.261194
## MinTemp            0.366245   0.042195   8.680  < 2e-16 ***
## Humidity          -0.099383   0.032470  -3.061 0.002391 **
## MonthFeb:Humidity -0.025880   0.050920  -0.508 0.611617
## MonthMar:Humidity -0.081594   0.039496  -2.066 0.039631 *
## MonthApr:Humidity -0.044248   0.046996  -0.942 0.347135
## MonthMay:Humidity  0.035445   0.047744   0.742 0.458383
## MonthJun:Humidity  0.078315   0.052637   1.488 0.137765
## MonthJul:Humidity  0.051360   0.051236   1.002 0.316883
## MonthAug:Humidity  0.079649   0.047321   1.683 0.093302 .
## MonthSep:Humidity -0.007641   0.049081  -0.156 0.876378
## MonthOct:Humidity  0.091053   0.047287   1.926 0.055028 .
## MonthNov:Humidity  0.014932   0.041651   0.358 0.720201
## MonthDec:Humidity -0.021128   0.041150  -0.513 0.607989
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.172 on 326 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.6454, Adjusted R-squared:  0.6128
## F-statistic: 19.78 on 30 and 326 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Response: Evaporation
##                  Df  Sum Sq Mean Sq  F value    Pr(>F)
## Month            11 1478.85  134.44  28.4865 < 2.2e-16 ***
## Weekday           6   50.51    8.42   1.7837 0.1017458
## MinTemp           1  588.63  588.63 124.7247 < 2.2e-16 ***
## Humidity          1  519.30  519.30 110.0348 < 2.2e-16 ***
## Month:Humidity   11  163.47   14.86   3.1488 0.0004588 ***
## Residuals       326 1538.54    4.72
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 4: Summary Statistics Table for second model

term	estimate	std.error	statistic	p.value
MinTemp	0.366	0.042	8.680	0.000
Humidity	-0.099	0.032	-3.061	0.002

Table 5: Anova table for second model

term	df	sumsq	meansq	statistic	p.value
Month	11	1478.848	134.441	28.487	0.000
Month:Humidity	11	163.467	14.861	3.149	0.000
Weekday	6	50.508	8.418	1.784	0.102

From Table 4 and Table 5, it can be shown that Weekday has the highest p-value of 0.102, hence it will be removed in the next model.

2.2.3 Third model with Day of the Week & Maximum Temperature removed

The following predictors are used:

- Month
- Minimum temperature in degrees Celsius
- Relative humidity, as measured at 9am
- Interaction term between Month and Relative humidity

```
##
## Call:
## lm(formula = Evaporation ~ Month + MinTemp + Humidity + Month:Humidity,
##     data = mwc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0316 -1.1560 -0.1263  1.0184 10.6597
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        8.589140   2.202471   3.900 0.000117 ***
## MonthFeb           0.822148   3.297575   0.249 0.803268
## MonthMar           5.263051   2.610525   2.016 0.044596 *
## MonthApr           1.971572   3.040391   0.648 0.517136
## MonthMay          -4.377344   3.261415  -1.342 0.180461
## MonthJun          -8.376118   3.924447  -2.134 0.033547 *
## MonthJul          -5.360039   3.479608  -1.540 0.124412
## MonthAug          -7.102852   3.189591  -2.227 0.026625 *
## MonthSep          -1.243475   3.090815  -0.402 0.687712
## MonthOct          -6.158396   3.068813  -2.007 0.045585 *
## MonthNov          -1.036904   2.737218  -0.379 0.705066
## MonthDec           0.926791   2.748164   0.337 0.736149
## MinTemp            0.368846   0.041819   8.820  < 2e-16 ***
## Humidity          -0.099750   0.031724  -3.144 0.001815 **
## MonthFeb:Humidity -0.021806   0.050276  -0.434 0.664760
## MonthMar:Humidity -0.079813   0.039166  -2.038 0.042360 *
## MonthApr:Humidity -0.047469   0.046050  -1.031 0.303377
## MonthMay:Humidity  0.035145   0.046597   0.754 0.451246
## MonthJun:Humidity  0.083313   0.052006   1.602 0.110113
## MonthJul:Humidity  0.054069   0.050199   1.077 0.282219
## MonthAug:Humidity  0.089054   0.047045   1.893 0.059234 .
## MonthSep:Humidity  0.003411   0.048049   0.071 0.943452
## MonthOct:Humidity  0.089443   0.046676   1.916 0.056194 .
## MonthNov:Humidity  0.013451   0.040881   0.329 0.742336
## MonthDec:Humidity -0.022341   0.040556  -0.551 0.582087
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.175 on 332 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.638,  Adjusted R-squared:  0.6119
## F-statistic: 24.38 on 24 and 332 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
```

```
##
## Response: Evaporation
##                  Df  Sum Sq Mean Sq  F value    Pr(>F)
## Month            11 1478.85  134.44  28.4160 < 2.2e-16 ***
## MinTemp           1  608.93  608.93 128.7068 < 2.2e-16 ***
## Humidity          1  510.03  510.03 107.8030 < 2.2e-16 ***
## Month:Humidity   11  170.74   15.52   3.2808 0.0002758 ***
## Residuals       332 1570.74    4.73
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 6: Summary Statistics Table for third model

term	estimate	std.error	statistic	p.value
MinTemp	0.369	0.042	8.820	0.000
Humidity	-0.100	0.032	-3.144	0.002

Table 7: Anova table for third model

term	df	sumsq	meansq	statistic	p.value
Month	11	1478.848	134.441	28.416	0
Month:Humidity	11	170.742	15.522	3.281	0

From Table 6 and Table 7, it can be shown that all of the current predictors have significant effect to the amount of evaporation.

Therefore, the third model is the final model with the following significant terms:
- Month
- Minimum temperature in degrees Celsius
- Relative humidity, as measured at 9am
- Interaction term between Month and Relative humidity

The terms differ slightly from what was concluded from the bivariate analyses. From the bivariate analyses, all predictors except **Weekday**, show some moderate relationship with the amount of evaporation. However based on our final model, both **Weekday** and **MaxTemp** are not significant in predicting the evaporation amount. As for **Weekday** variable, it is very clear that there is no relationship between the variable and the amount of evaporation (from Figure 2).

However, even when there is a moderate, positive linear between **MaxTemp** and **Evaporation** (from Figure 4), **MaxTemp** is not one of the significant predictors in the final model. A possible explanation for this is due to the high correlation between 2 predictors **MinTemp** and **MaxTemp**. From Figure 3 and 4, it can be seen that both **MaxTemp** and **MinTemp** has a positive, moderate, linear relationship with **Evaporation**. This might suggest a high correlation between the two predictors, which is then verified by Figure 6. In this case, **MaxTemp** loses its predictive power because it does not explain new variation in **Evaporation**, which is already explained by **MinTemp**. Or in other words, **MaxTemp** becomes redundant since it does not provide more useful information in predicting the amount of evaporation.

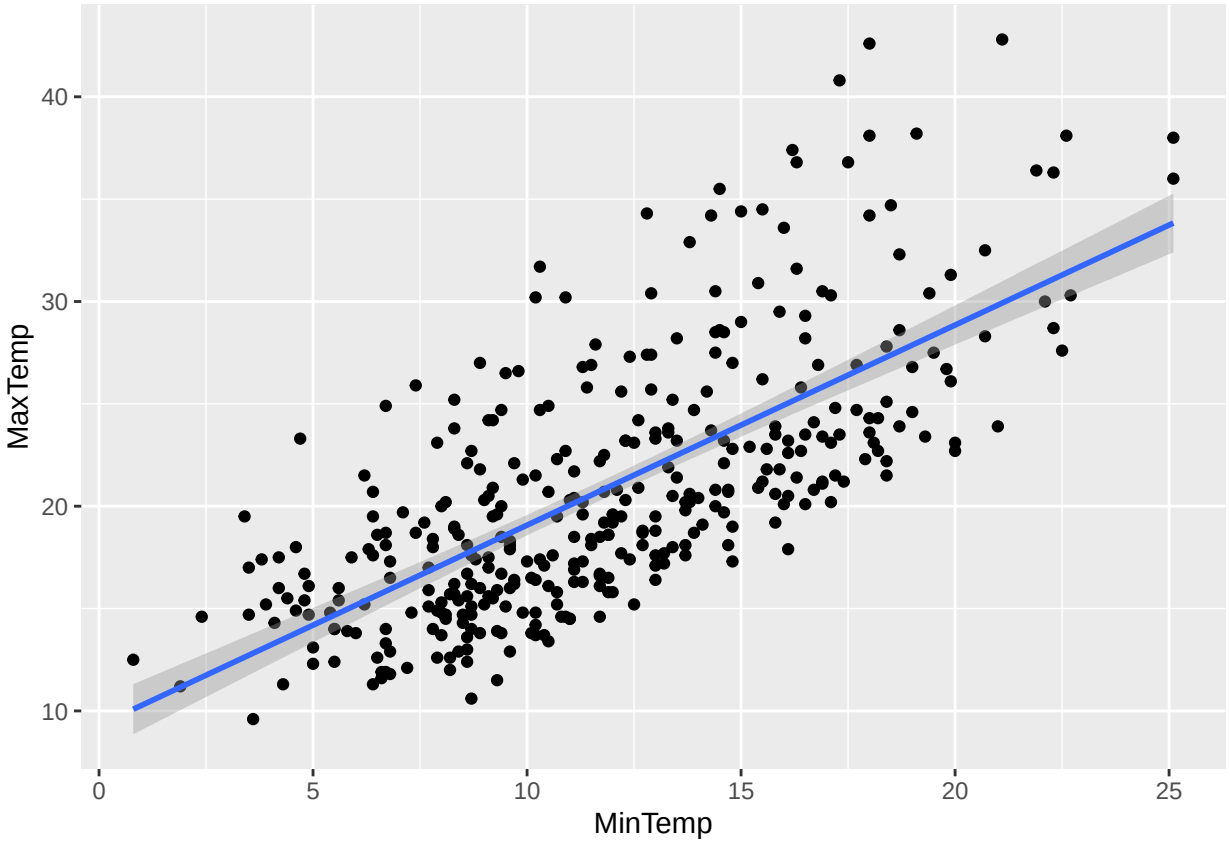

Figure 6: Moderate, positive linear relationship between MaxTemp and MinTemp

The following assumptions of the model are also tested: Linearity, Homoscedasticity, Normality and Independence. The first 3 assumptions are satisfied. As weather information of one day can affect the subsequent days, the independence assumption is not satisfied. (Refer to ***Appendix*** for more detailed assessment)

3 Results

3.1 Model interpretation

This part of the report will explain the summary statistics table of the final linear regression model obtained in the method subsection 2.2.3

Firstly, the intercept of the model in this case indicates that if the month is January, and minimum temperature and relative humidity are both 0, we will get an estimated evaporation of 8.589 mm. Note that as January is the reference month, if the month is different from January, we will need to add the coefficients of the corresponding month. For example, if the month is October, and the coefficient for **MonthOct** is -6.158, this means that if minimum temperature and relative humidity are both 0, we will get an estimated evaporation of 8.589 - 6.158 = 2.431 mm

Secondly, the coefficient of **MinTemp** is 0.369. This means that if everything else remains the same (month and humidity), an increase of 1 degree Celsius in minimum temperature will increase evaporation by 0.369 mm

The model gets more difficult to interpret when it involves Month and Humidity. As our model uses January as the reference month, the interpretation will depend on the corresponding month.

- Assume that we keep the month unchanged, when humidity increases by 1%,
  - If the month is January, the evaporation will decrease by 0.0998 mm. This is indicated by the humidity coefficient of -0.0998
  - However, if the month is anything other than January, we will need to consider the corresponding coefficient of **Month:Humidity**. For example, if the month is June, the evaporation change will be (-0.0998) + (0.0833) = -0.0165 mm. This will result in a decrease of 0.0165 mm in amount of evaporation. The value 0.0833 is the coefficient of **MonthJun:Humidity**.
- Assume that we keep the humidity unchanged at $h(\%)$, and the only thing changed is the month difference:
  - From reference month January to February, the evaporation change will be 0.822 + (-0.022) * **h** (mm). The value of 0.822 comes from the coefficient of **MonthFeb**, and the value of -0.022 comes from the coefficient from **MonthFeb:Humidity**
  - From non reference month March to April, the evaporation change will be (**MonthApr** - **MonthMar**) + (**MonthApr:Humidity** - **MonthMar:Humidity**) * **h** = (1.972) - (5.263) + ((-0.047) - (-0.08)) * **h** = -3.291 +0.033 * **h** (mm)

Another way we can interpret the model is comparing the month. If minimum temperature and relative humidity are both 0, March will have the highest evaporation of 8.589 + 5.263 = 13.852 mm; and June will have the lowest evaporation of 8.589 - 8.376 = 0.213 mm. If relative humidity increases by 1%, the evaporation change of (-0.0998) + (0.0894) = -0.0194 mm occur in October (most positive), and evaporation change of (-0.0998) + (-0.0798) = -0.180 mm occur in March (most negative). This is equivalent to March having the highest decrease in evaporation, and October having the lowest decrease in evaporation for every 1% increase in relative humidity.

In short, we can use the model to predict the evaporation as follows:

- For January: Evaporation = **(Intercept)** + **MinTemp**(*x1*) + **Humidity**(*x2*)
- For other months, for example November: Evaporation = **(Intercept)** + **MonthNov** + **MinTemp**(*x1*) + (**Humidity** + **MonthNov:Humidity**)(*x2*)

with *x1* is minimum temperature in Deg C and *x2* is relative humidity measured at 9am (%)

4 Discussion

4.1 Prediction

MWC provides some days with the following extreme scenarios as shown in table 8 and seek predictions for the amount of evaporation using our linear model.

Table 8: Data table for forecasting

Date	Month	Min Temp (DegC)	Max Temp (DegC)	Relative Humidity (%)
2020-02-29	Feb	13.8	23.2	74
2020-12-25	Dec	16.4	31.9	57
2020-01-13	Jan	26.5	44.3	35
2020-07-06	Jul	6.8	10.6	76

As can be seen from table 9, the amount of evaporation is expected to be highest on the third scenarios on January 13th 2020 at 14.872 mm with prediction ranging from 10.105 mm to 19.640 mm. This is expected because this scenario is on January which has the highest mean evaporation; has high minimum temperature, which positively correlates with evaporation; and has low relative humidity, which negatively correlates with evaporation. On the other extreme is the last case scenario which happens on July 6th 2020 (With July having low mean evaporation, low minimum temperature and high maximum temperature). It has the lowest evaporation of 2.265 mm with prediction ranging from -2.111 mm to 6.642 mm (the range is effectively from 0 mm to 6.642 mm since evaporation has to be a positive number).

If there is more than 10mm of evaporation at MWC's Cardinia Reservoir, the corporation takes temporary measures to ensure a continuous supply of water, including transferring water from its Silvan Reservoir upstream. We can say with 95% confidence that:

- For January 13, 2020, lower boundary is > 10 mm. Hence, there will be more than 10mm of evaporation. Action will need to be taken
- For February 29, 2020 and July 6, 2020, upper boundary is < 10 mm. Hence, there will not be more than 10mm of evaporation. No action needed

The other scenario on December 25, 2020, we have 95% confidence that the evaporation will fall in the range of 4.209 mm to 13.003 mm. Hence, we are not able to conclude if evaporation is greater or smaller than 10 mm.

Table 9: Predictions for the amount of evaporation, in mm, for some particular days

Date	Lower boundary (mm)	Predicted evaporation (mm)	Upper boundary (mm)
2020-02-29	1.089	5.506	9.923
2020-12-25	4.209	8.606	13.003
2020-01-13	10.105	14.872	19.640
2020-07-06	-2.111	2.265	6.642

5 Conclusion

MWC, responsible for water supply management in Melbourne, has noted a reliability problem with current estimations of evaporation in their reservoirs due to the changing climate of the city. Accurate forecasts of evaporation are crucial for MWC to effectively manage the water supply at Cardinia Reservoir.

MWC aims to create a new evaporation prediction model using weather data for Melbourne between July 1st, 2018 and June 30th, 2019. The report initially presents bivariate summaries of various temporal and meteorological factors that may have a notable influence on evaporation. The analysis shows that both Minimum Temperature and Maximum Temperature have a moderate, positive linear relationship with evaporation, while Relative Humidity has a negative linear relationship. Moreover, there appears to be a significant difference in evaporation across the different months of the year, with January having the highest median evaporation and June having the lowest median evaporation. However, the difference in evaporation between the days of the week is not substantial, and there is barely any noticeable variation in the median evaporation across the weekdays.

Using backward selection method, four factors were identified as significant: Month, Minimum Temperature in degrees Celsius, Relative Humidity at 9am expressed as a percentage, and the interaction between Month and Relative Humidity. The selected model incorporates these four factors.

The final model's terms deviate somewhat from the findings of the bivariate analyses. Although Maximum Temperature was found to have a moderate, positive linear relationship with evaporation, it was not considered a significant predictor of evaporation in the final model. One potential explanation for this discrepancy is the high correlation between Maximum Temperature and Minimum Temperature. In this case, Maximum Temperature may be considered redundant, as it does not offer any additional information for predicting the amount of evaporation, which is already explained by Minimum Temperature.

The report presents the expected evaporation and the predicted range of forecasts for some extreme scenarios, allowing MWC to make informed decisions about whether temporary measures are necessary to ensure a continuous supply of water. For instance, for the scenario on January 13, 2020, with a very high minimum temperature of 26.5 degrees Celsius and low 9am Relative Humidity of 35%, the model predicts with 95% confidence that the evaporation amount will be greater than 10 mm. In such situations, it is recommended to transfer water from Silvan Reservoir upstream to ensure a continuous supply of water.

The model has some limitations that need to be addressed. Although all the selected factors are significant in the model, the independence assumption must be met for the linear model to be considered valid. A valid independence assumption means that the data needs to be unrelated, implying that knowledge of one day's information does not provide information about another day's information. However, this is not the case in this scenario because weather information such as temperature or humidity on one day can significantly impact subsequent days. Furthermore, all the data was collected from the past, where significant climate events may have affected the entire year. Consequently, the data may not be appropriate for future forecasts. These constraints imply that the model should be used with caution and verified using other methods before making significant decisions.

6 Appendix

6.1 Code

```
knitr::opts_chunk$set(fig.pos ="H", out.extra = "", echo = TRUE)
#METHOD SECTION

#Loading libraries
pacman::p_load(tidyverse, tidymodels, readr, stringr, knitr, lubridate)

#Read csv file
mwc <- read_csv("melbourne.csv")
#Cleaning data, adding month and weekday column, remove other columns not needed in analysis
mwc <-mwc %>%
  mutate (Date = ymd (Date))%>%
  mutate (Month = factor(month(Date, label = TRUE), ordered = FALSE)) %>%
  mutate (Weekday = factor(wday(Date, label = TRUE), ordered = FALSE)) %>%
  select (Date, Month, Weekday,
          `Minimum temperature (Deg C)`,
          `Maximum Temperature (Deg C)`,
          `9am relative humidity (%)`, `Evaporation (mm)`)

# Rename column for easier use
mwc <- mwc %>%
  rename (c(MaxTemp = `Maximum Temperature (Deg C)`,
            MinTemp = `Minimum temperature (Deg C)`,
            Humidity = `9am relative humidity (%)`,
            Evaporation = `Evaporation (mm)`))

# Show the dataset
mwc %>% sample_n(10) %>%
  kable(caption = "Weather data with relevant factors (10 random rows)",
        align = "c",
        booktabs = TRUE,
        longtable = TRUE)

# Bivariate analyses

#Side-by-side boxplot of amount of evaporation for each month
mwc %>%
  ggplot(aes(x=Month, y = Evaporation, fill = Month)) +
  geom_boxplot() +
  labs(y = "Evaporation (mm)")

#Side-by-side boxplot of amount of evaporation for each Day of the Week
mwc %>%
  ggplot(aes(x=Weekday, y = Evaporation, fill = Weekday)) +
  geom_boxplot()+
  labs(y = "Evaporation (mm)")

#Scatter plot of amount of evaporation against minimum temperature
mwc %>%
```

```
  ggplot(aes(x = MinTemp, y = Evaporation)) +
  geom_point() + geom_smooth(method = "lm")+
  labs(y = "Evaporation (mm)", x = "Minimum temperature (Deg C)")

#Scatter plot of amount of evaporation against maximum temperature
mwc %>%
  ggplot(aes(x = MaxTemp, y = Evaporation)) +
  geom_point() +geom_smooth(method = "lm") +
  labs (x = "Maximum Temperature (Deg C)", y = "Evaporation (mm)")

#Scatter plot of amount of evaporation against relative humidity measured at 9am
mwc %>%
  ggplot(aes(x = Humidity, y = Evaporation)) +
  geom_point() + geom_smooth(method = "lm")+
  labs(x = "9am relative humidity (%)", y = "Evaporation (mm)")



## Model selection

# Build model using all predictors
mwc_1_lm <- lm(Evaporation ~ Month + Weekday + MinTemp +
                 MaxTemp + Humidity + Month:Humidity,data = mwc )
# Show the summary statistics & anova for first model
summary(mwc_1_lm)
anova(mwc_1_lm)
# Show summary statistics for the first model
mwc_1_lm %>% tidy() %>%
  filter (term %in% c("MinTemp", "MaxTemp", "Humidity"))%>%
  arrange(p.value)%>%
  kable(caption = "Summary Statistics Table for first model",
        digits = 3,
        longtable = TRUE,
        booktabs = TRUE)

# Show anova table for the first model
anova(mwc_1_lm) %>%tidy() %>%
  filter (term %in% c("Month", "Weekday", "Month:Humidity"))%>%
  arrange(p.value)%>%
  kable(caption = "Anova table for first model",
        digits = 3,
        longtable = TRUE,
        booktabs = TRUE)

# Build model using all predictors except for MaxTemp
mwc_2_lm <- lm(Evaporation ~ Month + Weekday + MinTemp +
                 Humidity + Month:Humidity,data = mwc )
summary(mwc_2_lm)
anova(mwc_2_lm)

# Show summary statistic table for second model with maximum temperature removed
mwc_2_lm %>% tidy() %>%
  filter (term %in% c("MinTemp", "Humidity"))%>%
```

```
  arrange(p.value)%>%
  kable(caption = "Summary Statistics Table for second model",
        digits = 3,
        longtable = TRUE,
        booktabs = TRUE)

# Show anova table for the second model
anova(mwc_2_lm) %>%tidy() %>%
  filter (term %in% c("Month", "Weekday", "Month:Humidity"))%>%
  arrange(p.value)%>%
  kable(caption = "Anova table for second model",
        digits = 3,
        longtable = TRUE,
        booktabs = TRUE)

# Build model using all predictors except for MaxTemp and Weekday
mwc_3_lm <- lm(Evaporation ~ Month + MinTemp + Humidity +
                 Month:Humidity,data = mwc )
summary(mwc_3_lm)
anova(mwc_3_lm)

# Show summary statistic table for third model with MaxTemp and weekday removed
mwc_3_lm %>% tidy() %>%
  filter (term %in% c("MinTemp", "Humidity"))%>%
  arrange(p.value)%>%
  kable(caption = "Summary Statistics Table for third model",
        digits = 3,
        longtable = TRUE,
        booktabs = TRUE)

# Show anova table for the third model
anova(mwc_3_lm) %>%tidy() %>%
  filter (term %in% c("Month", "Month:Humidity"))%>%
  arrange(p.value)%>%
  kable(caption = "Anova table for third model",
        digits = 3,
        longtable = TRUE,
        booktabs = TRUE)

# Relationship between MaxTemp and MinTemp
mwc %>%
  ggplot(aes(x = MinTemp, y = MaxTemp)) + geom_point() + geom_smooth(method ="lm")



# DISCUSSION SECTION

# Prediction

# Create new data table for prediction
day_to_pred <- tibble (Date = c('2020-02-29','2020-12-25','2020-01-13','2020-07-06'),
                       Month = c('Feb','Dec','Jan','Jul'),
                       MinTemp = c(13.8,16.4,26.5,6.8),
```

```
                        MaxTemp = c(23.2,31.9,44.3,10.6),
                        Humidity = c(74,57,35,76))

# Using predict function to predict
results <- predict (mwc_3_lm, newdata = day_to_pred, interval = "prediction")

# Add the result data to data table and rename
mwc_predict <- day_to_pred %>%
  bind_cols(results[,2])%>%
  bind_cols(results[,1])%>%
  bind_cols(results[,3])%>%
  select(Date,'...6','...7','...8')%>%
  rename(c("Predicted evaporation (mm)" = '...7'),
         c("Lower boundary (mm)" = "...6"),
         c("Upper boundary (mm)" = "...8"))

# Show data to predict
day_to_pred %>%
  rename(c("Min Temp (DegC)" = MinTemp),
         c("Max Temp (DegC)" = MaxTemp),
         c("Relative Humidity (%)" = Humidity)) %>%
  kable(caption = "Data table for forecasting",
        longtable = TRUE,
        align = "c",
        booktabs = TRUE)

# Show result table of prediction
mwc_predict %>%
  kable(digits = 3,
        align = "c",
        caption = "Predictions for the amount of evaporation,
        in mm, for some particular days",
        longtable = TRUE,
        booktabs = TRUE)

# Model diagnostics

# Linearity assumption
plot(mwc_3_lm, which = 1)

# Homoscedasticity assumption
plot(mwc_3_lm, which = 3)

# Normality assumption
plot(mwc_3_lm, which = 2)
```

6.2 Model diagnostics

In this section of the report, the assumptions of the linear model are tested including: Linearity, Homoscedasticity, Normality and Independence. Each assumptions will be presented with a relevant plot (except for Independence) and an assessment

6.2.1 Linearity

```
# Model diagnostics

# Linearity assumption
plot(mwc_3_lm, which = 1)
```

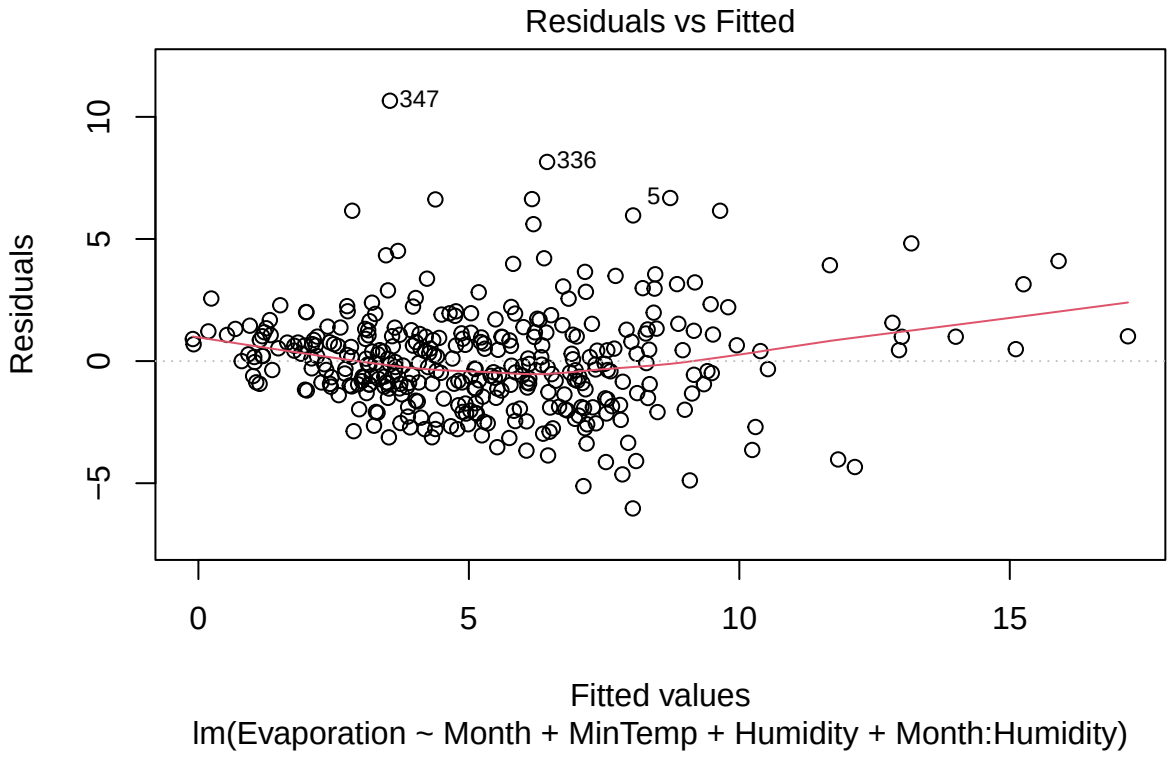

Figure 7: Plot of residuals vs fitted values for Linearity assumption check

From Figure 7, most of the points are scattered around the zero line. No curvature or trend observed. Hence, the linearity assumption is satisfied

6.2.2 Homoscedasticity

```r
# Homoscedasticity assumption
plot(mwc_3_lm, which = 3)
```

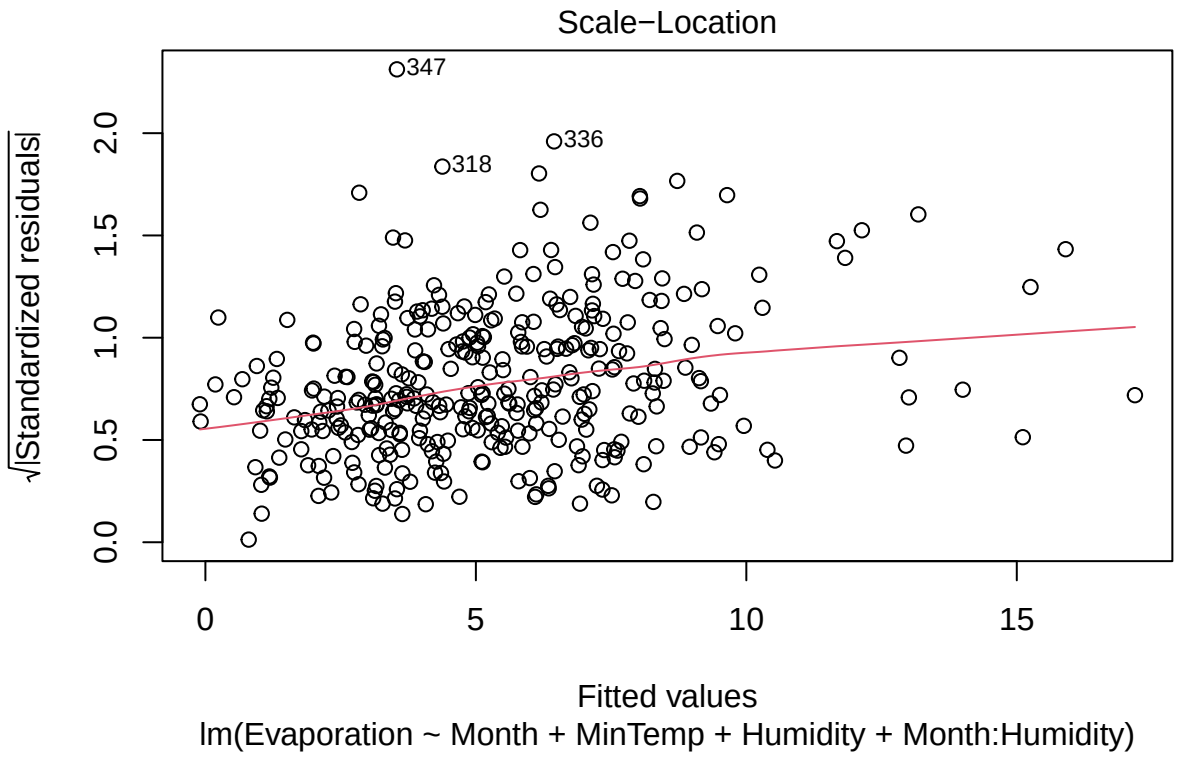

Figure 8: Plot of square root of standardized residuals against fitted values for Homoscedasticity assumption check

From Figure 8, it can be seen that the points are roughly equal spread from left to right of the plot. Hence, the Homoscedasticity assumption is reasonable.

6.2.3 Normality

```r
# Normality assumption
plot(mwc_3_lm, which = 2)
```

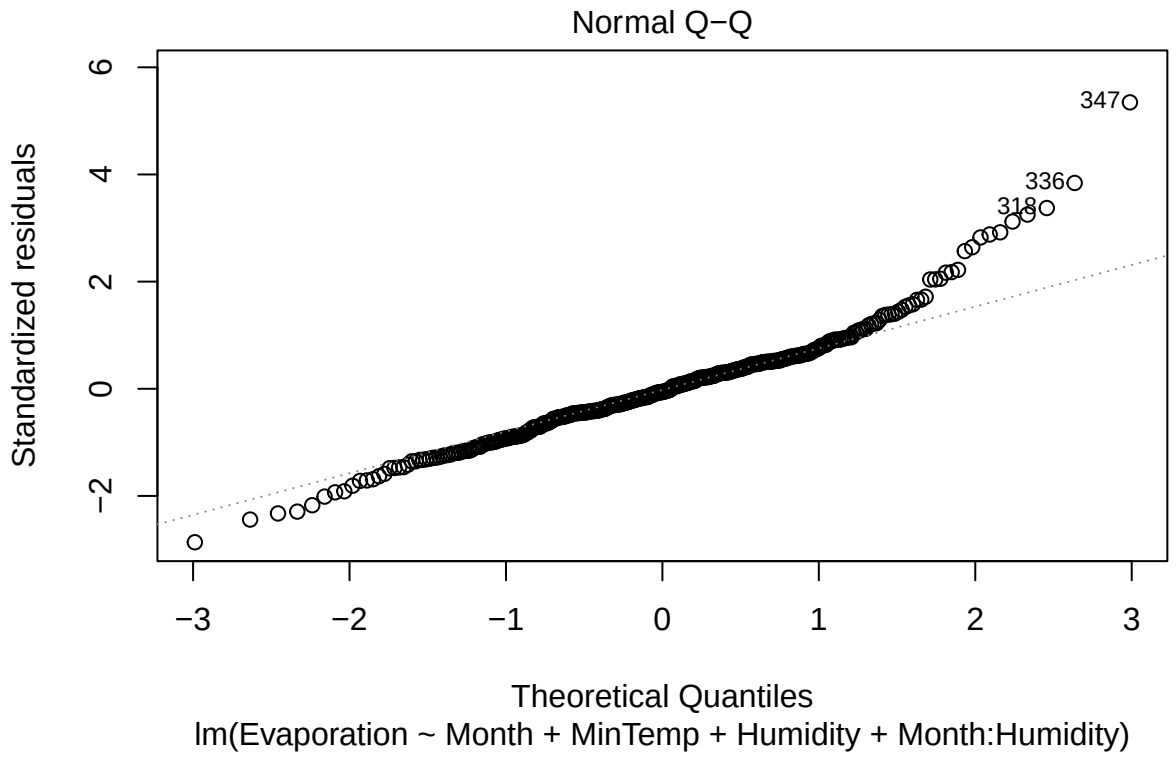

Figure 9: ormal QQ plot of the residuals for Normality assumption check

From Figure 9, it can be seen that the points are roughly distributed linearly, close to the dotted line except for a few number of points at the 2 tails. Hence, the Normal assumption is reasonable.

6.2.4 Independence

There is no particular plot that can be used to check the independence assumption. We need to justify based on how the data is obtained. The assumption can only be deemed satisfied if observations from one subject do not give more information about other subjects. In this case, because the weather information(temperature, humidity. . .) of one day can affect the weather of the following days, the independence is not satisfied.