MATHS 7107 Data Taming Assigmment 04

Possakorn Kittipipatthanapong (a1873765)

2023-03-28

Section00: Executive Summary

This project aimed to develop a new prediction model to enhance the accuracy of evaporation estimation by the Melbourne Water Corporation (MWC), being responsible for managing Melbourne, Australia's water supply. Following methodology, Melbourne Water Corporation could use these results to demonstrate the ability to make predictions for individual days at Cardinia reservoir, and present confidence intervals for these predictions.

First, analysis start with exploratory the outcome of the previous financial year. The Melbourne Water Corporation utilized this data to improve the management of Cardinia Reservoir. The analysis involved bivariate summaries and model selection methodology, resulting in only four significant predictors, including Month, Minimum temperature, Relative humidity, and the interaction between Month and Relative humidity.

The Minimum temperature factors had a positive influence on the level of evaporation, whereas Relative humidity had an opposite impact. The pattern analysis revealed that there was a higher evaporation rate in summer from December to April, particularly in March. However, the rate of evaporation decreased after May, with a significant reduction in evaporation during June.

To evaluate the model's effectiveness, the MWC performed general applications with extreme scenarios. The scenario on January 13, 2020, which had a high minimum temperature and low relative humidity, produced the highest anticipated evaporation. Conversely, the scenario in June, which had a lower minimum temperature and higher relative humidity, demonstrated the lowest amount of evaporation when compared to other scenarios.

The newly developed prediction model by the the Melbourne Water Corporation will help forecast evaporation based on important features to improve water supply management , create the consistency of the water management, and utilize this data to ensure an uninterrupted water supply from upstream, particularly during summer with high minimum temperatures or low relative humidity.

Section01: Methodology

In this section, we focused on how to analyse the Melbourne weather observations dataset using many methods to interpret including Bivariate summaries to find the variable relationship and summary statistics, and model selection following linear summary and ANOVA analysis.

Bivariate summaries

In our analysis, we will focus on following potential influences on amount of evaporation in a day including
- Month
- Day of the week
- Maximum temperature in degrees Celsius
- Minimum temperature in degrees Celsius
- Relative humidity, as measured at 9am.

So, the table will be prepared following the condition above and shown as the table below

Table 1: prepared melbourne evaporation data with related feature with first 5 rows

Month	DayOfWeek	max_temp_c	min_temp_c	relative_humidity	evaporation_mm
1	Tue	26.2	15.5	74	7.0
1	Wed	22.2	18.4	64	7.0
1	Thu	29.5	15.9	75	6.6
1	Fri	42.6	18.0	31	7.8
1	Sat	21.2	17.4	63	15.4

Exploring the relationship between response and each predictors is necessary to understand the background of the dataset and also increasing chance to select the appropriate predictors to our model. Then, the relationship of response with each predictors could visualize the association following the plot below;

Plot evaporation in mm with quantitative predictor

The relationship between quantitative predictors and response could display through the **Scatter plot** following the plot below;

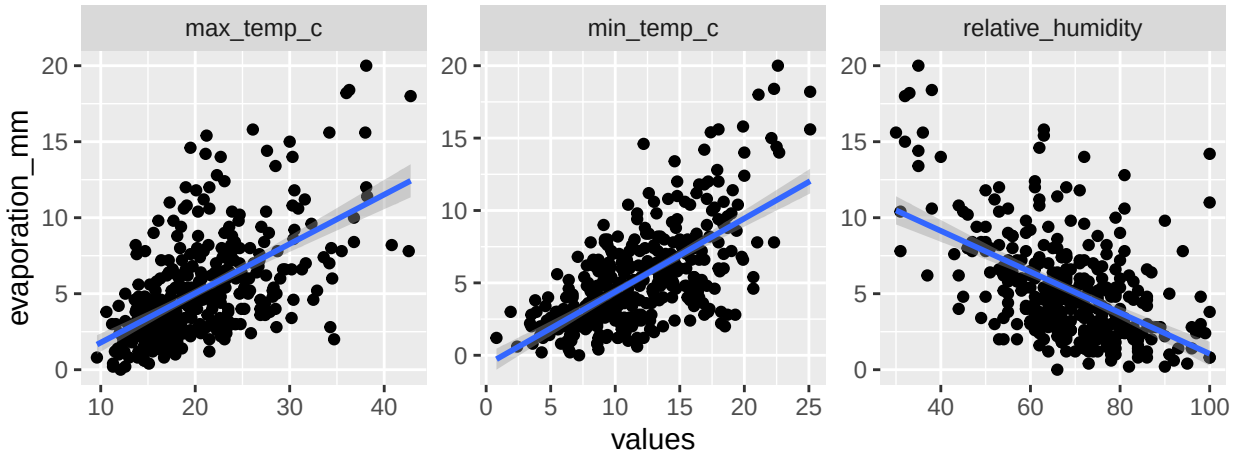

Figure 1: Scatter plot for all potential quantitative predictors

Following the Maximum temperature and Minimum temperature in degrees Celsius, they represent the moderate positive linear relationship. On the other hands, relationship between relative humidity and evaporation interpret the moderate negative values.

Plot evaporation in mm with categorical predictor

The relationship between categorical predictors and response could display through the **Box plot** following the plot below;

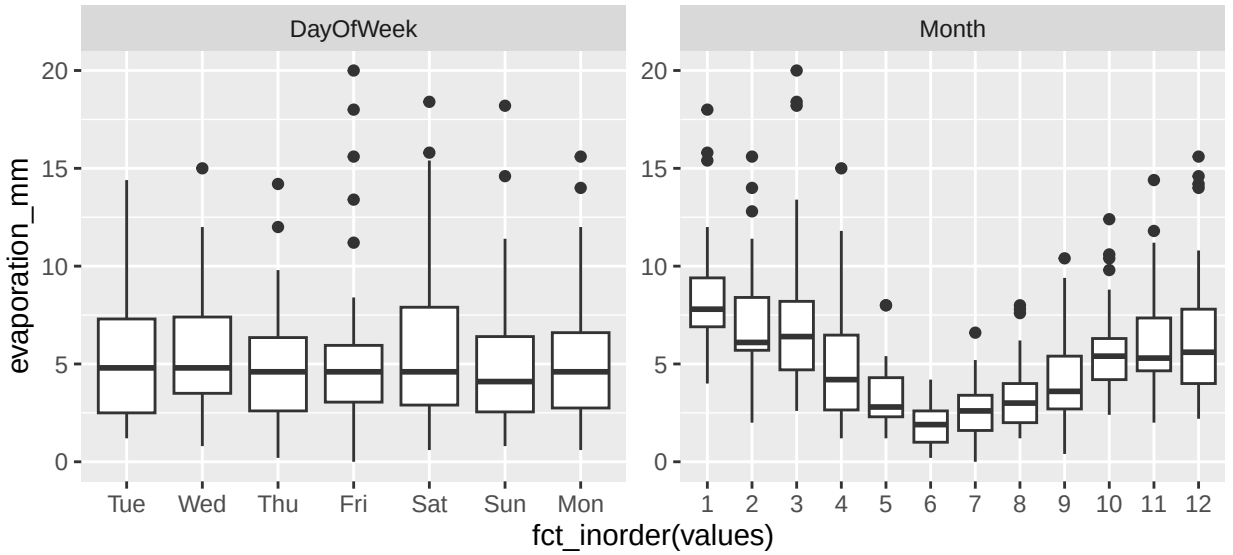

Figure 2: Box plot for all potential categorical predictors

For Day of Week predictor, they show the similar trend of median weight. Following the Month predictor, middle of the year since April until September represent the lower median weight compared to other periods.

Plot included interaction between each predictors

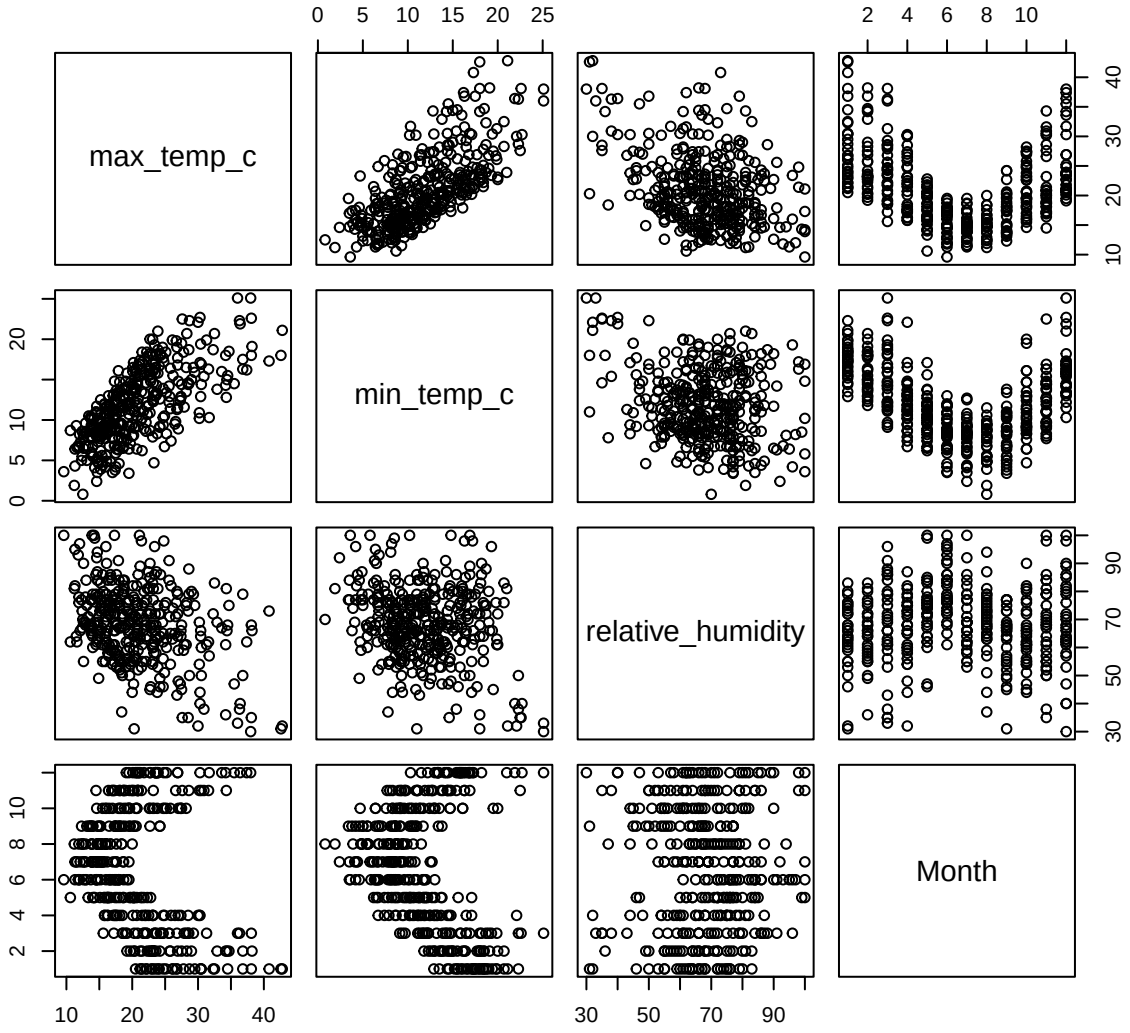

Figure 3: scatter plot for all related predictors

For Relationship between Month and other predictors, both of relationship with minimum temperature and maximum temperature provide the similar pattern compared to relationship of Month and evaporation. However, pattern of Month and relative humidity represent the unique pattern with highest mean average on June.

Model selection

Following the Melbourne dataset, we could build a model to predict the evaporation following the list of predictors from previous Bivariable summaries including with Month, Day of the week, Maximum temperature in degrees Celsius, Minimum temperature in degrees Celsius, and Relative humidity, as measured at 9am. And, also consider an interaction term of Month and

relative humidity. Therefore, this model will be predicted all of predictors with significant effect to evaporation following the procedure:

01 Fit a model containing all the possible predictors

$$\begin{aligned}\text{evaporation\_mm} = {} & \alpha + \beta_1(\text{Month}_2) + \\ & \beta_2(\text{Month}_3) + \beta_3(\text{Month}_4) + \\ & \beta_4(\text{Month}_5) + \beta_5(\text{Month}_6) + \\ & \beta_6(\text{Month}_7) + \beta_7(\text{Month}_8) + \\ & \beta_8(\text{Month}_9) + \beta_9(\text{Month}_{10}) + \\ & \beta_{10}(\text{Month}_{11}) + \beta_{11}(\text{Month}_{12}) + \\ & \beta_{12}(\text{DayOfWeek}_{\text{Mon}}) + \beta_{13}(\text{DayOfWeek}_{\text{Tue}}) + \\ & \beta_{14}(\text{DayOfWeek}_{\text{Wed}}) + \beta_{15}(\text{DayOfWeek}_{\text{Thu}}) + \\ & \beta_{16}(\text{DayOfWeek}_{\text{Fri}}) + \beta_{17}(\text{DayOfWeek}_{\text{Sat}}) + \\ & \beta_{18}(\text{max\_temp\_c}) + \beta_{19}(\text{min\_temp\_c}) + \\ & \beta_{20}(\text{relative\_humidity}) + \beta_{21}(\text{Month}_2 \times \text{relative\_humidity}) + \\ & \beta_{22}(\text{Month}_3 \times \text{relative\_humidity}) + \beta_{23}(\text{Month}_4 \times \text{relative\_humidity}) + \\ & \beta_{24}(\text{Month}_5 \times \text{relative\_humidity}) + \beta_{25}(\text{Month}_6 \times \text{relative\_humidity}) + \\ & \beta_{26}(\text{Month}_7 \times \text{relative\_humidity}) + \beta_{27}(\text{Month}_8 \times \text{relative\_humidity}) + \\ & \beta_{28}(\text{Month}_9 \times \text{relative\_humidity}) + \beta_{29}(\text{Month}_{10} \times \text{relative\_humidity}) + \\ & \beta_{30}(\text{Month}_{11} \times \text{relative\_humidity}) + \beta_{31}(\text{Month}_{12} \times \text{relative\_humidity}) + \\ & \epsilon \end{aligned} \tag{1}$$

02 Determine the p-value for inclusion of each predictor:

1) P-values for quantitative variables can be determined using the linear model summary.

```
##
## Call:
## lm(formula = evaporation_mm ~ Month + DayOfWeek + max_temp_c +
##     min_temp_c + relative_humidity + Month:relative_humidity,
##     data = melbourne_prep)
##
## 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 ***
## Month2                     1.122982   3.341422   0.336 0.737028
## Month3                     5.340251   2.630467   2.030 0.043155 *
## Month4                     1.729320   3.102811   0.557 0.577679
## Month5                    -4.255253   3.347211  -1.271 0.204537
```

```
## Month6                    -7.914716   3.972809  -1.992 0.047183 *
## Month7                    -4.930279   3.580302  -1.377 0.169442
## Month8                    -6.310577   3.222937  -1.958 0.051083 .
## Month9                    -0.544108   3.157664  -0.172 0.863298
## Month10                   -6.307800   3.112895  -2.026 0.043546 *
## Month11                   -1.080420   2.787061  -0.388 0.698525
## Month12                    0.667154   2.793904   0.239 0.811420
## DayOfWeekMon              -0.272388   0.432537  -0.630 0.529304
## DayOfWeekTue              -0.083051   0.436596  -0.190 0.849252
## DayOfWeekWed              -0.078214   0.436180  -0.179 0.857801
## DayOfWeekThu              -0.536148   0.435847  -1.230 0.219539
## DayOfWeekFri              -0.408977   0.443221  -0.923 0.356828
## DayOfWeekSat               0.499638   0.432760   1.155 0.249127
## max_temp_c                 0.017765   0.030507   0.582 0.560738
## min_temp_c                 0.357912   0.044596   8.026 1.86e-14 ***
## relative_humidity         -0.098209   0.032565  -3.016 0.002765 **
## Month2:relative_humidity  -0.026262   0.050976  -0.515 0.606776
## Month3:relative_humidity  -0.080822   0.039559  -2.043 0.041850 *
## Month4:relative_humidity  -0.043164   0.047080  -0.917 0.359914
## Month5:relative_humidity   0.034968   0.047799   0.732 0.464966
## Month6:relative_humidity   0.078436   0.052691   1.489 0.137560
## Month7:relative_humidity   0.049674   0.051370   0.967 0.334276
## Month8:relative_humidity   0.079397   0.047371   1.676 0.094686 .
## Month9:relative_humidity  -0.006753   0.049154  -0.137 0.890813
## Month10:relative_humidity  0.092502   0.047400   1.952 0.051853 .
## Month11:relative_humidity  0.015097   0.041694   0.362 0.717527
## Month12:relative_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
```

2) P-values for categorical variables, or interactions containing categorical variables, can be determined using an ANOVA.

03 Remove the predictor with the highest p-value for inclusion, unless all remaining predictors are significant at the 5

Following the determining the p-value for inclusion, there are two methods to remain only significant predictors including P-values for quantitative variables and P-values for categorical variables. For quantitative variables, max_temp_c represent the highest P-value, 0.56074, determined using the linear model summary. For categorical variables, DayOfWeek show the highest P-value equal to 0.1025018 in ANONA analysis. Therefore, these predictors will be removed to remain predictors having significant at the 5.

Table 2: Model selection for finding the appropriate feature - first loop

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Month	11	1478.84780	134.440709	28.428788	0.0000000
DayOfWeek	6	50.50824	8.418039	1.780076	0.1025018
max_temp_c	1	279.65057	279.650568	59.134817	0.0000000
min_temp_c	1	383.82986	383.829855	81.164534	0.0000000
relative_humidity	1	448.57149	448.571492	94.854779	0.0000000
Month:relative_humidity	11	160.95415	14.632196	3.094119	0.0005645
Residuals	325	1536.93610	4.729034	NA	NA

04 Update your model to include only the remaining predictors.

$$
\begin{aligned}
\text{evaporation\_mm} = \alpha &+ \beta_1(\text{Month}_2) + \\
&\beta_2(\text{Month}_3) + \beta_3(\text{Month}_4) + \\
&\beta_4(\text{Month}_5) + \beta_5(\text{Month}_6) + \\
&\beta_6(\text{Month}_7) + \beta_7(\text{Month}_8) + \\
&\beta_8(\text{Month}_9) + \beta_9(\text{Month}_{10}) + \\
&\beta_{10}(\text{Month}_{11}) + \beta_{11}(\text{Month}_{12}) + \\
&\beta_{12}(\text{min\_temp\_c}) + \beta_{13}(\text{relative\_humidity}) + \\
&\beta_{14}(\text{Month}_2 \times \text{relative\_humidity}) + \beta_{15}(\text{Month}_3 \times \text{relative\_humidity}) + \\
&\beta_{16}(\text{Month}_4 \times \text{relative\_humidity}) + \beta_{17}(\text{Month}_5 \times \text{relative\_humidity}) + \\
&\beta_{18}(\text{Month}_6 \times \text{relative\_humidity}) + \beta_{19}(\text{Month}_7 \times \text{relative\_humidity}) + \\
&\beta_{20}(\text{Month}_8 \times \text{relative\_humidity}) + \beta_{21}(\text{Month}_9 \times \text{relative\_humidity}) + \\
&\beta_{22}(\text{Month}_{10} \times \text{relative\_humidity}) + \beta_{23}(\text{Month}_{11} \times \text{relative\_humidity}) + \\
&\beta_{24}(\text{Month}_{12} \times \text{relative\_humidity}) + \epsilon
\end{aligned}
\tag{2}
$$

```
##
## Call:
## lm(formula = evaporation_mm ~ Month + min_temp_c + relative_humidity +
##     Month:relative_humidity, data = melbourne_prep)
##
## 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 ***
## Month2                     0.822148   3.297575   0.249 0.803268
## Month3                     5.263051   2.610525   2.016 0.044596 *
## Month4                     1.971572   3.040391   0.648 0.517136
## Month5                    -4.377344   3.261415  -1.342 0.180461
## Month6                    -8.376118   3.924447  -2.134 0.033547 *
```

```
## Month7                     -5.360039   3.479608  -1.540 0.124412
## Month8                     -7.102852   3.189591  -2.227 0.026625 *
## Month9                     -1.243475   3.090815  -0.402 0.687712
## Month10                    -6.158396   3.068813  -2.007 0.045585 *
## Month11                    -1.036904   2.737218  -0.379 0.705066
## Month12                     0.926791   2.748164   0.337 0.736149
## min_temp_c                  0.368846   0.041819   8.820  < 2e-16 ***
## relative_humidity          -0.099750   0.031724  -3.144 0.001815 **
## Month2:relative_humidity   -0.021806   0.050276  -0.434 0.664760
## Month3:relative_humidity   -0.079813   0.039166  -2.038 0.042360 *
## Month4:relative_humidity   -0.047469   0.046050  -1.031 0.303377
## Month5:relative_humidity    0.035145   0.046597   0.754 0.451246
## Month6:relative_humidity    0.083313   0.052006   1.602 0.110113
## Month7:relative_humidity    0.054069   0.050199   1.077 0.282219
## Month8:relative_humidity    0.089054   0.047045   1.893 0.059234 .
## Month9:relative_humidity    0.003411   0.048049   0.071 0.943452
## Month10:relative_humidity   0.089443   0.046676   1.916 0.056194 .
## Month11:relative_humidity   0.013451   0.040881   0.329 0.742336
## Month12:relative_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
```

Table 3: Model selection for finding the appropriate feature - first loop

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Month	11	1478.8478	134.440709	28.416038	0.0000000
min_temp_c	1	608.9318	608.931752	128.706760	0.0000000
relative_humidity	1	510.0328	510.032773	107.802993	0.0000000
Month:relative_humidity	11	170.7421	15.522010	3.280807	0.0002758
Residuals	332	1570.7438	4.731156	NA	NA

The p values in regression analysis play a crucial role in assessing whether the relationships observed in the sample data hold true for the entire population. The linear regression p value for each independent variable tests the null hypothesis that the variable has no correlation with the dependent variable. If there is no correlation, it is not possible to confirm that there is any impact on the population level based on the available evidence. On the other hand, the p-value is less than your significance level, your sample data provide enough evidence to reject the null hypothesis for the entire population.

The final model selection process identified several important predictors, such as Month, Minimum temperature in Celsius, Relative humidity, and the interaction between Month and Relative humidity. However, unlike in the bivariate analyses, Maximum temperature and day of the week were not considered significant predictors based on their P-values and were therefore dropped from the

model.

Model diagnostics

It is essential to validate all the model assumptions of the Linear Regression model, which includes defining the functional form. If any of these assumptions are not met, it is necessary to review and revise the model. There are four assumptions that we have made following;

Check the linearity assumption

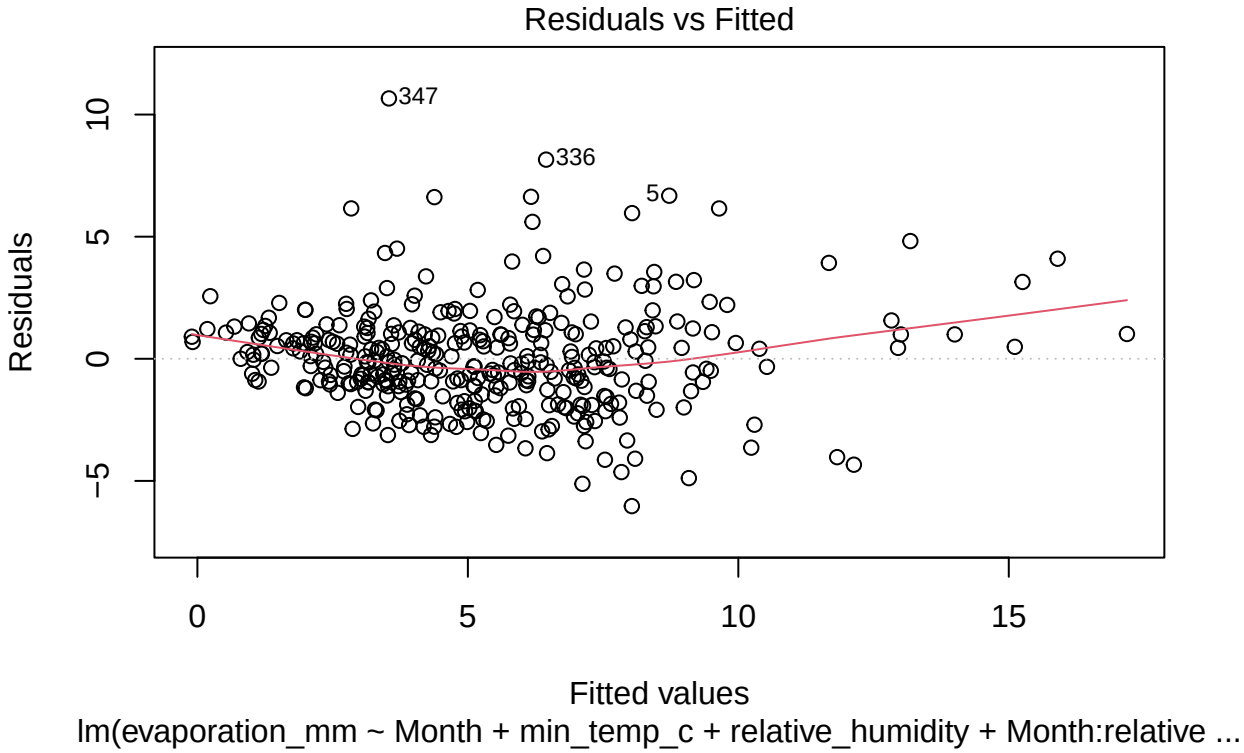

Figure 4: residual versus fitted plot - check Linearity on evaporation model

The assumption of linearity justified because the red reference present the mostly straight and only few points are threw off the guideline.

Check the homoscedasticity assumption

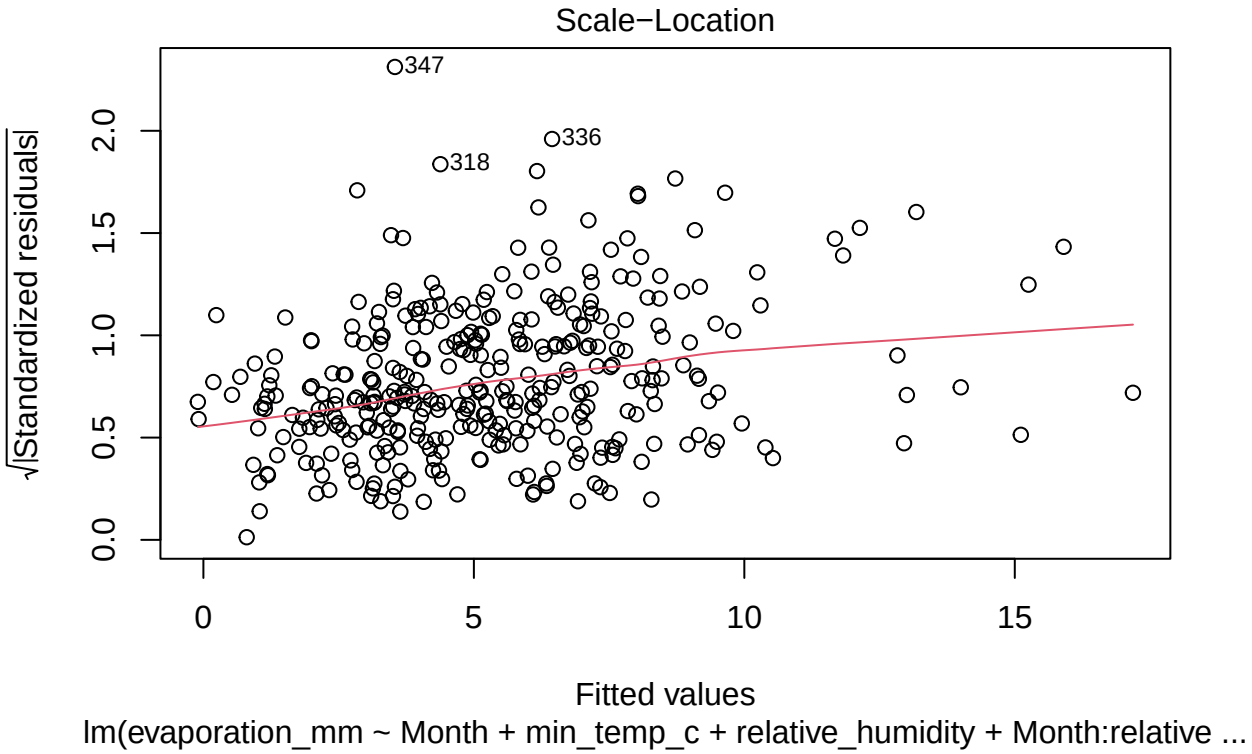

Figure 5: square root of the standardised residual versus fitted plot - check homoscedasticity on evaporation model

The assumption of homoscedasticity justified because there are no apparent trends and the reference line visual the roughly straight and little flat. Therefore, it could indicate that this model constantly spread.

Check the Normality assumption

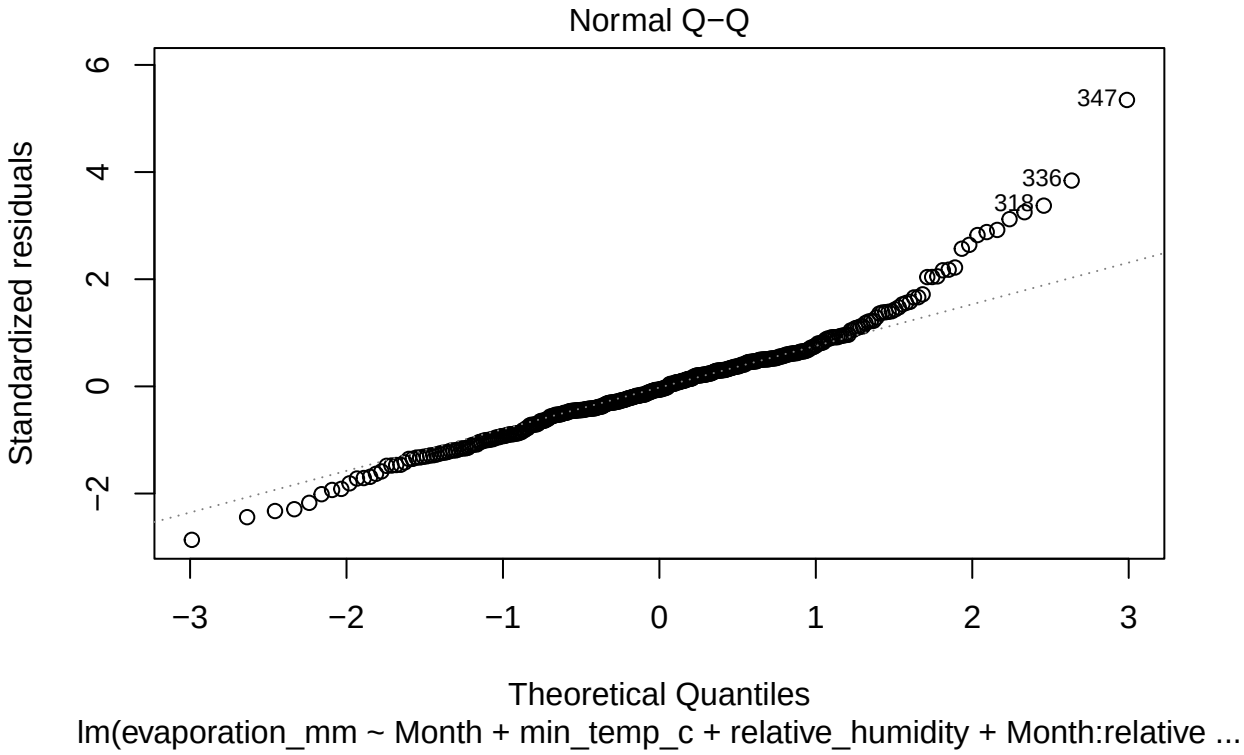

Figure 6: normal QQ-plot - check Normality on evaporation model

The assumption of Normality justified because the points should lie along the dotted line. And, there are few points that lie less than -1.5 on the x-axis and greater than 1.5 on axis drift away from the reference line. That are not significant because of minority of that group.

Check the independence assumption

Following the information due to Melbourne's weather observations data set, there are a few possible dependence impacts between each data points including effect of the weather on the curernt date will affect to coming day or specific location to collect all the data points. Therefore, it is the limitation that need to acknowledge and prevent to decide with biased and unreliable estimates of the regression coefficients.

Section02: Results

Model interpretation

In this section, a model need to interpret the coefficients related to each predictor. For the intercept of model, it mean that when all of the predictors are zero with reference category on January, the expected value of evaporation in mm is 8.589. For increasing the minimum temperature by one degrees Celsius, the evaporation with other zero predictors will increase by 0.369 mm. On the other hands, increasing the relative humidity by one unit will decrease evaporation 0.1 mm.

Following categorical predictors, there are two variables put in to the model including Month and interaction between month and relative humidity.
For Month predictor, there are eleven coefficients following the **Month01** as the reference category. As the coefficient of **Month06** equal to -8.376, it presents the estimate the mean weight of evaporation on Month06 being the lowest value lower than mean weight of evaporation on Month01. However, the mean weight of evaporation on **Month03**, 5.263, being the biggest value higher than mean weight of evaporation on Month01. With **P-value of Month6 and Month3** equal to **0.034 and 0.045** respectively, the indicators represent the significant difference compared to Month01.
Finally, if we select the **Month03** with assumption that others predictors equal to zero, the evaporation will be **intercept + coefficient on Month3** as **8.589 + 5.263 = 13.852** mm.

For interaction between month and relative humidity, there are also eleven coefficients following **Month01 interacted with relative humidity** as the reference category. Following the coefficient of **Month10:relative_humidity** equal to 0.089443, it presents the estimate the mean weight of evaporation being the biggest value higher than mean weight of evaporation on Month01 with relative_humidity. Meanwhile, **interaction term between Month03 and relative humidity**, -0.800 , is the lowest value less than mean weight of evaporation on Month01 with relative_humidity.
For example, if we select the **Month10** assuming that **relative humidity equal to 50 and others are zero**, the evaporation will be equal to **intercept + (coefficients of relative humidity + coefficients of Month10:relative humidity) X relative humidity** = **8.589 + (-0.100 + 0.089) x 50** = **8.039** mm.

Prediction

MWC is keenly interested in utilizing your model in various scenarios, including extreme ones, and would like your predictions on the amount of evaporation, measured in millimeters, for specific types of days described below. Use case 01: February 29, 2020, if this day has a minimum temperature of 13.8 degrees and reaches a maximum of 23.2 degrees, and has 74% humidity at 9am.
Use case 02: December 25, 2020, if this day has a minimum temperature of 16.4 degrees and reaches a maximum of 31.9 degrees, and has 57% humidity at 9am.
Use case 03: January 13, 2020, if this day has a minimum temperature of 26.5 degrees and reaches a maximum of 44.3 degrees, and has 35% humidity at 9am.
Use case 04: July 6, 2020, if this day has a minimum temperature of 6.8 degrees and reaches a maximum of 10.6 degrees, and has 76% humidity at 9am.

Following the model Prediction, results represent the difference outcome from different following the table above;
On **02 Feb, 2020** , it show the expected evaporation, **5.506** mm with confidence interval range from **1.089** mm to **9.923** mm.
On **25 Dec, 2020** , it show the expected evaporation, **8.606** mm with confidence interval range from **4.209** mm to **13.003** mm.

Table 4: Predicted table results with prediction intervals

Cases	Date	Expected evaporation(mm)	Lower Boundary Interval - evaporation(mm)	Upper Boundary Interval - evaporation(mm)
01	2020-02-29	5.506144	1.089134	9.923155
02	2020-12-25	8.605772	4.208554	13.002990
03	2020-01-13	14.872285	10.104616	19.639953
04	2020-07-06	2.265493	-2.111493	6.642478

On **13 Jan, 2020** , it show the expected evaporation, **14.872** mm with confidence interval range from **10.105** mm to **19.64** mm.
On **06 July, 2020** , it show the expected evaporation, **2.265** mm with confidence interval range from **-2.111** mm to **6.642** mm.

Section03: Discussion

Interpreting the results from the prediction, it show the lowest expected evaporation, **2.265 mm** with confidence interval range from **-2.111 mm** to **6.642 mm**, on 06 July 2020. However, predicted result On 13 Jan 2020 is the highest expected evaporation equal to **14.872 mm** with confidence interval range from **10.105 mm** to **19.640 mm**.

In the cases that the evaporation level at MWC's Cardinia Reservoir exceeds 10mm, the corporation implements temporary measures to ensure a consistent water supply. **On 13 Jan 2020**, we could say with 95% confidence to transferring water from its Silvan Reservoir, located upstream. However, this will not occur on other predicted days.

Section04: Conclusion

The objective of this projects is creating the new prediction model to perform the better accuracy of evaporation estimation by the Melbourne Water Corporation ('MWC'), the organization responsible for managing Melbourne, Australia's water supply has been tasked with creating a report that pertains to evaporation. After interpreting the outcome following the information on previous financial year, Melbourne Water Corporation utilizes this data to enhance the management of Cardinia Reservoir in the southeastern region of the city.

Following Bivariate summaries and model selection methodology, there are only four remaining significant predictors including with Month, Minimum temperature, Relative humidity, and interaction between Month and Relative humidity. For Minimum temperature predictors, it present the positive effect to amount of evaporation. However, relative humidity absolutely show the opposite impact. The pattern of the month from December to April, in the Summer, shows a higher rate of evaporation, especially in March. However, the rate of evaporation decreases after May, with significantly less evaporation occurring in June. Nonetheless, it is important to acknowledge that the independent assumption has certain limitations that should be taken into account to avoid making misleading decisions.

To analyze and evaluate model, the Melbourne Water Corporation perform the general application with extreme scenarios to measure the outcome. The case on 13 January 2020 with high minimum temperature and less relative humidity show the highest expected evaporation. While, the case during June having less minimum temperature and high relative humidity represent the lowest amount of evaporation compared to other scenarios.

In conclusion, Melbourne Water Corporation has developed a new predictive model that can forecast evaporation based on significant features to enhance water supply management and ensure the consistency of the water supply in the city. The data from this model can be utilized to guarantee an uninterrupted water supply from upstream, especially during summer when there are high minimum temperatures or low relative humidity.

Appendix

```
knitr::opts_chunk$set(echo = FALSE
                      , message = FALSE
                      , warning = FALSE
                      , fig.pos = 'H'
                      , out.extra = ''
                      , collapse = TRUE
                      )
## Methodology: library selection
library(tidyverse)
# library(skimr)
# library(mlbench)
library(knitr)
library(kableExtra)
library(equatiomatic)
library(corrplot)
## Methodology: import dataset
melbourne <- read.csv("sc/melbourne.csv")
## Methodology: Bivariate summaries
melbourne_prep <- melbourne %>%
  mutate(Month = factor(month(Date), ordered = FALSE),
         DayOfWeek = factor(wday(Date, label = TRUE), ordered = FALSE)
         ) %>%
  rename(max_temp_c = "Maximum.Temperature..Deg.C.",
         min_temp_c = "Minimum.temperature..Deg.C.",
         relative_humidity = "X9am.relative.humidity....",
         evaporation_mm = "Evaporation..mm."
         ) %>%
  select(Month,
         DayOfWeek,
         max_temp_c,
         min_temp_c,
         relative_humidity,
         evaporation_mm
         )
kable(head(melbourne_prep,5), caption = "prepared melbourne evaporation data with related featur
  kable_styling(position = "left", latex_options = "hold_position")
# Bivariate summaries: quantitative predictors
melbourne_prep %>%
  select(max_temp_c,
         min_temp_c,
         relative_humidity,
         evaporation_mm
         ) %>%
  gather(-evaporation_mm, key = "predictors", value = "values") %>%
  ggplot(aes(x = values, y = evaporation_mm)) +
```

```
  geom_point() +
  geom_smooth(method='lm') +
  facet_wrap(vars(predictors), scales = "free")
# Bivariate summaries: categorical predictors
melbourne_prep %>%
  select(Month,
         DayOfWeek,
         evaporation_mm
         ) %>%
  gather(-evaporation_mm, key = "predictors", value = "values") %>%
  ggplot(aes(x = fct_inorder(values), y = evaporation_mm)) +
  geom_boxplot() +
  facet_wrap(vars(predictors), scales = "free")
# Bivariate summaries: multiple scatter plots
pairs(melbourne_prep[, c("max_temp_c", "min_temp_c", "relative_humidity","Month")])
# Model selection: first loop - fit model
evaporation_model <- lm(evaporation_mm ~
                          Month +
                          DayOfWeek +
                          max_temp_c +
                          min_temp_c +
                          relative_humidity +
                          Month:relative_humidity
                        , data = melbourne_prep)
extract_eq(evaporation_model, wrap = TRUE, terms_per_line = 2)
# Model selection: first loop - Determine the p-value - linear model summary
summary(evaporation_model)
# Model selection: first loop - Determine the p-value - ANONA analysis
kable(anova(evaporation_model), caption = "Model selection for finding the appropriate feature -
  kable_styling(position = "left", latex_options = "hold_position")
# Model selection: first loop - fit model
evaporation_model <- lm(evaporation_mm ~
                          Month +
                          min_temp_c +
                          relative_humidity +
                          Month:relative_humidity
                        , data = melbourne_prep)
extract_eq(evaporation_model, wrap = TRUE, terms_per_line = 2)
# Model selection: check loop - Determine the p-value - linear model summary
summary(evaporation_model)
# Model selection: check loop - Determine the p-value - ANONA analysis
kable(anova(evaporation_model), caption = "Model selection for finding the appropriate feature -
  kable_styling(position = "left", latex_options = "hold_position")
## Model diagnostics - check assumption Linearity
plot(evaporation_model, which = 1)
## Model diagnostics - check assumption homoscedasticity
plot(evaporation_model, which = 3)
```

```
## Model diagnostics - check assumption Normality
plot(evaporation_model, which = 2)
## Model Prediction - create the use case dataset
input_usecase <- tibble(
  Month = as.factor(month(as.Date(c("2020-02-29","2020-12-25","2020-01-13","2020-07-06")))),
  min_temp_c = c(13.8,16.4,26.5,6.8),
  max_temp_c = c(23.2,31.9,44.3,10.6),
  relative_humidity = c(74,57,35,76)
)

## Model Prediction - create the prediction table
predict_evaporation <- data.frame(Cases = c('01','02','03','04'),
                                  Date = c("2020-02-29","2020-12-25","2020-01-13","2020-07-06"),
                                  as.tibble(predict(evaporation_model, input_usecase, interval =
                                  ) %>%
  rename('Lower Boundary Interval - evaporation(mm)' = "lwr",
         'Upper Boundary Interval - evaporation(mm)' = "upr",
         'Expected evaporation(mm)' = "fit"
         )

## Model Prediction - visualize the predicted table output
kable(predict_evaporation, caption = "Predicted table results with prediction intervals") %>%
  kable_styling(position = "left", latex_options = "hold_position") %>%
  column_spec(3:5, width = "3.9cm")
```