Assessment 3- CASE STUDY

Name Zahra Ahmadpour

Student ID A1863476

Course Data Taming, Modelling and Visualization

Number of words excluding references 2784

Table of Contents

1. Executive summary

In this research, evaporation (mm) in the Cardina reservoir in the city's South East has been predicted with linear regression, so that in case of evaporation is more than 10 mm in MWC's Cardina reservoir, suitable solutions, including water transfer from Silvan reservoir, for continuous water supply.

According to the statistical calculations with R analysis of the linear regression of evaporation based on the month, minimum temperature and humidity, the results obtained based on the 95% confidence interval showed that in January 2020, the amount of evaporation at the lowest level is more than 10 (mm) that necessary arrangements should be made to continuous supply.

According to the evaporation's predictions in December, February and July, which the evaporation at the highest level are less than 10 mm, so there is no concern for water supply in the remaining months of 2020.

2. Methods

- **Dataset:**

The dataset of the analysis is the Melbourne weather observations, for all days of year 2019. The dataset is obtained from the Bureau of meteorology's "real time" of Australian government website. [1]

The observations of this data set are automatically generated and managed, so there may be some errors in the data and it is hard to determine the reliability of the dataset. If the data is not available for various reasons, the data is calculated and replaced based on the following days. [1]

- **Software:**

In this analysis, I have used R software, which is currently one of the most common software in data analysis. the features of R programming are:

1. R is free of cost and it has a 10000 different packages that help to analysis the data.
2. R is common in data analysis and it has strong graphic for data visualization.
3. R perform complex analysis on big data.
4. R is suitable for machine learning as well.

and too many other features. [2]

I used these packages: modelr, lubridate, DT, tidyr, stringr, tidyverse, gutenbergr, caret, mlbench, inspectdf, readr, nycflights13, moments, correlation, Hmisc, car, forecast, knitr.

- **Data preparation:**

In this project, I have examined the effect of day of the week, month, maximum temperature, minimum temperature and 9am related humidity on the Evaporation of the Cardina reservoir in the city's South East by using a linear relationship. Then I add the columns, day of the week and month exported from column Date and remove the unnecessary columns from dataset.

Table 1: Dataset description

Name		Description	Unit
Date		Include month, day and year	M/D/Y
Day of the week		Sunday, Monday, …	1, 2, …,7
Month		----	1,2,….,12
Temperature	Minimum	Minimum temperature in the 24 hours to 9am	Deg C
	Maximum	Maximum temperature in the 24 hours from 9am	
Evaporation		Evaporation in the 24 hours to 9am	mm
9 am Relative humidity		Relative humidity at 9 am	%

- **Univariate analysis:**

To check the skewness of continuous variables that could be resolved by a log transformation, first I will produce the histogram plot for each numeric variable to see the skewness of them then I will review the logarithm transformation if it is needed.

1. **Evaporation (mm):**

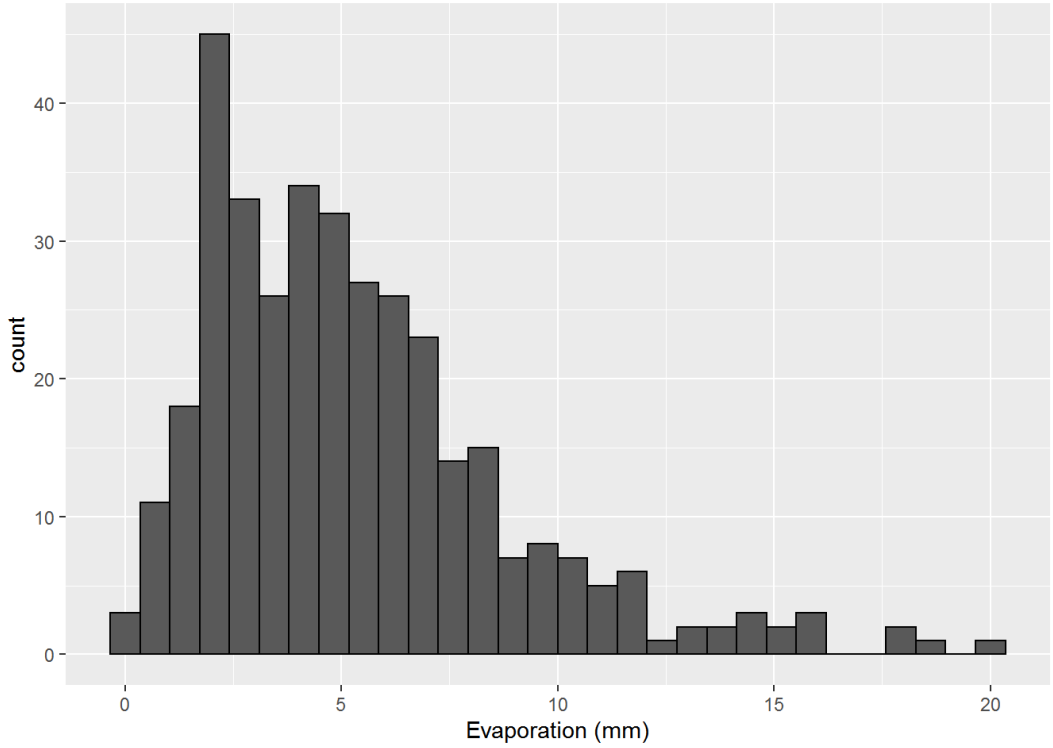


Figure 1: Histogram plot of Evaporation (mm)

The shape is right skewed because mean > median and (skewness=1.329)>1 (high skewed) then I will check the log transformation of evaporation:

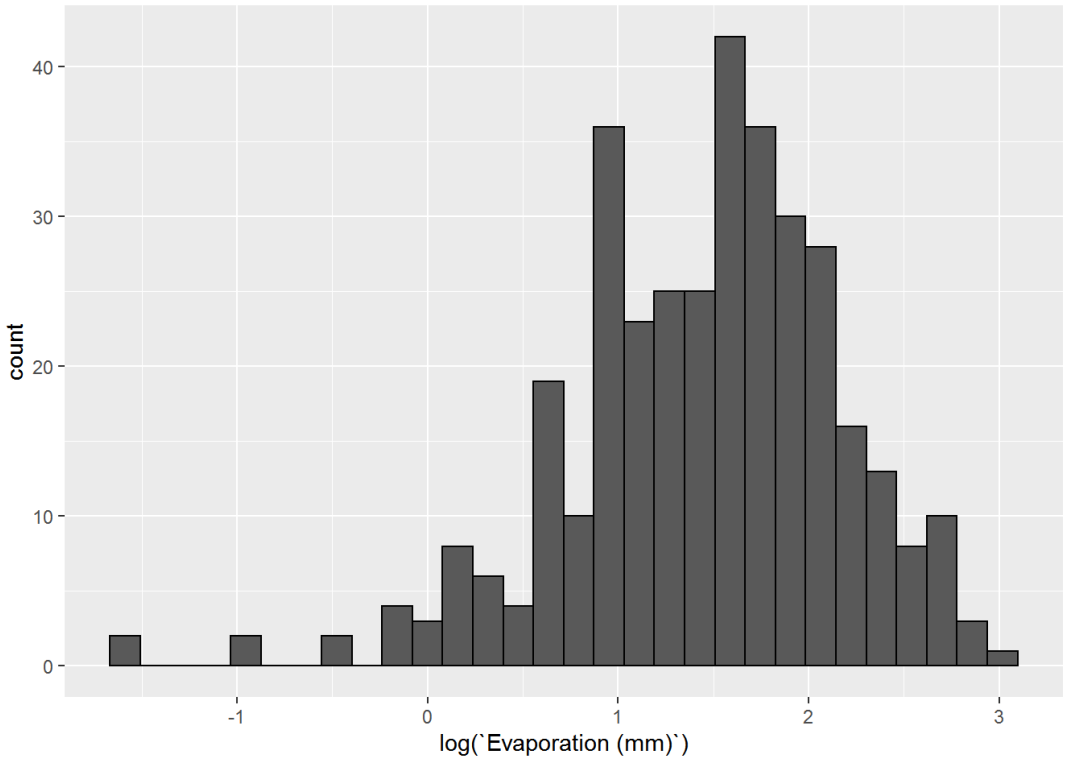


Figure 2: Histogram plot of log (Evaporation (mm))

After using the logarithm transformation, the shape is less skewed left skewed (median > mean), and the skewness decreased to -0.78 (moderate skewed), then I will use a logarithm of evaporation.

2. Maximum Temperature (Deg C):

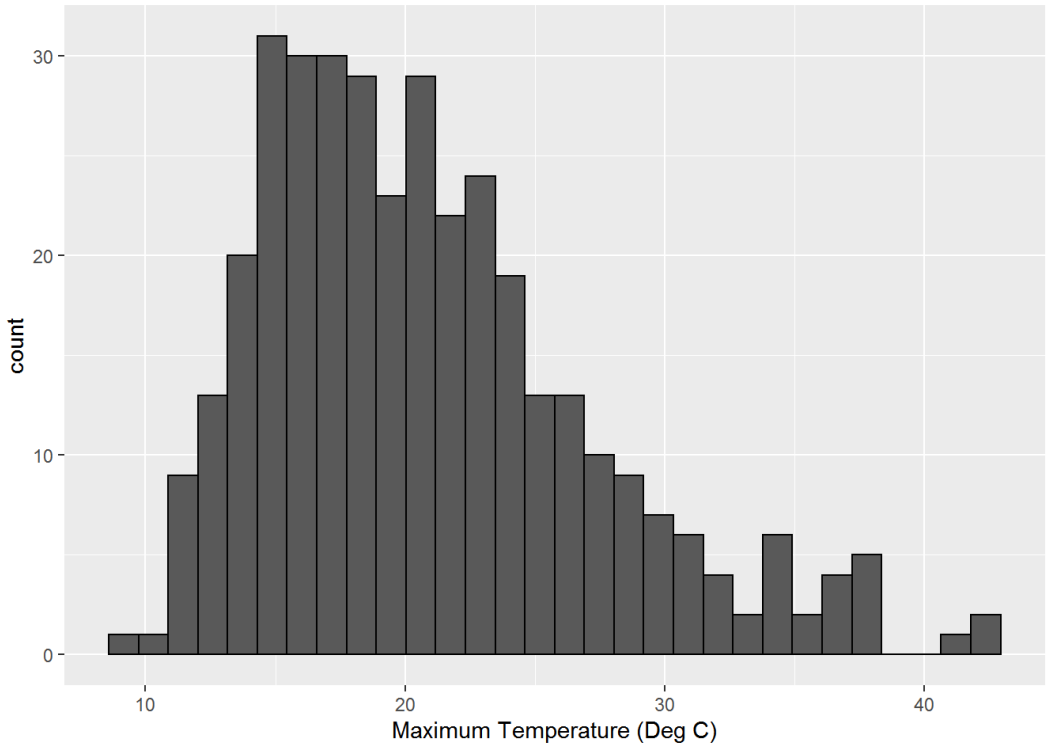


Figure 3: Histogram plot of Maximum temperature (Deg C)

The shape is right skewed because mean > median and 0.5< (skewness=0.942) <1 (moderate skewed), then I will check a log transformation of Maximum Temperature:

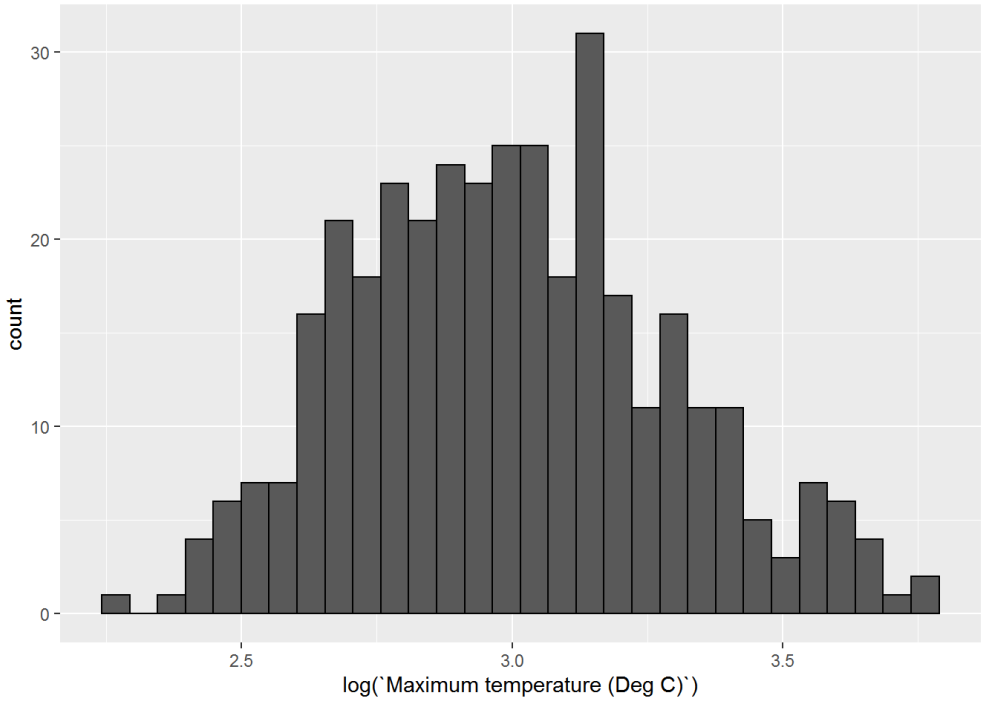


Figure 4: Histogram plot of log (Maximum temperature (Deg C))

After the logarithm transformation, the shape is almost symmetric and the skewness decreased to 0.24 (fairly symmetric), then I will use a logarithm of maximum Temperature.

3. Minimum temperature (Deg C):

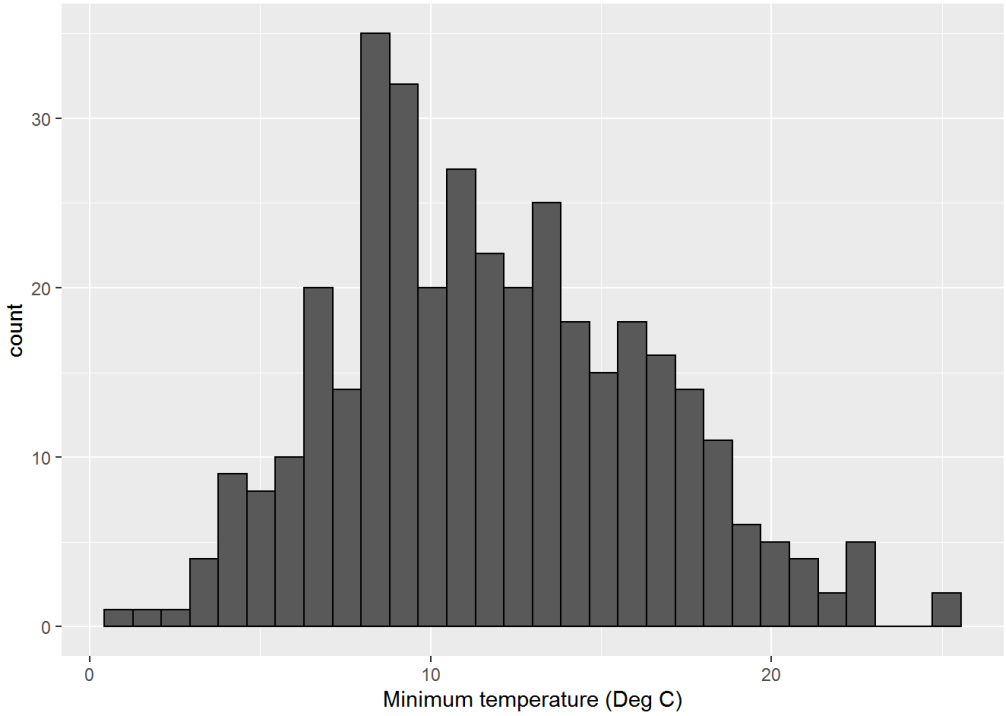


Figure 5: Histogram plot of Minimum temperature (Deg C)

The shape is fairly symmetric and the skewness is 0.31, then I will not add a log transformation.

4. 9am relative humidity (%):

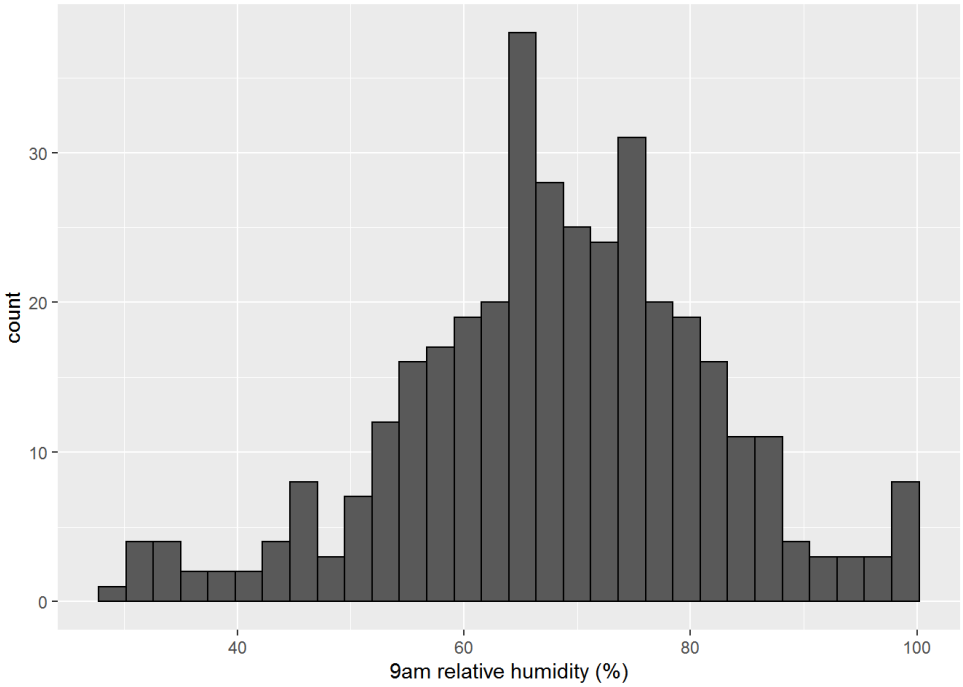


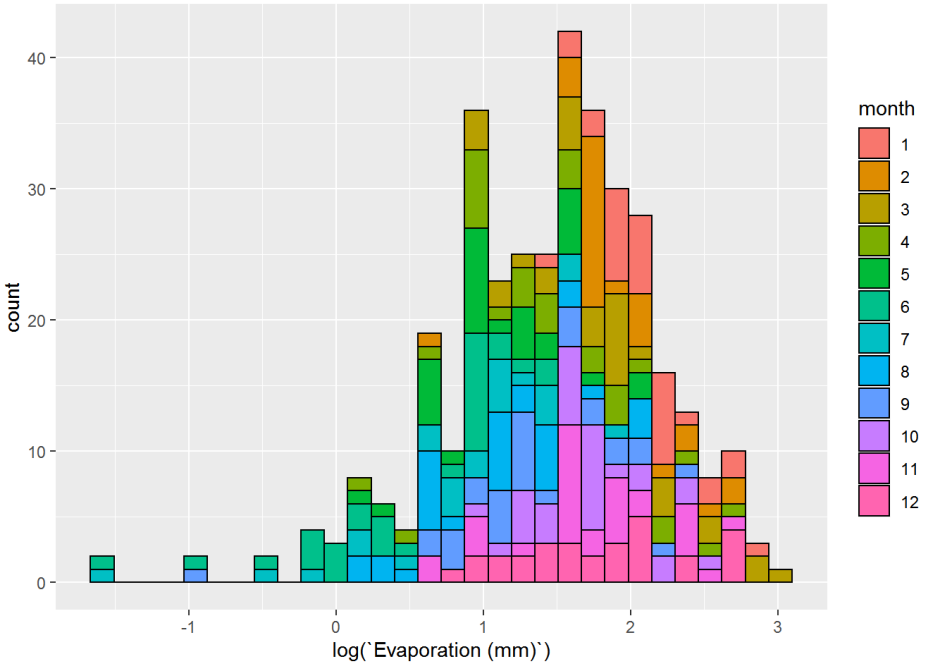
Figure 6: Histogram plot of 9am relative humidity (%)

The shape is almost symmetric and the skewness is -0.27, then I will not add a log transformation.

- **Bivariate analysis:**

In the bivariate analysis section, the linear relationship between the independent variables and the dependent variable was investigated, and the results are as follows:

1. Log (Evaporation (mm)) vs. month:



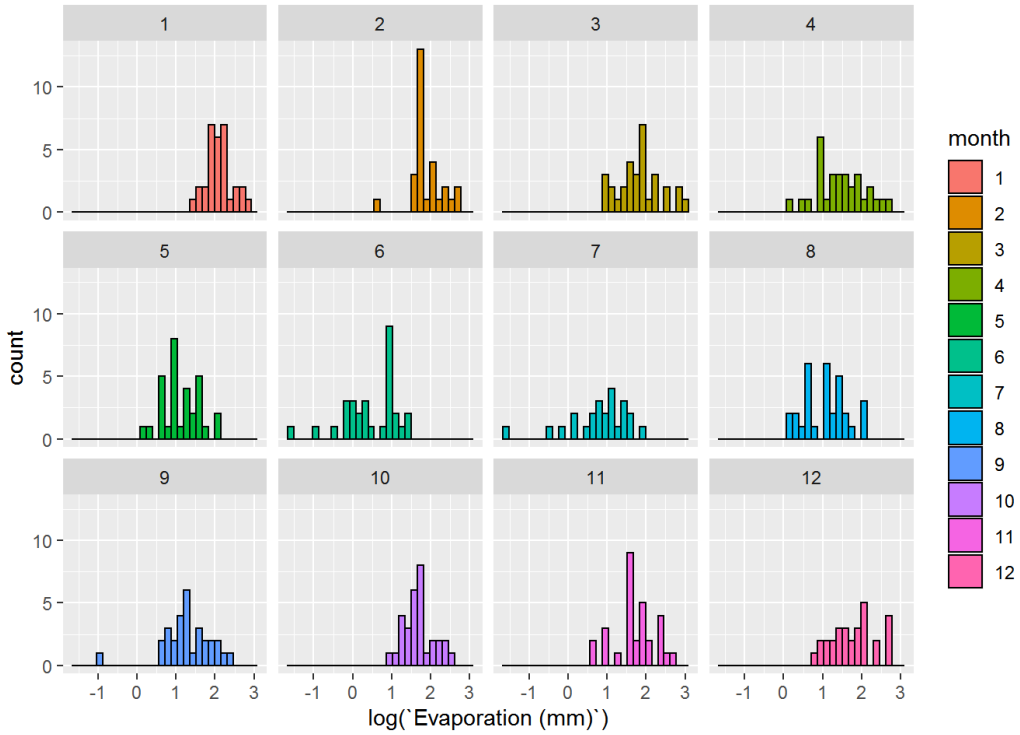


Figure 7: Histogram of log (Evaporation(mm)) by month

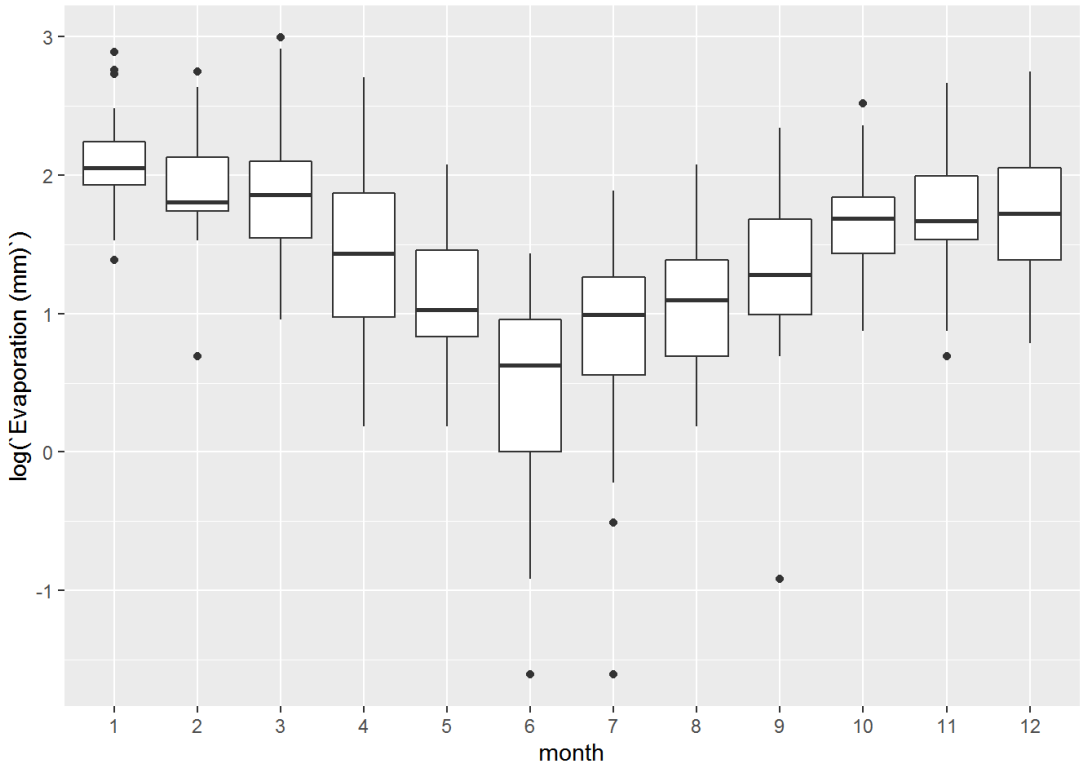


Figure 8: Box plot of log (Evaporation(mm)) against month

|r| = -0.12 (0.1 < |r| < 0.3) then there is a weak correlation between log (`Evaporation (mm)`) and month.

- **Shape:** It look likes the left-skewed because mean < median and also it might be multimodal. If check them one by one, January and April are almost right skewed and other months are left skewed and some almost symmetric.
- **Location:** The median is the line in the middle of each rectangle, we can see that January has the highest Evaporation, while Jun has the lowest. The maximum median is 2.1 (January) and minimum median is 0.2 (Jun).
- **Spread:** Maximum spread (height of the rectangle) is in Jun and minimum spread is in January.
- **Outliers:** We can see 4 outliers in January, 2 outliers in each month of Feb and July, 1 outlier in each month of Mar, Jun, Sep, Oct and Nov.

2. Log (Evaporation (mm)) vs. Day of the week:

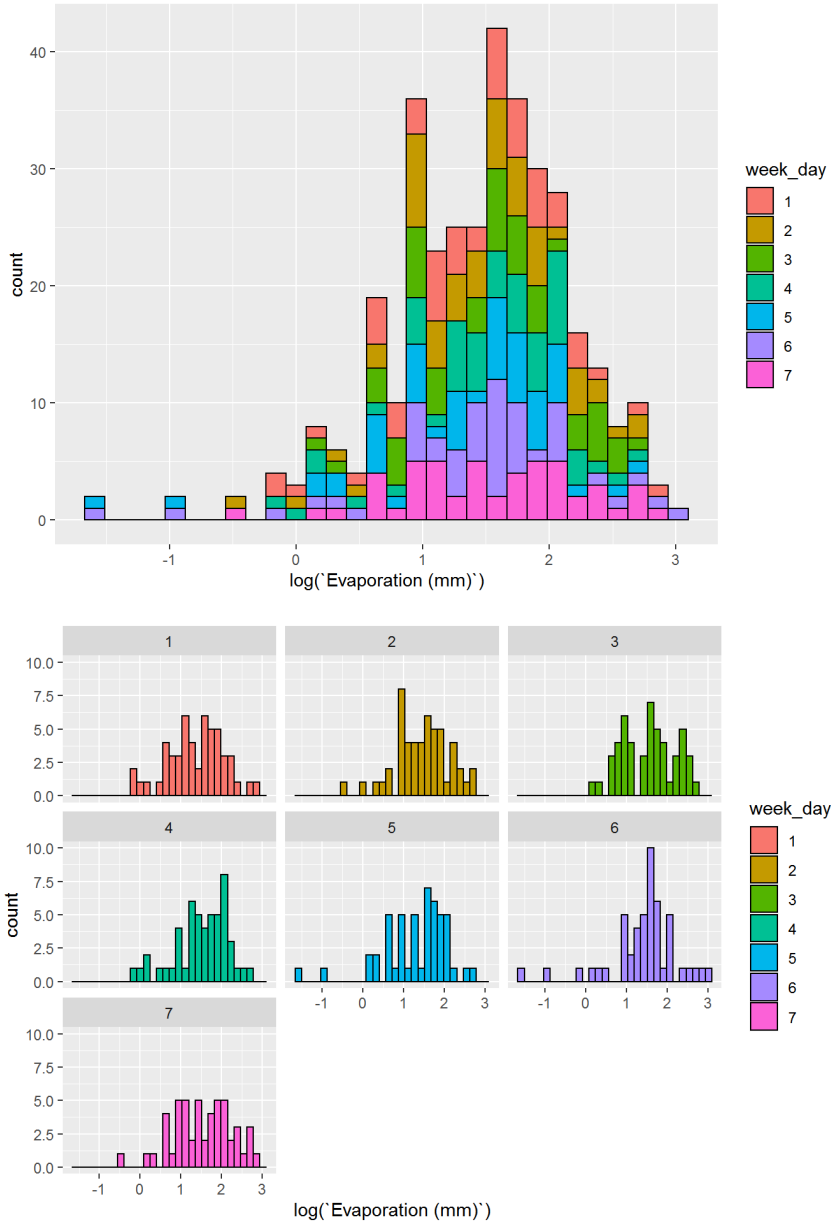


Figure 9: Histogram of log (Evaporation(mm)) by day of week

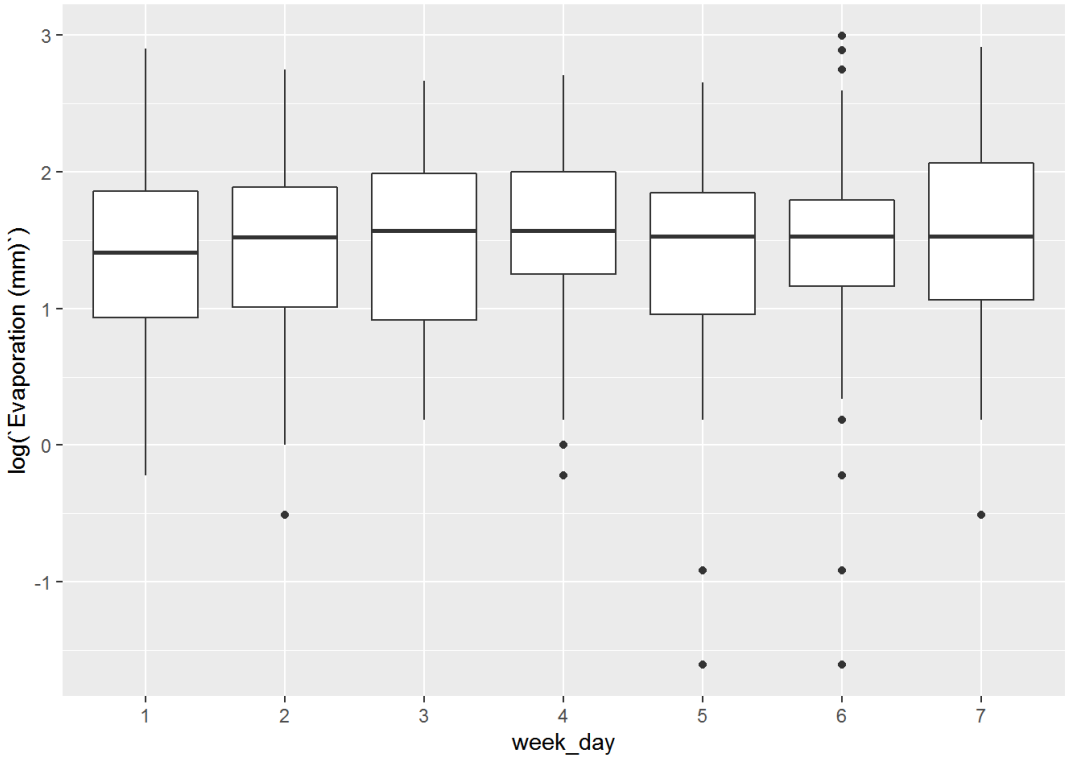


Figure 10: Box plot of log (Evaporation (mm)) against Day of week

|r| = -0.02 (|r| < 0.1) then there is not any correlation between log (`Evaporation (mm)`) and week of day. then it should be removed from linear model.

- **Shape:** It look likes the left-skewed because mean < median and also it might be multimodal. If check them one by one, Wednesday, Thursday and Friday are left skewed, Monday is right skewed and other days are almost symmetric.
- **Location:** The median is the line in the middle of each rectangle, we can see that Sunday has the lowest Evaporation, while Tuesday and Wednesday have the highest. The maximum median is 0.6 (Tuesday and Wednesday) and minimum median is 0.4 (Sunday).
- **Spread:** Maximum spread (height of the rectangle) is in Monday and minimum spread is in Friday.
- **Outliers:** We can see 7 outliers in Friday, 2 outliers in Wednesday and Thursday, 1 outlier in Monday and Saturday.

3. **Log (Evaporation (mm)) vs. log (Maximum temperature (Deg C)):**

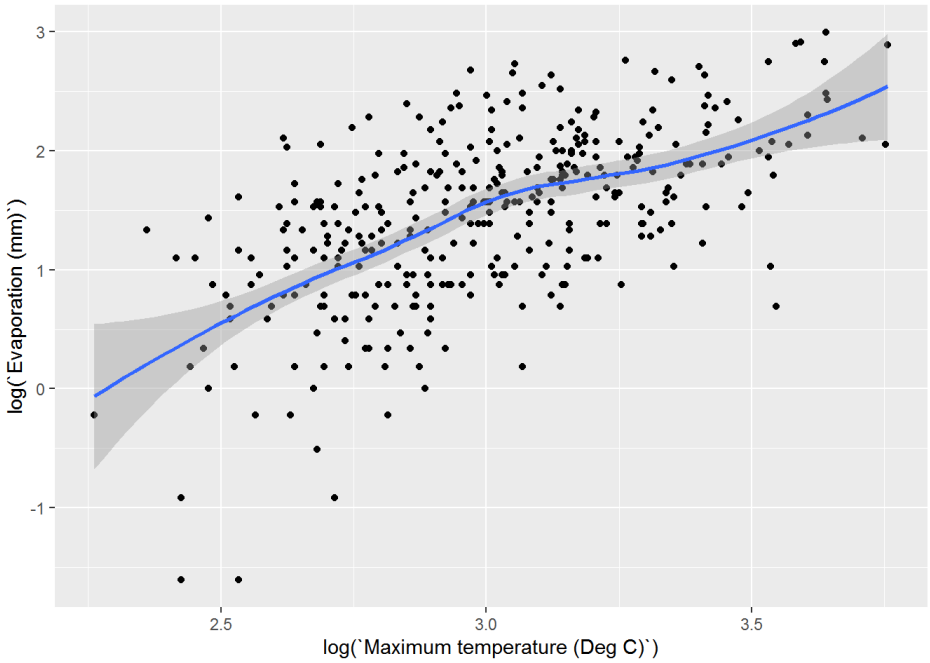


Figure 11: Scatter plot of log (Evaporation(mm)) against log (Maximum temperature (Deg C))

According to scatter plot and Pearson correlation that is |r|=0.59 (0.3<|r|<0.6) then there is a moderate linear relationship between log (Evaporation (mm)) and log (Maximum temperature (Deg C)).

4. **Log (Evaporation (mm)) vs. Minimum Temperature (Deg C):**

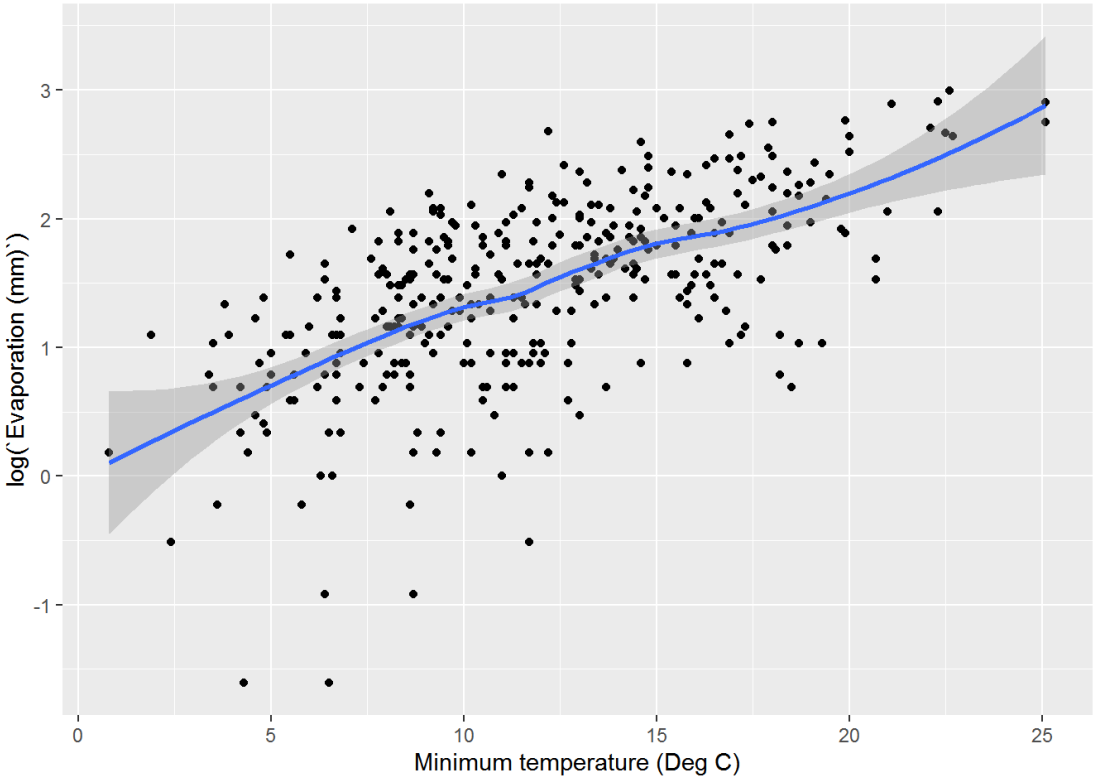


Figure 12: Scatter plot of log (Evaporation(mm)) against log (Minimum temperature)

According to scatter plot and Pearson correlation that is |r|=0.61 (|r|>0.6), then there is a strong linear relationship between log (Evaporation (mm)) vs. Minimum Temperature (Deg C).

5. Log (`Evaporation (mm)`) vs. `9am relative humidity (%)`:

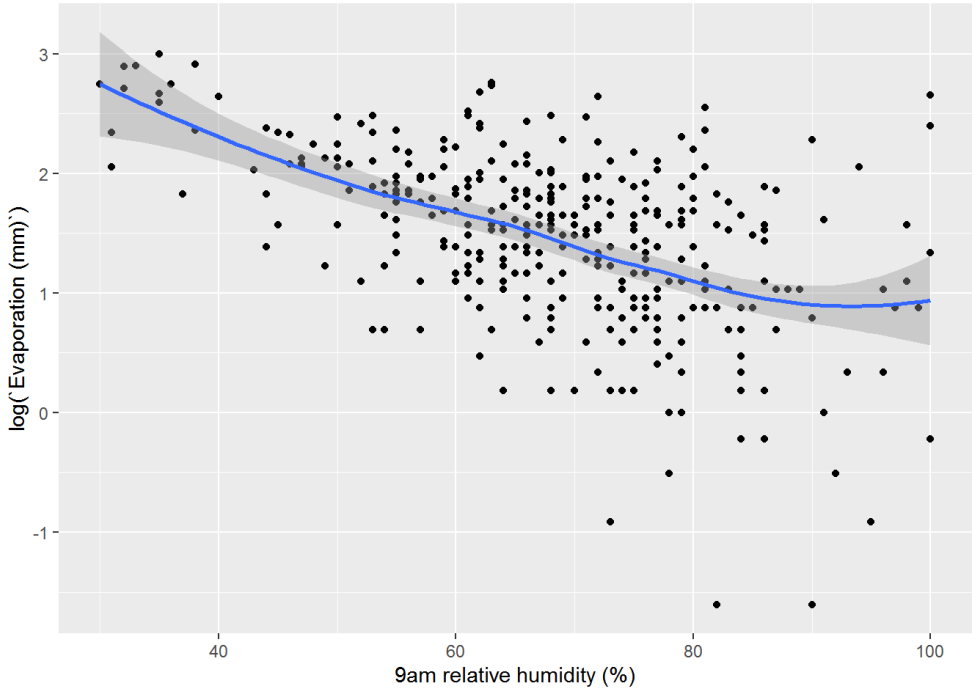


Figure 13: Scatter plot of log (Evaporation(mm)) against 9am relative humidity (%)

According to scatter plot and Pearson correlation that is |r|=0.52 (0.3<|r|<0.6) then there is a moderate linear relationship between log (Evaporation (mm)) vs. 9am relative humidity (%).

- **Check the independence of predictors:**

to review the independence of predictor variables, I calculated the correlation between them that the result is summarized in the table2:

Table 2: Correlation of predictors:

Predictor Variables		**Correlation**
Day of the week	month	-0.004
	Maximum temperature	0.023
	Minimum temperature	0.023
	9am humidity	-0.030
month	Maximum temperature	-0.230
	Minimum temperature	-0.273
	9am humidity	0.008
Maximum temperature	Minimum temperature	0.701
	9am humidity	-0.365
Minimum temperature	9am humidity	-0.239

The correlation between minimum temperature and log (maximum temperature) is 0.7 the they are highly related and one of them should be removed from our linear model. According to correlation of minimum and log of maximum temperature with evaporation, minimum temperature high related with evaporation (|r|=0.61) and log (maximum temperature) is moderate related to evaporation (|r|=0.59), then I prefer to remove the log (maximum temperature). All other predictors are not highly related.

- **Model Selection:**

In part of Model Selection, I first used all the variables of the day of the week, month, log (maximum temperature), minimum temperature, 9am humidity with the interaction of month and 9am humidity term to predict the Evaporation.

model1: log (Evaporation) = β_0 + β_1 * Day of the week +β_2 * Month + β_3 * log (Maximum temperature) + β_4 Minimum temperature + β_5 * 9am humidity + β_6 * (9am humidity) * month

Then, according to the P-value in the Summary of the model for numerical variables, maximum temperature with P-value>0.05 is not significant then I removed it from the model. By comparing the both model with ANOVA, I found this variable is not significant and removing it is true.

model2: log (Evaporation) = β_0 + β_1 * Day of the week +β_2 * Month + β_3 Minimum temperature + β_4 * 9am humidity + β_5 * (9am humidity) * month

Based on output of ANOVA for qualitative variables, P-value of day of the week is more than 0.05 and it would be removed too. Therefore, the final model is:

model 3: log (Evaporation) = β_0 + β_1 * Month + β_2 Minimum temperature + β_3 * 9am humidity + β_4 * (9am humidity) * month

To be sure about removing variables I used ANOVA and compare the both models, that P-value is not significant that means they are not different and deleting the variable is true.

Check the model for interaction term:

model 4: log (Evaporation) = β_0 + β_1 * Month + β_2 Minimum temperature + β_3 * 9am humidity

I also try to remove the interaction term and see the reflection of that on the model, but by comparing with ANOVA, I found this term is significant, then I will keep this item and my final model will be "melbourne_lm1". [3]

By comparing the bivariate results, there was not any linear relationship between Day of week and logarithm of Evaporation, and also in our final model the p-value of this term is not significant too.

The p-value of log of maximum temperature is not significant too and it should be removed but in bivariate analysis base on log of Evaporation the correlation is moderate and looks good. Since, the predictors should be independence. the log of maximum temperature has high correlation with minimum temperature and one of them can be in our linear model.

- **Model diagnostics (assumptions)**

1. **First assumption: check the linear model:**

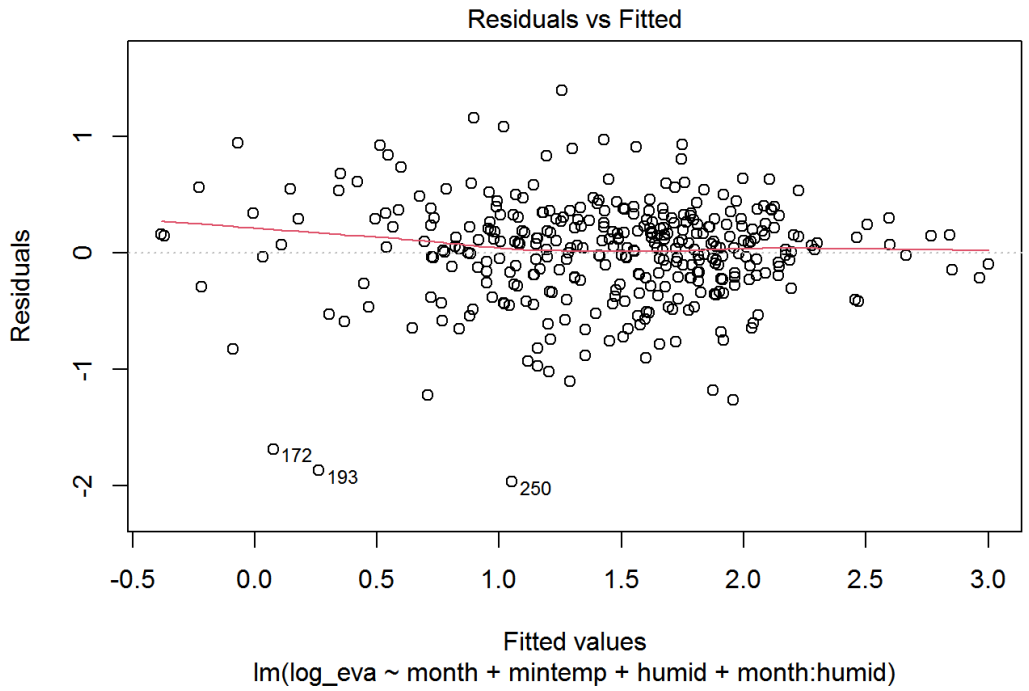


Figure 14: Residuals vs Fitted plot

When we move from left to right in residual vs fitted plot, the red line that define the pattern is almost straight (decreases a bit) and centered around zero, then it shows linear relationship between Evaporation (mm) as a response variable against the predictors is valid. [4]

2. **Second assumption: check the homoscedasticity:**

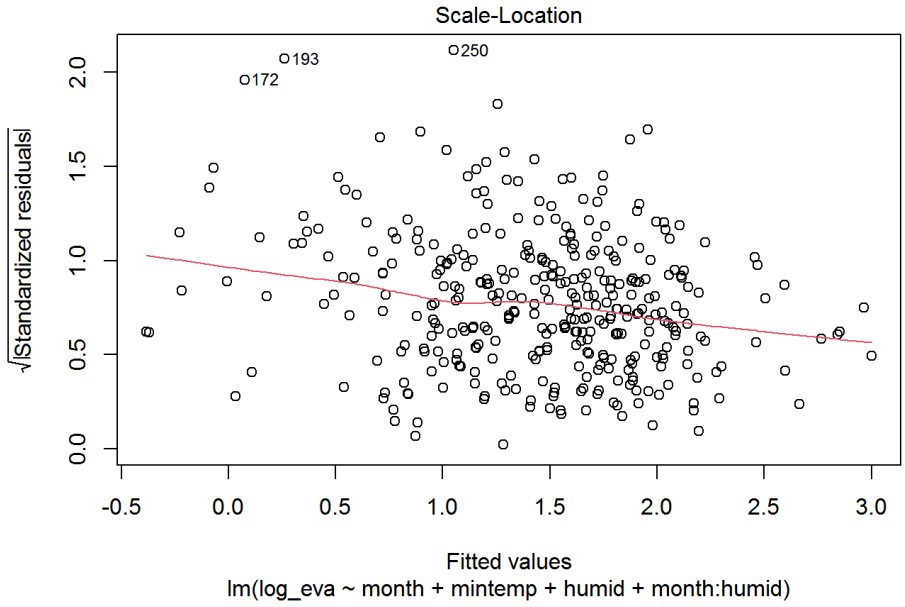


Figure 15: Scale - Location plot

Based on Scale-Location plot the red line from left to right is almost straight but it decreases a bit, and equally spread points. Therefore, the error's size is not constant in Evaporation(mm) against the independent variables. [5]

3. **Third assumption: check the normality:**

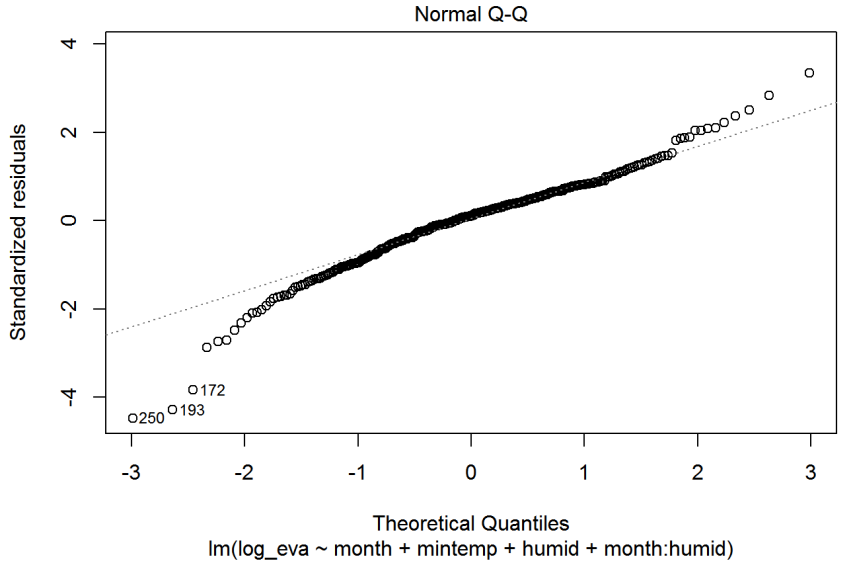


Figure 16: Normal Q-Q plot

The points between -1.1 to 1.8 in the normal Q-Q plot, are lied along the dotted line that is good because they are the most of our data. But the points less than -1.1 and greater than 1.8 on x-axis move away from the dotted line.

4. **Forth assumption: check the Independence of observations:**

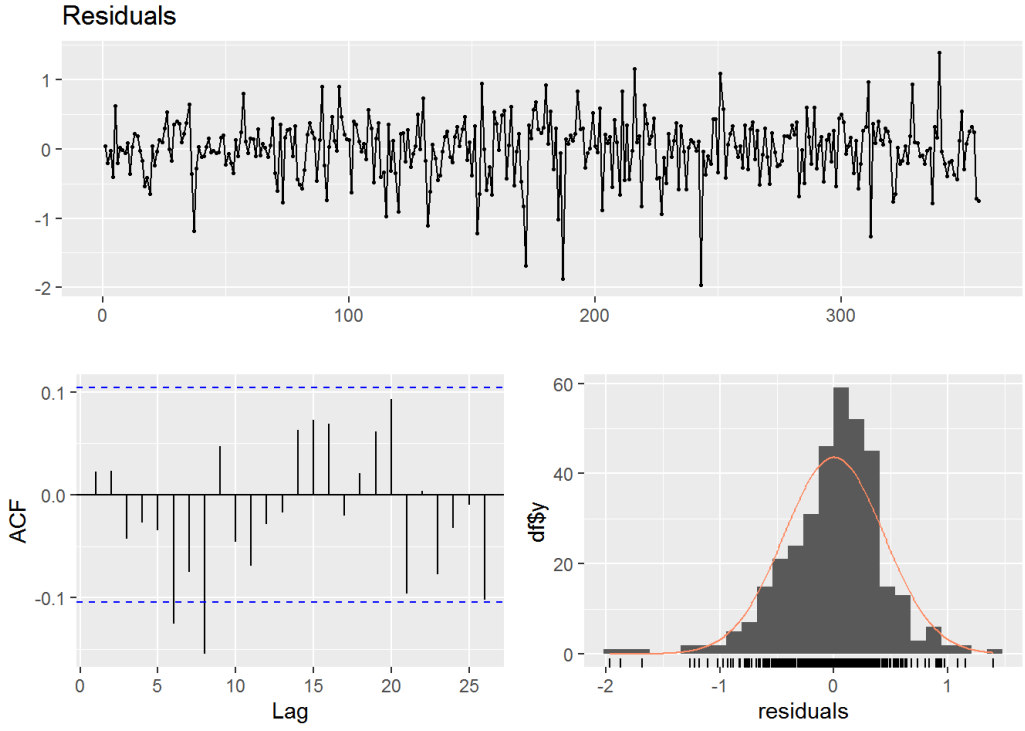


Figure 17: Residual time plot

All observations obtained based on automatic systems then it is hard to say they are 100% reliable but after checking the dataset with valid statistical method (in Appendix), From the output we can see that the test statistic is 1.946143 and the p-value is 0.226 that is more than 0.05, then the residuals in this model are not correlated and there are no hidden relationships including duplication in the data. [6]

By looking at residual time plots, the mean of the residuals is almost zero and there is no significant correlation in the residuals time plot, also, except of three outliers, the residual's variance remains the same along the days then it is constant (it can be seen on the histogram of the residuals). [7]

- The independent variables are not highly correlated with each other: Based on correlation between predictors, I found, there is high correlation between log (maximum temperature) and minimum temperature, then I remove it from linear model.
- Outliers and high leverage points and Influential values: By the Residuals vs. Leverage and Cook's distance plots, the values (outliers) that influence the results of the regression analysis, are 193, 223 and 347. [8]

The independent variables are not highly correlated with each other: Based on correlation between predictors, I found, there is high correlation between log (maximum temperature) and minimum temperature, then I remove it from linear model.

Outliers and high leverage points and Influential values:

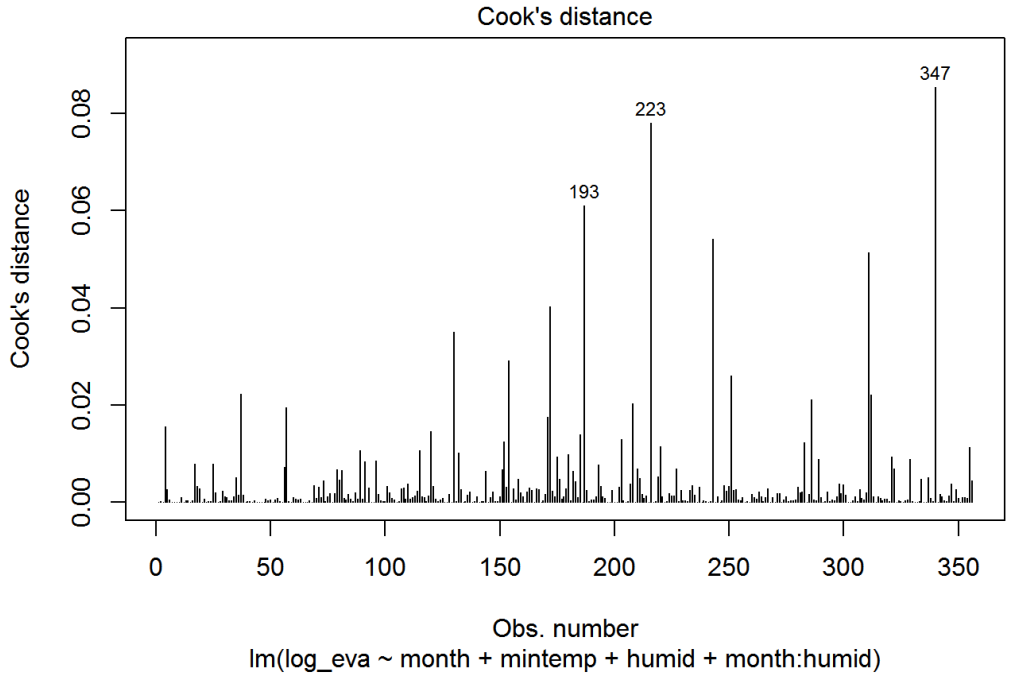


Figure 18: Cook,s distance plot

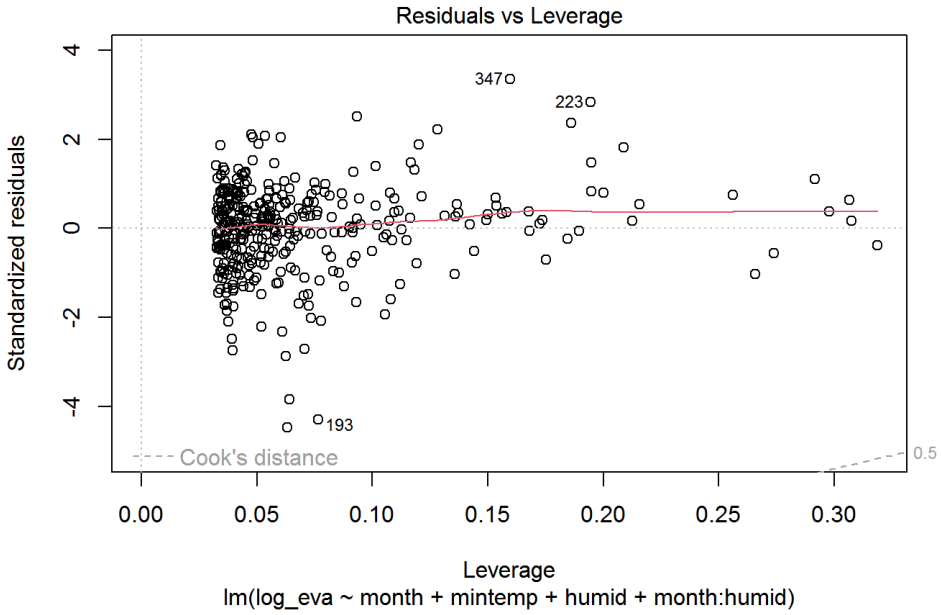


Figure 18: Residuals vs. Leverage plot

By the Residuals vs. Leverage and Cook's distance plots, the values (outliers) that influence the results of the regression analysis, are 193, 223 and 347. I check them in Appendix part. [8]

3. Results

- **Model Interpretation:**

log (Evaporation) = β_0 + β_1 * Month + β_2 Minimum temperature + β_3 * 9am humidity + β_4 * (9am humidity) * month

Intercept => β_0=1.91

slope of minimum temperature (Deg C) => β_1 = 0.048

slope of 9am humidity (%) => β_3= - 0.01

month (2 to 12) = 0 =>

Evaporation (mm) = exp. (1.91 + 0.048* minimum temperature (Deg C) - 0.01 * 9am humidity (%))

month (2 to 12) =1 =>

Evaporation (mm)) = exp. (1.91 + 0.048 * minimum temperature (Deg C) - 0.01 * 9am humidity (%) + 0.34 * month2 - 0.01 * month2: humid + 0.55 * month3 - 0.01 * (month3: humid) + 0.72 * month4 - 0.02 * (month4: humid) + ...)

By checking the ANOVA and summary of final model all of predictors are significant (p-value <0.05).

R-squared is 0.6291 (>0.5) that is Ok.

The minimum residual is -1.96847 and max is 1.39552. Distribution is almost normal.

the average increase in the response variable associated with a unit increase in the predictor variable.

Standard error is 0.454 on 331 degrees of freedom which is small.

Degree of freedom = n-k-1 = 365-3-1 = 364

4. Discussion

Based on our final linear model, predictions for specific days are summarized in table 3:

Table 3: Predictions of Evaporation(mm) with 95% confidence intervals.

Date	Minimum temperature (Deg C)	Maximum temperature (Deg C)	9am humidity (%)	Evaporation (mm)			
				fit	lwr	upr	range
2/29/2020	13.8	23.2	74	13.45	10.57	17.12	6.55
12/25/2020	16.4	31.9	57	5.31	4.22	6.67	2.45
1/13/2020	26.5	44.3	35	1.85	1.52	2.25	0.72
7/6/2020	6.8	10.6	76	7.71	6.24	9.53	3.29

Our predictions are based on 95% confidence interval, based on all 365 days of year with specific month, minimum temperature and 9am relative humidity.

According to table 3, we can see the Evaporation (mm) is minimum in July and is maximum in January. The next highest Evaporation are in December and February respectively. The range of Evaporation has a similar behavior too. Since, 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 then:

- Based our predictions, to ensure a continuous supply of water, we need transferring water from Silvan Reservoir upstream in January, because the lwr with 95% confidence is more than 10 mm.
- We don't need the transferring water from Silvan Reservoir upstream in February, July and December, because the upr with 95% confidence is less than 10 mm.

5. Conclusion

According to the results obtained from the linear regression analysis on Melbourne’s weather observations (2019) from the Bureau of meteorology’s “real time” of Australian government website, the prediction of evaporation for the year 2020, the maximum evaporation in Cardina reservoir in the city’s South East, in all months of the year except January is less than 10 (mm) that It removes the concerns for water supply in these months. In January, where the lowest evaporation rate is more than 10 (mm), I recommend the use of reserve and upstream sources for continuous water supply.

6. Appendix

I attached the R codes that I used to analysis as the html file (Assignment3-v4.html).

References

1. Notes to accompany Daily Weather Observations, Commonwealth of Australia , Bureau of Meteorology, http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml

2. 15 Features of R Programming you can't afford to overlook, https://techvidvan.com/tutorials/r-features/

3. 6. R Cookbook, 2nd Edition, 11.25 Comparing Models by Using ANOVA, https://rc2e.com/linearregressionandanova#recipe-id232

4. Linear Regression in R | An Easy Step-by-Step Guide, Published on February 25, 2020 by Rebecca Bevans. Revised on May 6, 2022. https://www.scribbr.com/statistics/linear-regression-in-r/

5. Simple Linear Regression | An Easy Introduction & Examples Published on February 19, 2020 by Rebecca Bevans. Revised on June 1, 2022. https://www.scribbr.com/statistics/simple-linear-regression/#:~:text=Linear%20regression%20models%20use%20a,relationship%20between%20two%20quantitative%20variables.

6. How to Perform a Durbin-Watson Test in R, https://www.statology.org/durbin-watson-test-r/

7. Forecasting: Principles and practice, https://otexts.com/fpp2/residuals.html

8. Articles - Regression Model Diagnostics, Linear Regression Assumptions and Diagnostics in R: Essentials, kassambara, 11/03/2018. http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/