Executive summary

As a result of a changing climate in Melbourne, Australia, previous models can no longer provide the predictive power that they once supplied. One such model involves the evaporation of the water in off-stream reservoir in Cardinia. Cardinia Reservoir is the second largest reservoir in Melbourne and primarily provides Melbourne's south, south-east, and Mornington peninsula areas with water. The problem was identified by its managers and owners, M*elbourne Water Corporation,* who have subsequently engaged *Adelaide University* to use data provided by the *Bureau of Meteorology* to construct a new model that has improved predictive capabilities for the new climate reality.

For this first model, a selection of time variables, temperature variables, a humidity variable, and a variable that represents the interaction between relative humidity and the time of year were used. A subsequent exploratory data analysis and bivariate analysis ensued which made clear some trends in the data and indicated areas of potential significance. The model was as follows:

$$y_{evapi} = \beta_{month}x_{monthi} + \beta_{MinTi}x_{MinTi} + \beta_{int}x_{monthi}x_{RH\%i} + \beta_{intercept} + \varepsilon_i$$

Its interpretation and a demonstration of its predictive value is provided as requested by *Melbourne water Corporation* (*page 10*). Assumptions of linearity, homoscedasticity, normal distribution of noise, and independence of errors have been summarised, considered, and appropriately analysed (*page 11*). The model was additionally compared to computer generated models, but no significant difference was established. This was a very simple model and its predictive capacity was limited because confidence intervals span a large proportion of the outcome variables historical values. Going forward it may be prudent to include the effects of other variables in the model to ascertain their significance and understand their relationship with other variables in the data. Additionally, it was suggested that to improve confidence in the data, more measuring devices be put in place to attain concordant results.

Method

Bivariate Summaries

A select few variables were taken from the data provided by *Melbourne Water Corporation* so their relationship with evaporation and potential predicative value could be scrutinized.

The variables selected were:

• Month (with levels starting from June when evaporation is at its minimum),
• Day of the week (with levels starting from Monday),
• Maximum temperature in degrees Celsius,
• Minimum temperature in degrees Celsius, and
• Relative humidity, as measured at 9am.

Model Selection

A linear model was built from the selected predictor variables as well as an additional interaction variable of month and humidity as measured at 9am. A backward stepwise regression enabled the selection of a model after four iterations. Using R, functions from the 'olsrr' and 'stat' libraries were then used to computer generate models for comparison Both functions arrived at the same model. The computer-generated model was then compared against our own model with an ANOVA and, although containing different predictors, was found to have no significant difference.

Model Diagnostics

A few assumptions have been made about the data that need to be addressed. We have assumed that the best model is a linear one (assumption of Linearity), that all noise in the data has the same variance (assumption of homoscedasticity), that the noise is normally distributed (assumption of normality), and that all error terms are independent (assumption of independence). These assumptions were verified as follows:

- Linearity

The residual value represents divergence from the linear model. A plot of fitted values against the residuals measures the divergence of the data from the model's mean response. Divergence from a flat line would indicate this assumption is not well supported.

- Homoscedasticity

The square root of the standardized residual value gives positive values that sit around a mean of zero and a variance of 1. It is a measure of noise in the residuals. If a plot of this value against fitted values reveals a trend, then this assumption is not well supported.

- Normality

The standardized residuals plotted against theorized quantiles speaks to the normality of the data. It maps the position of quantiles created in the standardized residuals and from a normal distribution of the fitted values. If the plot reveals curvature, then this assumption is not well supported.

- Independence

  The argument for independence can be found below in the discussion section.

Prediction

Melbourne Water Corporation has expressed an interest in seeing the selected model compute predictions with the following conditions:

- 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.
- 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.
- 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.
- 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.

These will be calculated and the accompanying confidence intervals provided. The confidence intervals represent the range of values in which the probability of finding the mean value for a specific prediction is 0.95.

Results

EDA and Bivariate Summaries

Evaporation (mm)

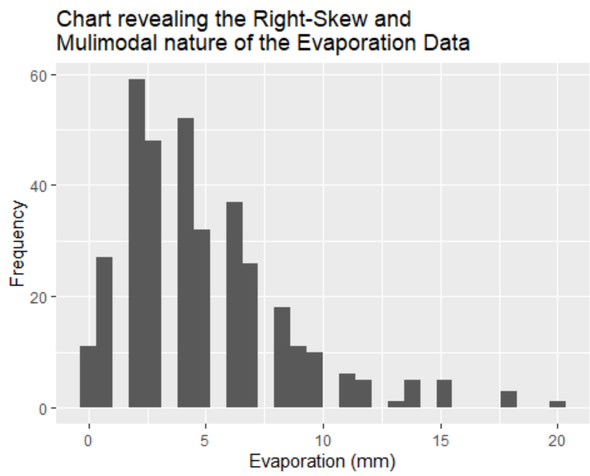

Figure 1: A histogram with evaporation in mm along the x-axis and it's frequency along the y-axis.

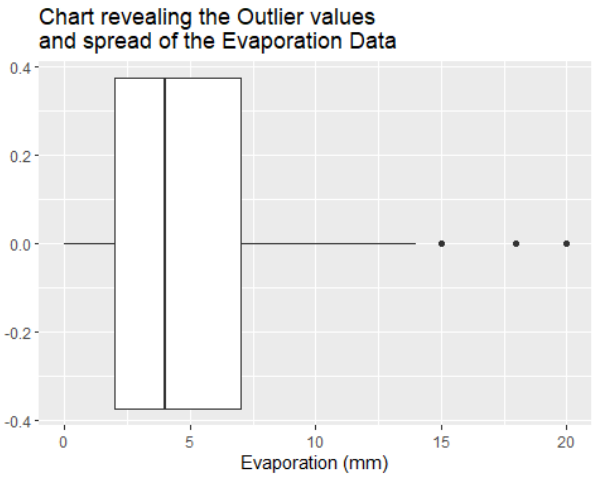

Figure 2: A boxplot with evaporation in mm along the x-axis.

The Evaporation data is multimodal and right skewed with 9 outliers; 5 at 15 mm, 3 at 18 mm, and 1 at 20 mm. The data sits on a domain of [0, 20] with a mean of 4.936 mm, and standard deviation of 3.504 mm. The median value was 4 mm, and the interquartile range is 5 mm.

Month

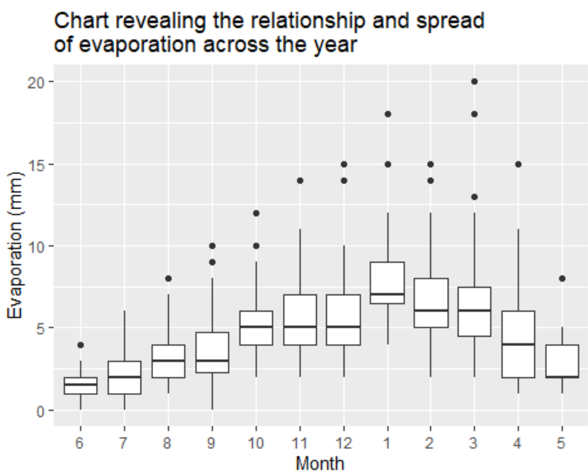

Figure 3: A side-by-side boxplot demonstrating the relationship between the categorical variable month and the quantitative variable evaporation (in mm). Months have been organised to start from the lowest median value in June.

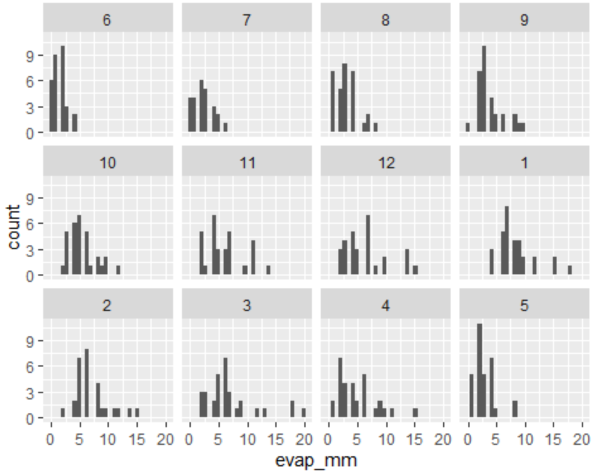

Figure 4: Additional histograms with evaporation in mm along the x-axis and frequency along the y-axis facted by month.

The levels of month start from January with a value of 1 and go through to December with a vlaue of 12. Each month appears to be multimodal and right skewed, the position along the domain changes throughout the year. There are many outliers throughout the year, but the month with the lowest variation was in June. Each month rests on a domain of (-1, 21)

Day of the week

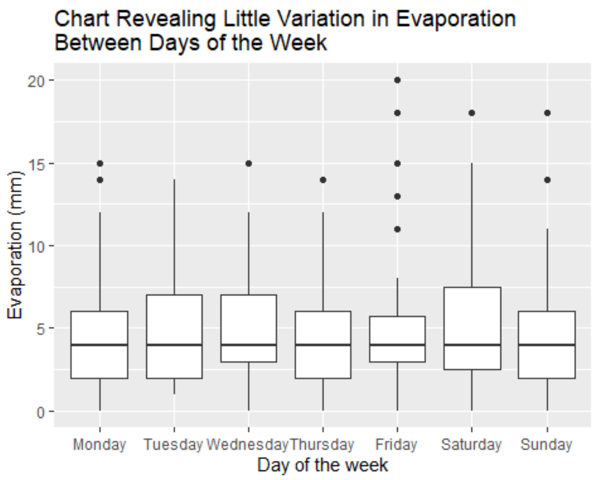

Figure 5: A side-by-side boxplot demonstrating the relationship between the categorical variable 'day of week' and the quantitative variable evaporation (in mm).

The day of the week is an unnatural and arbitrary measure of time, it does not appear to influence evaporation. An ANOVA was unable to find any significance in relation to evaporation (p-value = 0.5203).

Maximum Temperature (^{o}C)

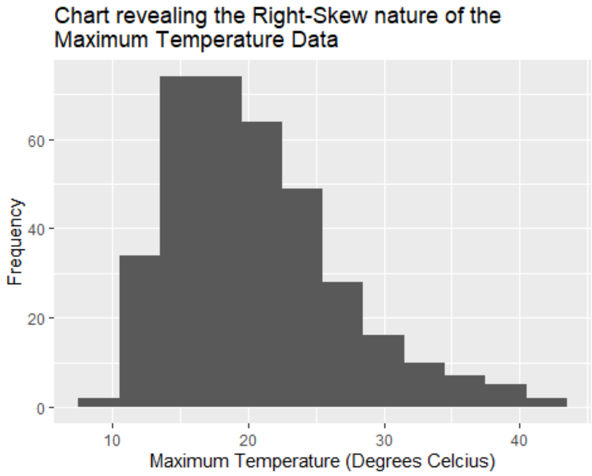

Figure 6: A histogram with maximum temperature (^{o}C) along the x-axis and frequency along the y-axis.

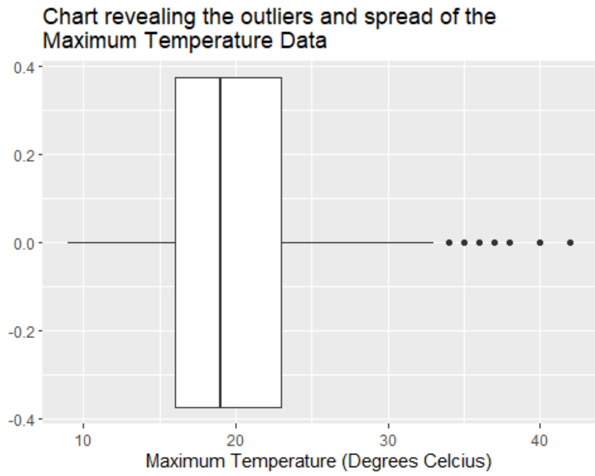

Figure 7: A boxplot of the Maximum Temperature (^{o}C) data.

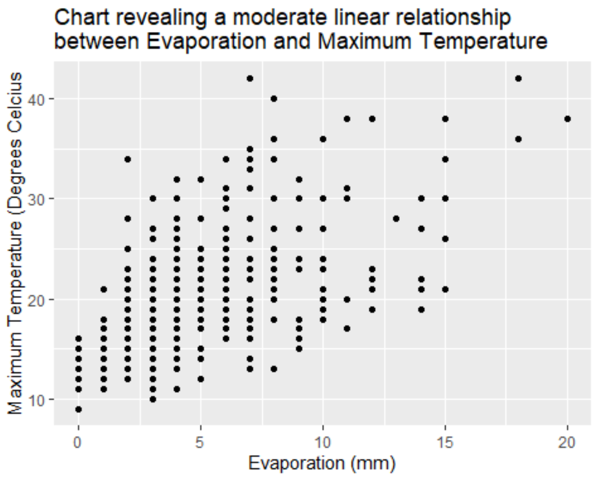

Figure 8: A scatter plot with quantitative variables evaporation (in mm) on the x-axis and Maximum Temperature (^oC) on the y-axis.

The Maximum Temperature data had a right skew and appeared to be unimodal on a domain of [9, 42]. The mean was 20.4ºC, with a standard deviation of 6.3ºC. The median was 19.0ºC and the interquartile range was 7ºC. The outliers represent values larger than (1.5 × IQR from the upper boundary of the 3rd quartile) 33.5ºC. This occurred twenty times over the period.

There is a positive linear trend of strong strength (correlation of 0. 58) between Evaporation and the Maximum Temperature.

Minimum Temperature (^oC)

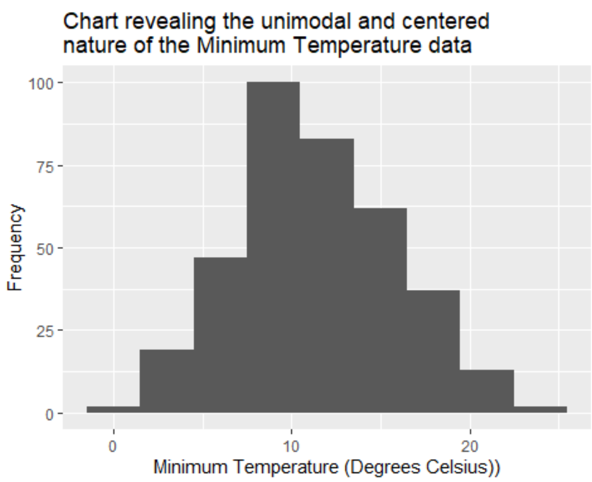

Figure 9: A histogram with the minimum temperature (^oC) along the x-axis and frequency along the y-axis.

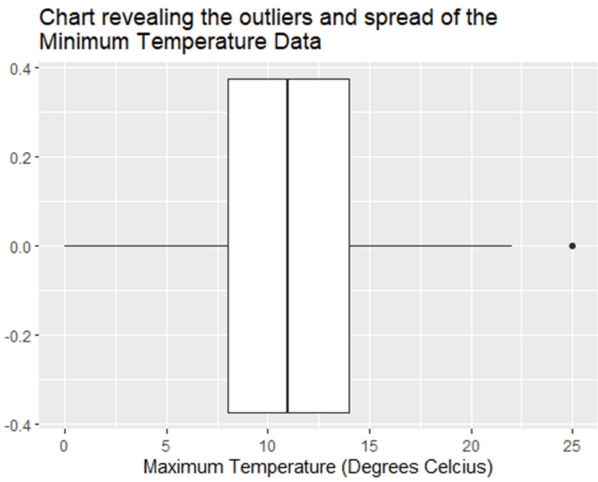

Figure 10: A boxplot of the Minimum Temperature (°C) data.

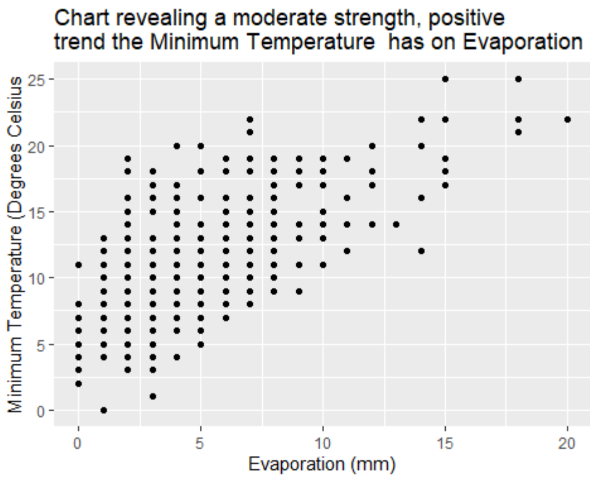

***Figure 11:** A scatter plot with quantitative variables evaporation (in mm) on the x-axis and Minimum Temperature (°C) on the y-axis.*

The Maximum temperature data was unimodal and fairly symmetrical, with the peak sitting shy of the mean of 14ºC by 1ºC, the standard deviation was 5 ºC. The data had a median of 11ºC, and an interquartile range of 6ºC. The data rested on a domain of [0, 25] with two outliers representing values larger than 23ºC; in this case there were two at 25ºC.

There is a negative linear relationship of moderate strength (correlation of 0.65) between Evaporation and the Minimum Temperature.

Relative Humidity at 9am (%)

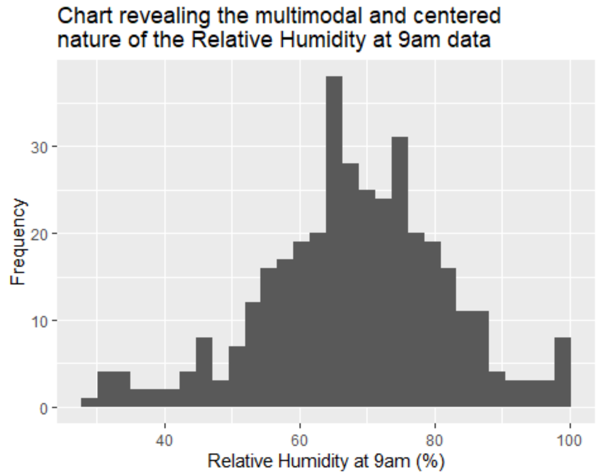

Figure 12: A histogram of the 9am Relative Humidity (%) data.

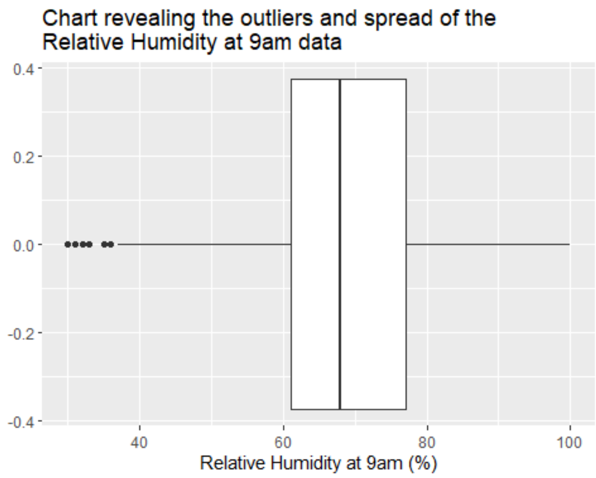

Figure 13: A boxplot of the 9am Relative Humidity (%) data.

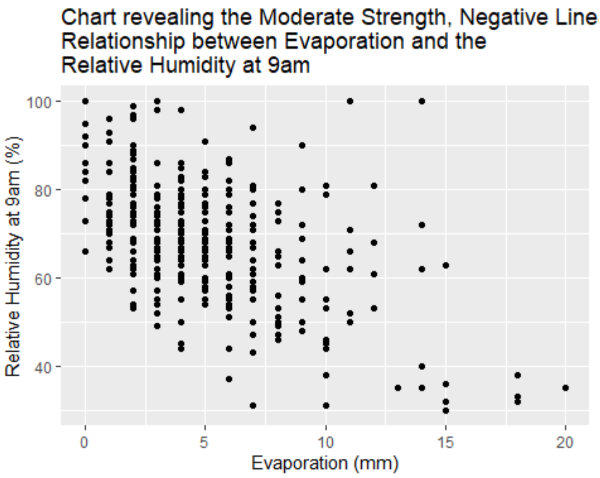

Figure 14: A scatter plot of quantitative variables evaporation in mm (on the x-axis) and the relative humidity at 9am as measured as a percentage (on the y-axis).

The Relative humidity (as measured at 9am) data was multimodal but fairly centred around the mean of 68.33%, with a standard deviation of 13.56%. The data had a median of 68.00%, and an interquartile range of 16%. The data rested on a domain of [30, 100] with nineteen outliers representing values lower than 45%.

There is a negative linear relationship of moderate strength (correlation of 0.53) between Evaporation and Relative Humidity.

Model Selection

Models generated in the stepwise regression:

Step 1:

$$y_{evapi} = \beta_{month}x_{monthi} + \beta_{MaxT}x_{MaxTi} + \beta_{MinTi}x_{MinTi} + \beta_{RH\%}x_{RH\%i} + \beta_{day}x_{dayi} + \beta_{int}x_{monthi}x_{RH\%i} + \beta_{intercept} + \varepsilon_i$$

Using this initial model, an ANOVA and the summary function were used to confirm significance with a p-value less than 0.05. The largest p-value was 0.52015 and belonged to quantitative Relative Humidity term, so it was removed

Step 2:

$$y_{evapi} = \beta_{month}x_{monthi} + \beta_{MaxT}x_{MaxTi} + \beta_{MinTi}x_{MinTi} + \beta_{day}x_{dayi} + \beta_{int}x_{monthi}x_{RH\%i} + \beta_{intercept} + \varepsilon_i$$

The summary function revealed the Maximum Temperature quantitative variable had the highest p-value (0.34033) and so it was removed.

Step 3

$$y_{evapi} = \beta_{month}x_{monthi} + \beta_{MinTi}x_{MinTi} + \beta_{day}x_{dayi} + \beta_{int}x_{monthi}x_{RH\%i} + \beta_{intercept} + \varepsilon_i$$

An ANOVA revealed the next term to be removed was the categorical Day variable with a p-value of 0.2769.

Step 4

$$y_{evapi} = \beta_{month}x_{monthi} + \beta_{MinTi}x_{MinTi} + \beta_{int}x_{monthi}x_{RH\%i} + \beta_{intercept} + \varepsilon_i$$

All remaining variables were found to be statistically significant.

Table 1: A summary of the functions used to generate p-values for each variable in the model.

Function used	Variable	p-value
ANOVA	Month	< 2.2e-16
ANOVA	Month : Relative Humidity interaction term	< 2.2e-16
Summary function	Minimum Temperature	4.10e-16

Two separate computer-generated models were also created. In both cases the output model was:

$$y_{evapi} = \beta_{month}x_{monthi} + \beta_{MinTi}x_{MinTi} + \beta_{RH\%}x_{RH\%i} + \beta_{int}x_{monthi}x_{RH\%i} + \beta_{intercept} + \varepsilon_i$$

An ANOVA found no significant difference between the model we made and the model above.

Model Diagnostics

The assumptions were appropriately scrutinized. The residuals of the fitted values were largely linear for the bulk of the data; suggesting the assumption of linearity holds. However, we note that the assumption of linearity does become less clear when fitted values are larger than 10 mm.

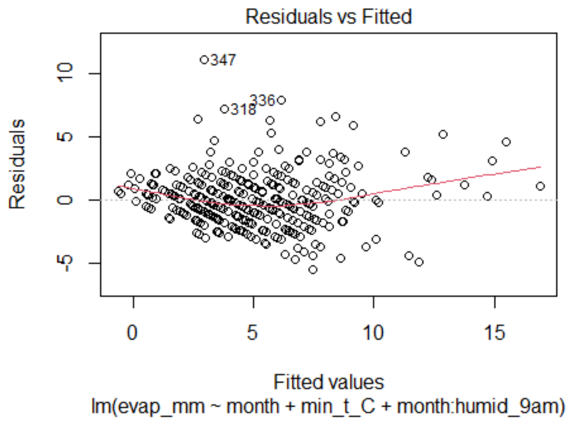

Figure 15: A plot of fitted values from our model on the x-axis against the corresponding residual values on the y-axis.

The square-root of standardized residuals of the fitted values has a slight slope but nothing exponential and appears to become less prominent across the domain. The assumption of homoscedasticity seems to be reasonable.

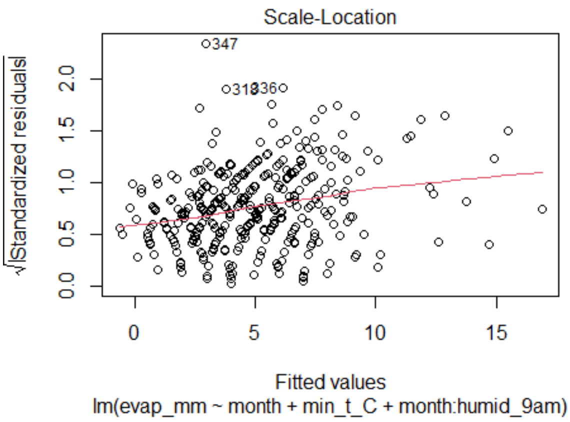

Figure 16: A plot of residual values from our model on the x-axis against the corresponding (standardized residuals)$^{0.5}$ on the y-axis

The standardized residuals mapped against the theoretical quantiles from a normal distribution of fitted values holds linearity on the domain of [-2, 2]. The assumption of a normal noise distribution appears quite reasonable.

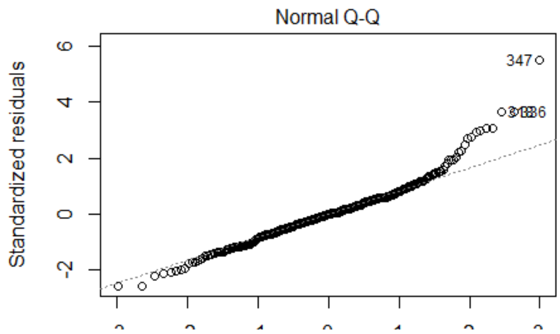

Figure 17: A plot of theoretical quantiles from a normal distribution of our model's fitted values on the x-axis against the corresponding standardized residuals on the y-axis

Model Interpretation

$$y_{evapi} = \beta_{month} x_{monthi} + \beta_{MinTi} x_{MinTi} + \beta_{int} x_{monthi} x_{RH\%i} + \beta_{intercept} + \varepsilon_i$$

The model above can also be written as below when you substitute in the coefficients:

$$y_{evapi} = (0.63358 + A) x_{monthi} + (0.35538) x_{MinTi} + (0.63358 + B) x_{monthi} x_{RH\%i} + (0.63358) + \varepsilon_i$$

Where 'A' and 'B' represent the alteration that needs to be made to the coefficient in order to account for the specific month y is to be calculated for. When taken in parts:

- When it comes to $\beta_{intercept}$, the value can be thought of as something of a baseline representing a subset of categorical data that's equivalent to when all other subsets have a value of zero. In this instance, June is that baseline/ reference column (when all subsets = 0).
- When the x_{MinTi} variable increases by 1 degree, it is expected that evaporation will increase by 0.35538 mm.
- When it is June (and A, B, and x_{monthi} =0), we expect evaporation to increase by 0.63358. Change the month and you move off the reference and need to alter the coefficients of A and B.
- As for the interaction term, this is indicating that in June (when B and x_{monthi}=0), would expect an increase in evaporation of 0.63358. Change the month and the B value and humidity value will influence the coefficient. But with a humidity of 1% in July (when x_{monthi}= 1, and B = - 0.02334) the evaporation will increase by 0.61024 mm.

These relationships allow our model to predict evaporation when you give A, B, and each x variable context about what you want to predict.

Discussion

Bivariate Summaries

Month

This variable is the largest meaningful measure of time we have for an analysis of this nature. It encompasses climate cycles very well because it's the climate cycles are governed by the same rules - the position around the sun. This was not always the case, for this we can thank a Pope who was a stickler for the position of Earth on Easter. Now the calendar year maps well to the seasons, which you can see in *figure 3,* and the variable will therefore hold significance for any given year. In terms of the **assumption of independence**, the month observation value from one subject could not give information about the month observations value for another as it is a label that divides the data in time (figure 3). Its ability to give information regarding other observations cannot be determined with any level of certainty, there are no hard and fast rules in weather systems. There are only probable relations, and just how probable is difficult to ascertain; for example, a hot day in January does not necessarily mean another day in January will be hot. So, there is a strong argument for independence of error.

It was very likely that the month would have a significant influence on evaporation as all the other variables that BOM collects are highly dependent on it; such as temperature, humidity, rainfall, etc. It therefore came as no surprise that Month was included in the final model, nor that the interaction term with humidity was significant (*table 1*). Particularly when considering the variation observed in evaporation between months in the bivariate analysis (*figure 3).*

Day

The days of the week are an arbitrary and unnatural division of time. There are no weather cycles with a duration of 7 days that are widely known about. The side-by-side boxplot also showed no great variation, so it comes as no surprise that it was not included in the final model (*figure 5*).

Maximum Temperature (ºC)

When it comes to phase changes, there are two main factors: temperature and pressure. When pressure is constant, increasing the temperature will result in more evaporation. As water molecules become more energetic, or get hotter, they begin to break free of the hydrogen bonds between themselves and other nucleophiles surrounding them. This is quite a strong bond considering it takes 100 ºC at sea-level for pure water to boil. There are many other factors that influence the threshold energy required, but maximum temperature was expected to be in the model as it had a positive linear relationship with evaporation (*figure 6*). Interesting that it was the minimum and not the maximum temperature that was included. However, it is expected that there are collinearity effects occurring between the month and interaction terms that account for this.

Minimum Temperature(ºC)

The minimum temperature marks an important variable because it indicates the amount of energy the water starts the day with. The higher this value is, the less time is needed for the water to acquire the remaining energy needed to evaporate, and the more evaporation you could have in a day as supported by *figure 11*. This was therefore another of the variables that were hypothesized to be significant in producing the model. In terms of the **assumption of independence**, there is no way the observations from one day could give information about another unless there were consistent systematic errors with the measurement device. The observation for one day cannot give information with any certainty about even the next day, that is the nature of weather predictions.

Relative Humidity as measured at 9am

The amount of water the air can hold is limited. Higher relative humidity makes it difficult for water molecules to become gaseous as the energy required to remain in that state at reasonable temperatures is dissipated away through interaction with other molecules in the air to the point that water molecules begin to hydrogen bond once more, form back into liquids, and sink out of the sky as a result of gravity and buoyant forces (observable in figure 14). Taking the measurement in the morning as apposed to the afternoon is important, as less evaporation would be expected by 9 am than by 3pm; indicating a reasonable minimum and the capability for saturation that day. The saturation of air is well understood and dependent on temperature. It's variation through the seasons is also predictable. It is interesting that the computer-generated models found the variable to be significant but ours did not; our initial hypothesis was in agreement with the computer-generated model. Perhaps the minimum temperature and the mathematical equation for relative humidity in conjunction with the interaction term simply give a proportional amount of information; the ANOVA comparing the two found no significant difference indicating this is likely to be true (*table 1*). In terms of the **assumption of independence**, unless there were consistent systematic error the relative humidity measured on one morning could not give you information with any certainty about observations of another day.

Prediction

Confidence intervals indicate a 95% chance of finding the mean amount of evaporation using the model, while prediction intervals indicate a 95% chance of finding a specific instance of an evaporation measurement. As the question relates to specific dates, prediction intervals are more relevant. They will indicate the range of values you would need to include in order to be 95% certain the true value has been included. See below the predictions made for the scenarios requested by *Melbourne Water Corporation* with the accompanying prediction intervals (*table 2*).

Table 2: The predictions and prediction intervals calculated for the four scenarios presented in the order listed in the method section.

Scenario	Fit	Lower bound	Upper bound	Range
Month = 2 (Feb) Min Temp = 13.8 Humidity = 74%	5.356955 mm	0.912332 mm	9.801579 mm	~ 8.9
Month = 12 (Dec) Min Temp = 16.5 Humidity = 57%	8.346807 mm	3.920063 mm	12.77355 mm	~ 8.9
Month = 1 (Jan) Min Temp = 26.5 Humidity = 35%	14.50894 mm	9.708499 mm	19.30938 mm	~ 9.6
Month = 7 (Jul) Min Temp = 6.8 Humidity = 76%.	2.062808 mm	- 2.343371 mm	6.468986 mm	~ 8.7

Regarding the question of which of the above scenarios would warrant transferring water from Silvan Reservoir to the Cardinia Reservoir due to a very specific evaporative event resulting in evaporation of more than 10 mm, the December and January scenario's prediction intervals encompass values over 10 mm, while the February and July scenario's do not. This suggests water will need to be transferred for the December and January scenarios, but not February or July. This should come as little surprise as January and December have an angle of incidence that allows a greater amount of energy to enter the system while February and July do not. More energy will make it harder to dissipate energy away through convection, and potentially increase the duration of radiation from rocks or concrete over-night which will influence the minimum water temperature for the next day. February and July are associated with cooler air temperatures in this part of the world which aids convection in dissipating energy from the water.

The model's capabilities appear to be quite limited. In order to be 95% certain of finding the mean, you need to include a range that spans over a third of the domain observed in previous years (*Figure 1*). This highlights the need for a more refined model that can reduce those confidence intervals by improving the error margins between observed values and predicted values. This could be through improved measurement, additional variables (such as those listed below), or interaction terms (many of the observations relate to each other directly through known mathematical equations or simply by understanding the phenomena as is the case with the relationship between month/ angle of incidence for energy coming into the system and humidity).

Evaporation is known to be influenced by a number of variables:

- Time of year / Season – as the Earth orbits the sun, the angle of incident light on our atmosphere changes as a result of the tilt in the Earth's axis. This will result in cyclic fluctuations in the amount of energy absorbed by the reservoir.
- Temperature – the warmer the water becomes the more likely it is to break free of the surface tension and evaporate away. Air temperature also determines the upper limit of how much water can exist in a gaseous state; allowing us to calculate a relative humidity.
- Altitude – the lower the air pressure is the more vacuum like it becomes. Air pressure is therefore evaporation is inversely proportionate to air pressure.
- Wind speed – Moving air molecules enable the transferal of energy to water molecules on the surface of the water. The more wind there is the less energy a water molecule needs in order to turn into a gas.
- Landscape – fauna acts as a wind block and plays a role in moderating temperature. land shape can also direct wind and plays a role in local weather systems.
- Proximity to civilization – the heat retention properties of concrete results in in temperatures being higher in built up areas.
- The salt and heavy metal concentration – these both make water 'heavier '. One sodium atom will hold 6 water molecules and many heavy metals can hold as many as 8. The amount of energy required to evaporate these bound water molecules therefore increases.
- Etc.

Many of the variables may well have collinearity (which would indicate the information that can provide the model is already account for with another variable, but this should be confirmed with a follow up analysis. The variables that represent the fundamental structure of the system are likely to be significant in an improved model. This would include things like the minimum temperature, minimum humidity, minimum wind speed, perhaps the rate of change in air pressure, concentration of salts and heavy metals; perhaps even with co-efficients for specific molecules and their interaction with other molecules in the system. When it comes to interaction terms, even minimum temperature may have a significant interaction term with the month as the season (and angle of incident light) dictates the amount of energy that can enter the system. Many variables that depend on the season, such as wind speeds, rainfall, cloud cover, foliage and detritus (which will influence effective windspeeds and water cover), etc. should also have their interaction with month assessed. Furthermore, there are relationships that are mathematically related in the data, such as that of temperature and pressure (influencing not only evaporation, but also windspeed). Additionally, introducing multiple measuring devices that produce concordant results would support the validity of measurements and give greater confidence in the data.

Conclusion

An exploratory data analysis and bivariate analysis were conducted to draw conclusions about the data and any relationships that exist with evaporation. These variables were then used in a stepwise backwards regression and compared with computer generated models. The models' assumptions were all reasonably satisfied meaning the model could be used to make predictions regarding the four scenarios provided by *Melbourne Water Corporation*. The large range found between the upper and lower bound of confidence intervals indicate there is room for improvement. Several variables that could hold promise were provided, and the role interaction terms may have in improving the model was conveyed.

References

Grolemund, G & Wickham, H 2011, 'Dates and Times Made Easy with {lubridate}', *Journal of Statistical Software*, vol. 40, no. 3, pp. 1-25.

Hebbali, A 2020, 'olsrr: Tools for Building OLS Regression Models', <https://CRAN.R-project.org/package=olsrr>.

R Core Team, 2021, *R: A language and environment for computing and statistical computing* R foundation for statistical computing Vienna, Austria.

Rushworth, A 2021, 'Inspectdf: Inspection, Comparison and Visualisation of Data Frames', *R package version 0.0.11*, <https://CRAN.R-project.org/package=inspectdf>.

Wickham, H 2019, 'stringr: Simple, Consistent Wrappers for Common String Operations', R package version 1.4.0, <https://CRAN.R-project.org/package=stringr>.

Wickham, H, Averick, M, Bryan, J, Winston, C, McGowan, LDA, François, R, Grolemund, G, Hayes, A, Henry, L, Hester, J, Kuhn, M, Pedersen, TL, Miller, E, Bache, SM, Müller, K, Ooms, J, Robinson, D, Seidel, DP, Spinu, V, Takahashi, K, Vaughan, D, Wilke, C, Woo, K & Yutani, H 2019, 'Welcome to the {tidyverse}', *Journal of Open Source Software*, vol. 4, no. 43, p. 1686.

Wickham, H, François, R, Henry, L & Müller, K 2021, 'dplyr: A Grammar of Data Manipulation', R package version 1.0.7, <https://CRAN.R-project.org/package=dplyr>.

Appendix

```
library("tidyverse")
library("dplyr")
library("lubridate")
library("stringr")
library("inspectdf")
library("olsrr")
```

(Grolemund & Wickham 2011; Hebbali 2020; R Core Team 2021; Rushworth 2021; Wickham 2019; Wickham et al. 2019; Wickham et al. 2021)

Table A1: The data prior to cleaning

Date	Minimum temperature (Deg C)	Maximum Temperature (Deg C)	Rainfall (mm)	Evaporation (mm)	Sunshine (hours)	Direction of maximum wind gust	Speed of maximum wind gust (km/h)	Time of maximum wind gust	9am Temperature (Deg C)	9am relative humidity (%)	9am cloud amount (oktas)	9am wind direction	9am wind speed (km/h)	9am MSL pressure (hPa)	3pm Temperature (Deg C)	3pm relative humidity (%)	3pm cloud amount (oktas)	3pm wind direction	3pm wind speed (km/h)	3pm MSL pressure (hPa)
2019-01-1	15.5	26.2	0.0	7.0	11.0	S	35	17:44:00	19.8	74	7	S	6	1013.0	24.4	45	1	SSW	11	1011.5
2019-01-2	18.4	22.2	0.0	7.0	7.5	SSW	39	15:23:00	19.5	64	8	SSE	7	1013.9	21.4	62	1	SSW	19	1012.9
2019-01-3	15.9	29.5	0.0	6.6	9.3	SSW	26	14:53:00	18.1	75	8	S	2	1012.6	24.6	60	0	SSW	13	1009.9
2019-01-4	18.0	42.6	0.0	7.8	12.2	NW	54	12:03:00	29.5	31	0	NNE	9	1005.5	42.0	16	1	NW	15	1001.0
2019-01-5	17.4	21.2	0.4	15.4	5.8	SSW	39	08:24:00	18.0	63	7	S	13	1013.5	19.1	58	7	S	11	1013.4
2019-01-6	14.6	22.1	1.4	6.4	13.3	SSW	33	11:12:00	17.7	55	1	SW	9	1020.4	20.6	48	1	SSW	13	1019.5

```
colnames(mwc)
#simplify column names:
mwc <- rename(mwc,
    date = "Date",
    min_t_C = "Minimum temperature (Deg C)",
    max_t_C = "Maximum Temperature (Deg C)",
    rainfall_mm = "Rainfall (mm)",
    evap_mm = "Evaporation (mm)",
    sun_hrs = "Sunshine (hours)",
    gust_dir = "Direction of maximum wind gust",
    gust_kmh = "Speed of maximum wind gust (km/h)",
    gust_time = "Time of maximum wind gust",
    temp_9am_C = "9am Temperature (Deg C)",
    humid_9am = "9am relative humidity (%)",
    cloud_9am = "9am cloud amount (oktas)",
    wind_dir_9am = "9am wind direction",
    wind_speed_9am = "9am wind speed (km/h)",
    pres_9am = "9am MSL pressure (hPa)",
    temp_3pm_C = "3pm Temperature (Deg C)",
    humid_3pm = "3pm relative humidity (%)",
    cloud_3pm = "3pm cloud amount (oktas)",
    wind_dir_3pm = "3pm wind direction",
    wind_speed_3pm = "3pm wind speed (km/h)",
    pres_3pm = "3pm MSL pressure (hPa)"
    )
#create day, month, year columns
mwc <- mwc %>%
 mutate(date = as.Date(date),
```

```
    day = day(date), month = month(date), year = year(date))
#evaporation brought next to date to represent the subject of "evaporation on a given day".
mwc <- mwc[,c(1, 5, 24, 23, 22, 2, 3, 4,  6:21)]
#make day = day of week
mwc$day <- weekdays(mwc$date)
#Set value types
#low level categorical = factor, year could be considered a character as in the scheme of things time is infinite but we only have from 2019 so in this context it'll be a factor
mwc$year <-  factor(mwc$year)
month_levels <- c("6", "7", "8", "9", "10", "11", "12", "1", "2", "3", "4", "5")
mwc$month <-  factor(mwc$month, levels = month_levels) #setting levels from lowest average to highest
day_levels <- c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "Sunday")
mwc$day <-  factor(mwc$day, levels = day_levels)
mwc$gust_dir <- factor(mwc$gust_dir)
mwc$wind_dir_9am <- factor(mwc$wind_dir_9am)
mwc$wind_dir_3pm <- factor(mwc$wind_dir_3pm)
#All measurements appear to have a set amount of significant figures suggesting all other variables are discrete =integers
mwc$evap_mm <- as.integer(mwc$evap_mm)
mwc$min_t_C <- as.integer(mwc$min_t_C)
mwc$max_t_C <- as.integer(mwc$max_t_C)
mwc$rainfall_mm <- as.integer(mwc$rainfall_mm)
mwc$sun_hrs <- as.integer(mwc$sun_hrs)
mwc$gust_kmh <- as.integer(mwc$gust_kmh)
mwc$temp_9am_C <- as.integer(mwc$temp_9am_C)
mwc$humid_9am <- as.integer(mwc$humid_9am)
mwc$cloud_9am <- as.integer(mwc$cloud_9am)
mwc$wind_speed_9am <- as.integer(mwc$wind_speed_9am)
mwc$pres_9am <- as.integer(mwc$pres_9am)
mwc$temp_3pm_C <- as.integer(mwc$temp_3pm_C)
mwc$wind_speed_9am <- as.integer(mwc$wind_speed_9am)
mwc$humid_3pm <- as.integer(mwc$humid_3pm)
mwc$cloud_3pm <- as.integer(mwc$cloud_3pm)
mwc$wind_speed_3pm <- as.integer(mwc$wind_speed_3pm)
mwc$pres_3pm <- as.integer(mwc$pres_3pm)
inspect_types(mwc) #Looking good, time variables already sorted for me.
## # A tibble: 4 x 4
##   type            cnt  pcnt col_name
##   <chr>         <int> <dbl> <named list>
## 1 integer          16 66.7  <chr [16]>
## 2 factor            6 25    <chr [6]>
## 3 Date              1  4.17 <chr [1]>
## 4 hms difftime      1  4.17 <chr [1]>
```

```
#Tibble ready for analysis
```

Table A2: The tidy data

date	evap_mm	year	month	day	min_t_C	max_t_C	rainfall_mm	sun_hrs	gust_dir	gust_kmh	gust_time	temp_9am_C	humid_9am	cloud_9am	wind_dir_9am	wind_speed_9am	pres_9am	temp_3pm_C	humid_3pm	cloud_3pm	wind_dir_3pm	wind_speed_3pm	pres_3pm
2019-01-01	7	2019	1	Tuesday	15	26	0	11	S	35	17:44:00	19	74	7	S	6	1013	24	45	1	SSW	11	1011
2019-01-02	7	2019	1	Wednesday	18	22	0	7	SSW	39	15:23:00	19	64	8	SSE	7	1013	21	62	1	SSW	19	1012
2019-01-03	6	2019	1	Thursday	15	29	0	9	SSW	26	14:53:00	18	75	8	S	2	1012	24	60	0	SSW	13	1009
2019-01-04	7	2019	1	Friday	18	42	0	12	NW	54	12:03:00	29	31	0	NNE	9	1005	42	16	1	NW	15	1001
2019-01-05	15	2019	1	Saturday	17	21	0	5	SSW	39	08:24:00	18	63	7	S	13	1013	19	58	7	S	11	1013
2019-01-06	6	2019	1	Sunday	14	22	1	13	SSW	33	11:12:00	17	55	1	SW	9	1020	20	48	1	SSW	13	1019

EDA and Bivariate analysis

```
ggplot(mwc, aes(x = evap_mm)) +
 geom_histogram() +
 labs( title = "Chart revealing the Right-Skew and \nMulimodal nature of the
Evaporation Data",
    x = "Evaporation (mm)",
    y = "Frequency")
ggplot(mwc, aes(x = evap_mm)) +
 geom_boxplot() +
  labs( title = "Chart revealing the Outlier values \nand spread of the
Evaporation Data",
    x = "Evaporation (mm)")
summary(mwc$evap_mm)
sd(mwc$evap_mm, na.rm = TRUE)
filter(mwc, evap_mm >=15)
```

month

```
ggplot(mwc, aes(x = month, evap_mm)) +
 geom_boxplot()+
  labs( title = "Chart revealing the relationship and spread \nof evaporation
across the year",
    y = "Evaporation (mm)",
    x = "Month")
#evap (Quant) vs month (cat) (side by side box-plot)
ggplot(mwc, aes(evap_mm)) + geom_histogram() + facet_wrap(~month)
```

Day of week

```
# not expecting to find anything informative here
ggplot(mwc, aes(x = day, evap_mm)) +
 geom_boxplot() +
  labs( title = "Chart Revealing Little Variation in Evaporation \nBetween Days
of the Week",
    y = "Evaporation (mm)",
    x = "Day of the week")
#evap (Quant) vs day (cat) (side by side box-plot)
unique(mwc$day)
day_signif<- lm(evap_mm~day, data =mwc)
anova(day_signif)
```

Maximum Temp in deg C

```
ggplot(mwc, aes(x = evap_mm, max_t_C)) +
 geom_point()+
  labs( title = "Chart revealing a moderate linear relationship \nbetween
Evaporation and Maximum Temperature",
    x = "Evaporation (mm)",
    y = "Maximum Temperature (Degrees Celcius")
#evap (Quant) vs Max T (quant) (scatterplot)
ggplot(mwc, aes(max_t_C)) +
 geom_histogram(binwidth=3) +
  labs( title = "Chart revealing the Right-Skew nature of the Maximum
Temperature Data",
    x = "Maximum Temperature (Degrees Celcius)",
    y = "Frequency")
#right skewed
ggplot(mwc, aes(log(max_t_C)))+ geom_histogram(binwidth=0.1)
#normalizes well with log(x)
summary(mwc$max_t_C)
sd(mwc$max_t_C, na.rm = TRUE)
filter(mwc, max_t_C > 33.5)
cor(mwc$evap_mm, mwc$min_t_C, use = "complete.obs")
```

Minimum Temp in deg C

```
# As above but less of a slope is what I'm predicting. Colder days might
have larger wind speeds though...
ggplot(mwc, aes(x = evap_mm, min_t_C)) + geom_point() +
  labs( title = "Chart revealing a moderate strength, positive \ntrend the
Minimum Temperature  has on Evaporation ",
    x = "Evaporation (mm)",
    y = "Minimum Temperature (Degrees Celsius")
#evap (Quant) vs Min T (quant) (scatterplot)
ggplot(mwc, aes(min_t_C)) + geom_histogram(binwidth=3) +
 labs( title = "Chart revealing the unimodal and centered \nnature of the
Minimum Temperature data",
    x = "Minimum Temperature (Degrees Celsius))",
    y = "Frequency")
#fairly normal distribution
summary(mwc$min_t_C)
sd(mwc$min_t_C, na.rm = TRUE)
filter(mwc, min_t_C > 23)
cor(mwc$evap_mm, mwc$min_t_C, use = "complete.obs")
```

Relative Humidity at 9am

```
ggplot(mwc, aes(x = evap_mm, humid_9am)) + geom_point() +
 labs( title = "Chart revealing the Moderate Strength, Negative Linear
\nRelationship between Evaporation and the \nRelative Humidity at 9am",
    x = "Evaporation (mm)",
    y = "Relative Humidity at 9am (%)")
#evap (Quant) vs humid 9am (quant) (scatter plot)
# indeed, higher humidity results in less evaporation
```

```
summary(mwc$humid_9am)
sd(mwc$humid_9am, na.rm = TRUE)
filter(mwc, humid_9am <45)
cor(mwc$evap_mm, mwc$humid_9am, use = "complete.obs")
```

Model selection

```
#Evaporation (in mm) on a given day in Melbourne (our evap mm is a daily
measure from melbourne)
evap_model1 <- lm(evap_mm ~ month + max_t_C + min_t_C + humid_9am + day +
month:humid_9am, data = mwc)
summary(evap_model1)
anova(evap_model1)
#remove humid (0.52015)

evap_model2 <- lm(evap_mm ~ month + min_t_C + day + month:humid_9am, data
= mwc)
summary(evap_model2)
anova(evap_model2)
#remove maxT (0.34033)

evap_model3 <- lm(evap_mm ~ month + min_t_C + day + month:humid_9am, data
= mwc)
summary(evap_model3)
anova(evap_model3)
#remove day (0.2769)

evap_model4 <- lm(evap_mm ~ month + min_t_C + month:humid_9am, data = mwc)
summary(evap_model4)
anova(evap_model4)

#Check it:

step(evap_model1, direction = "backward")
ols_step_backward_p(evap_model1, prem= 0.05)
ols_check <-  lm(evap_mm ~ month + min_t_C + humid_9am + month:humid_9am, data = mwc)
step_check <-  lm(evap_mm ~ month + min_t_C + humid_9am + month:humid_9am, data = mwc)
#the above two functions found the same answer. so is ours statistically
differnt?
evap_model4 <- lm(evap_mm ~ month + min_t_C + month:humid_9am, data = mwc)
anova(ols_check, evap_model4)
#ANOVA finds no statistically significant difference between these models.
```

Model Diagnostics

```
plot(evap_model4, which = 1) #linearity

plot(evap_model4, which = 3) #homoscedasticity

plot(evap_model4, which = 2) #noise is normally distributed
```

Prediction

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.

```
predict(evap_model4, newdata = tibble(month = "2", min_t_C = 13.8,
humid_9am = 74), interval = "prediction")
```

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.

```
predict(evap_model4, newdata = tibble(month = "12", min_t_C = 16.4,
humid_9am = 57), interval = "prediction")
```

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

```
predict(evap_model4, newdata = tibble(month = "1", min_t_C = 26.5,
humid_9am = 35), interval = "prediction")
```

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.

```
predict(evap_model4, newdata = tibble(month = "7", min_t_C = 6.8,
humid_9am = 76), interval = "prediction")
```

```
citation("tidyverse")
citation("dplyr")
citation("lubridate")
citation("stringr")
citation("inspectdf")
citation("olsrr")
```