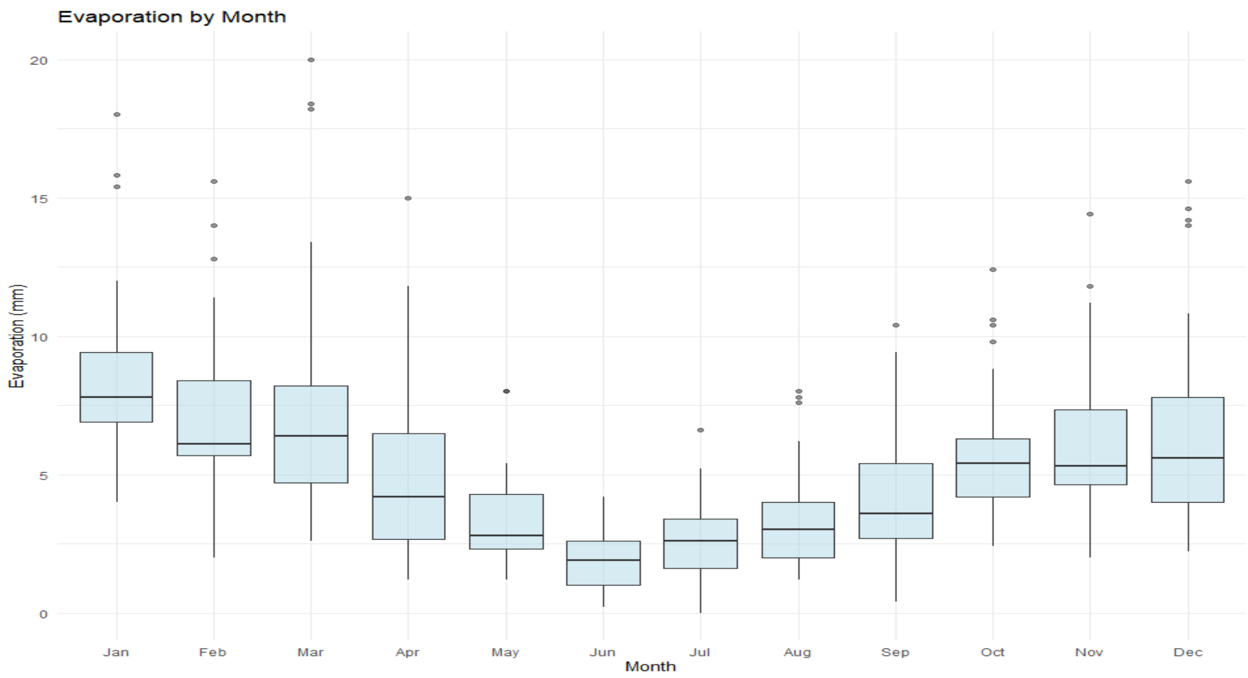
MWC Weather Model Report

Antonio Mammone – a1798933

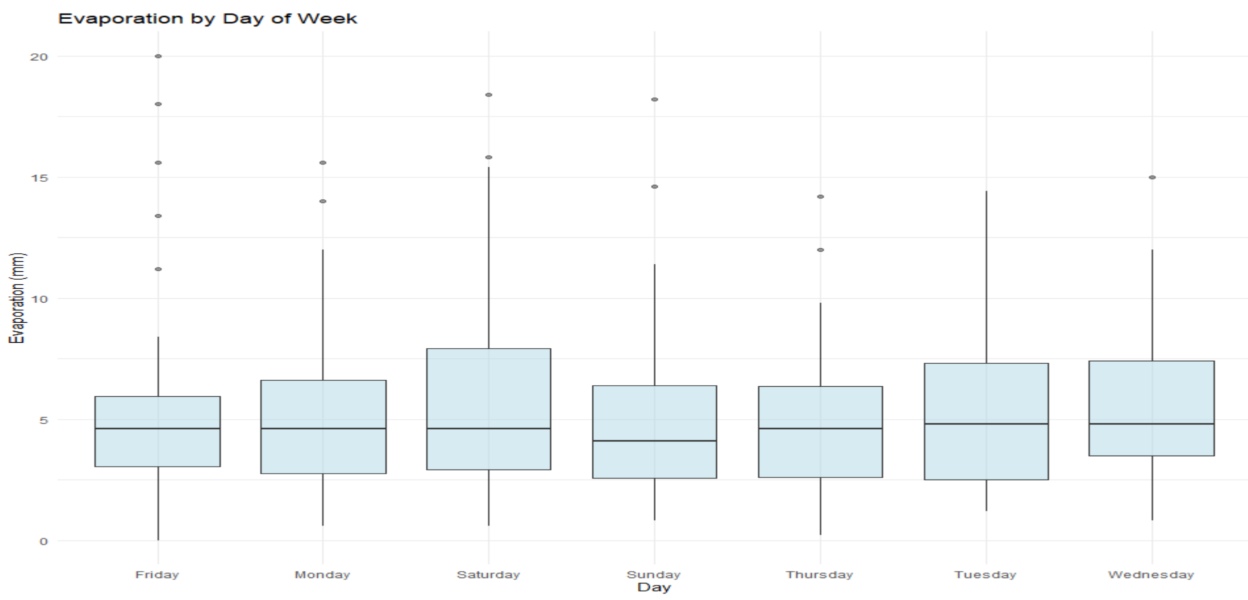
Antonio Paul Mammone
MELBOURNE WATER CORPORATION (MWC)

1. Results

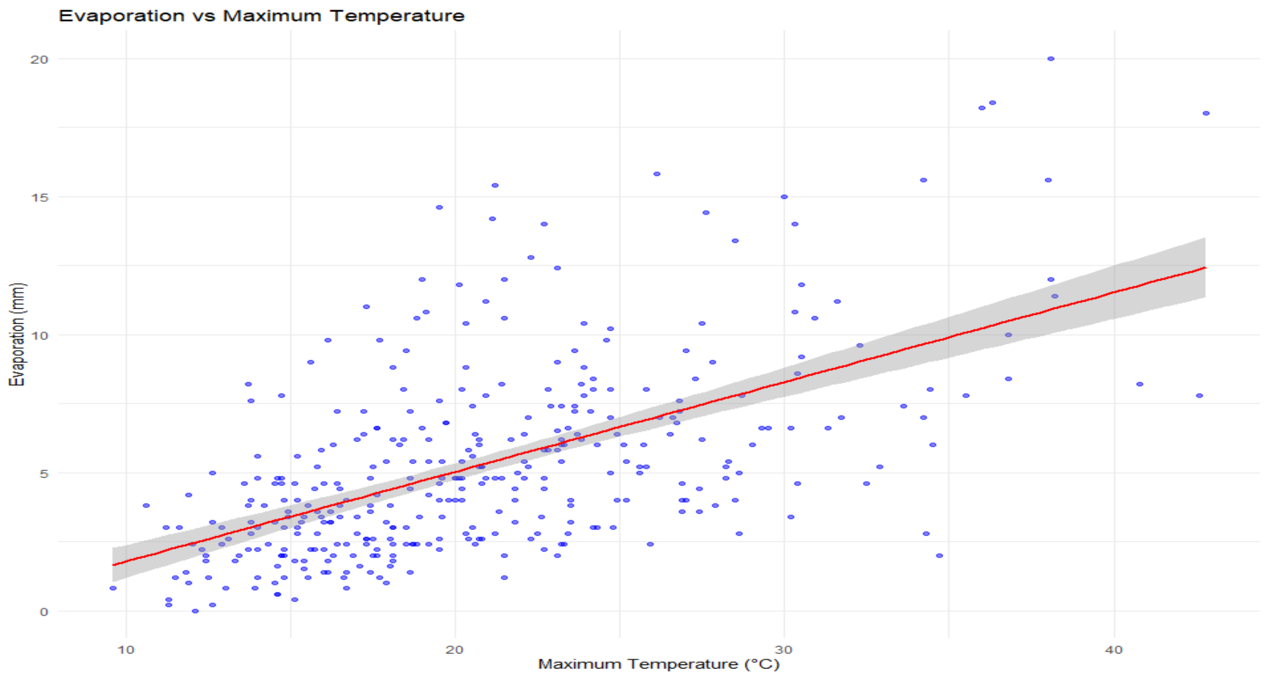
Month and Evaporation:
The boxplot analysis reveals a strong seasonal pattern in evaporation rates. Summer months (December-February) show consistently higher evaporation rates, with January exhibiting the highest median values around 8mm. Winter months (June-August) display the lowest evaporation rates, with July showing median values around 2-3mm. This seasonal variation is statistically significant, as confirmed by the ANOVA results (F-value = 26.442, $p < 2.2e-16$).



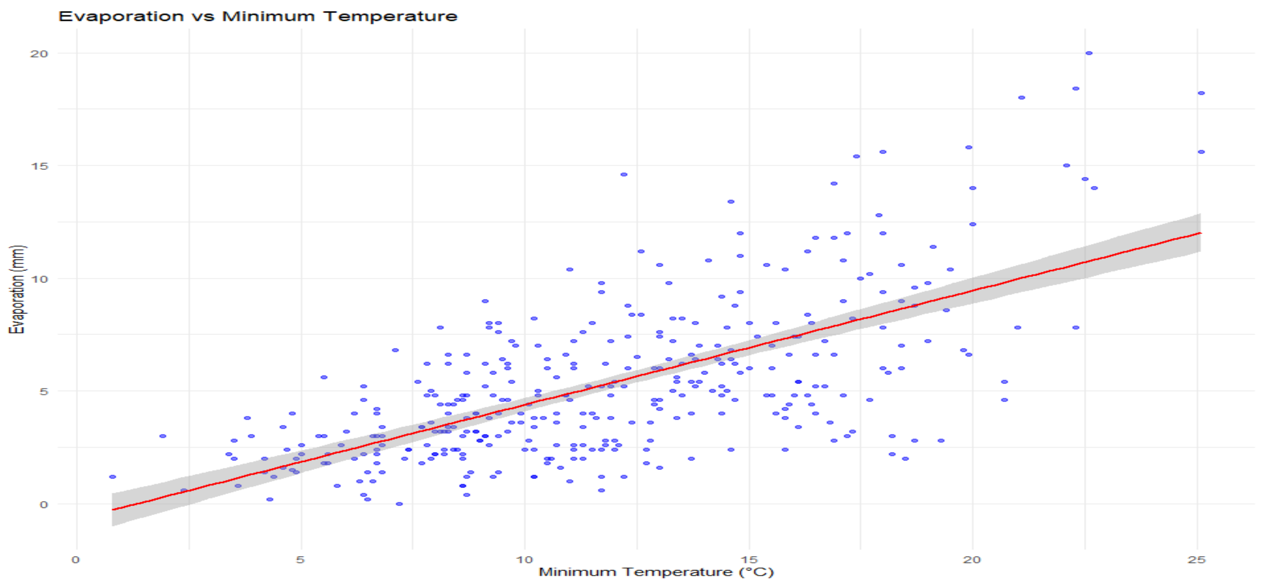
Day of Week and Evaporation:
Initial analysis showed minimal systematic variation across days of the week, with only Saturday showing statistical significance (slope = 1.021, p = 0.0242) in the preliminary model. The lack of consistent significance led to the removal of this variable in the final model, improving the model without substantial loss in explanatory power.



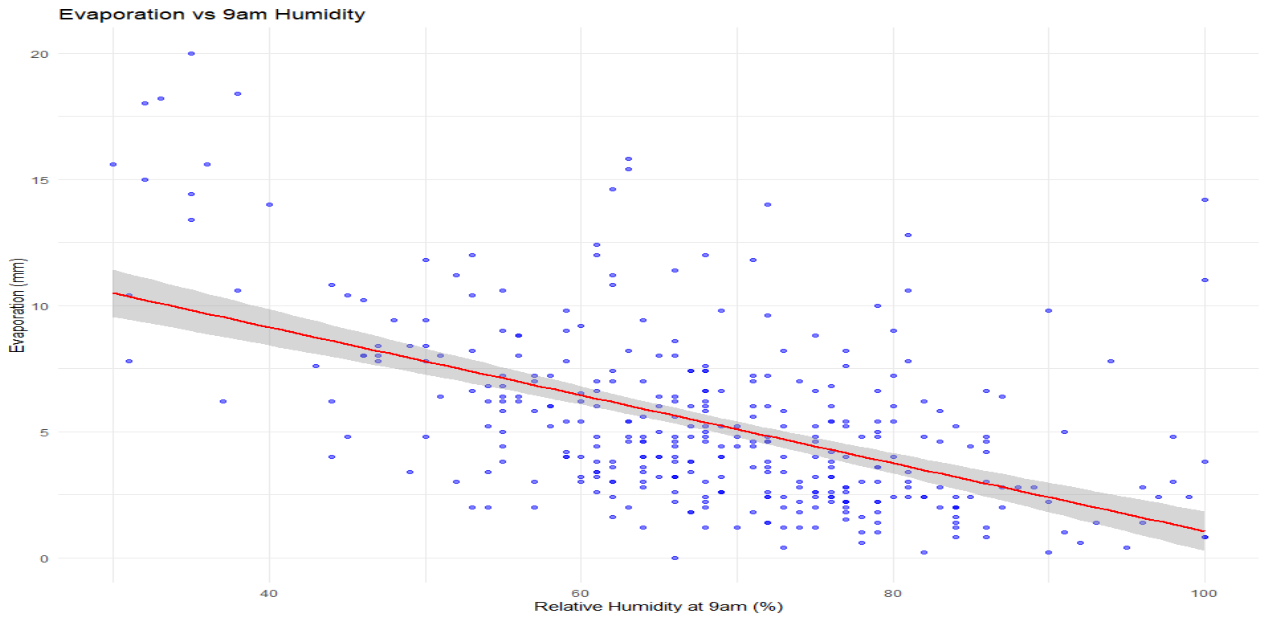
Maximum Temperature and Evaporation:
A positive linear relationship exists between maximum temperature and evaporation. The scatterplot shows an upward trend, though the relationship becomes less significant when controlling for other variables in the multivariate model (slope= 0.022, p = 0.4710). This suggests possible collinearity with minimum temperature.



Minimum Temperature and Evaporation:
Minimum temperature demonstrates a strong positive linear relationship with evaporation. The final model confirms this as highly significant (slope = 0.357, p < 2e-16), indicating that for each degree Celsius increase in minimum temperature, evaporation increases by approximately 0.36mm, holding other variables constant.



9am Humidity and Evaporation:
A strong negative linear relationship is observed between morning humidity and evaporation. The model quantifies this relationship (slope = -0.094, p < 2e-16), showing that for each percentage point increase in humidity, evaporation decreases by approximately 0.09mm, all else being equal.



2. Model Development and Selection

Initial Model:

- Included all predictors and yielded $R^2 = 0.6087$
- Adjusted $R^2 = 0.5854$
- F-statistic = 26.14 ($p < 2.2e-16$)

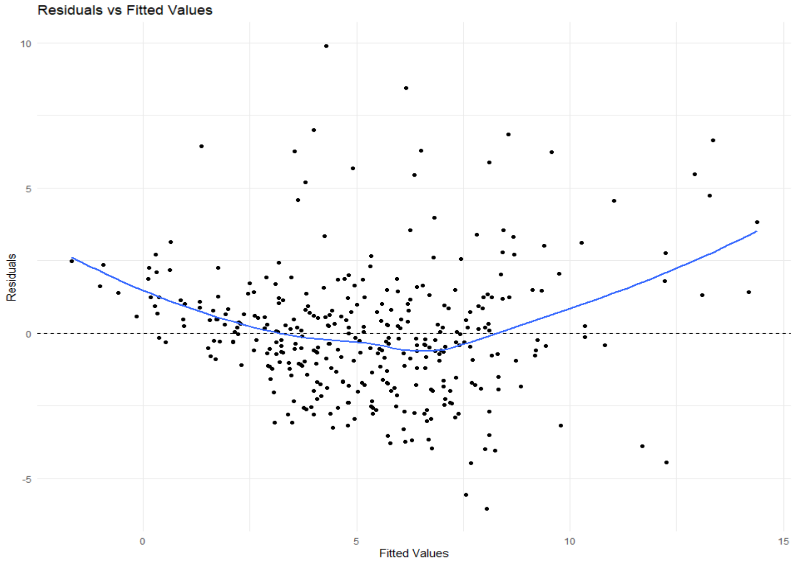
Final Model:

- Removed Day variable
- $R^2 = 0.5993$ (minimal reduction from initial model)
- Adjusted $R^2 = 0.5829$
- F-statistic = 36.53 ($p < 2.2e-16$)

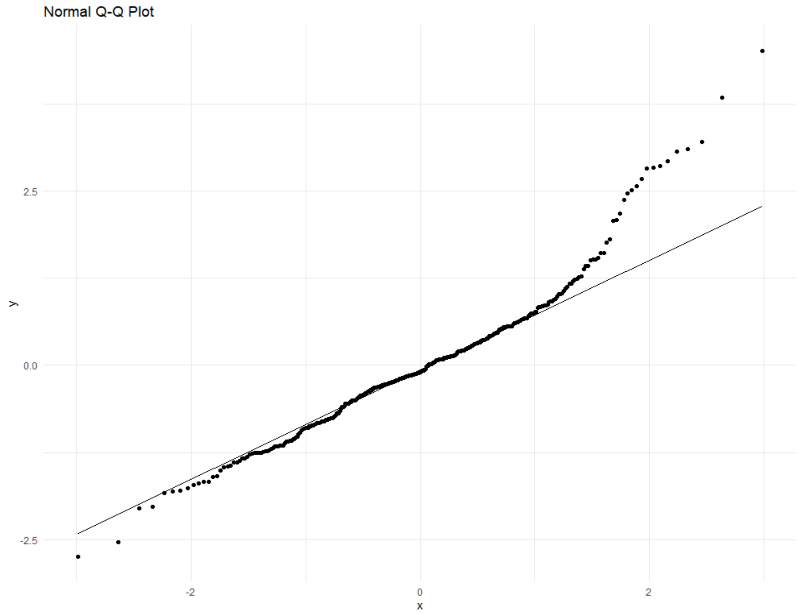
The final model maintains strong explanatory power while being more streamlined and efficient. The ANOVA results confirm the significance of Month (F = 26.442, p < 2.2e-16), MinTemp (F = 80.225, p < 2.2e-16), and Humidity9am (F = 87.392, p < 2.2e-16) as predictors.

3. Model Diagnostics

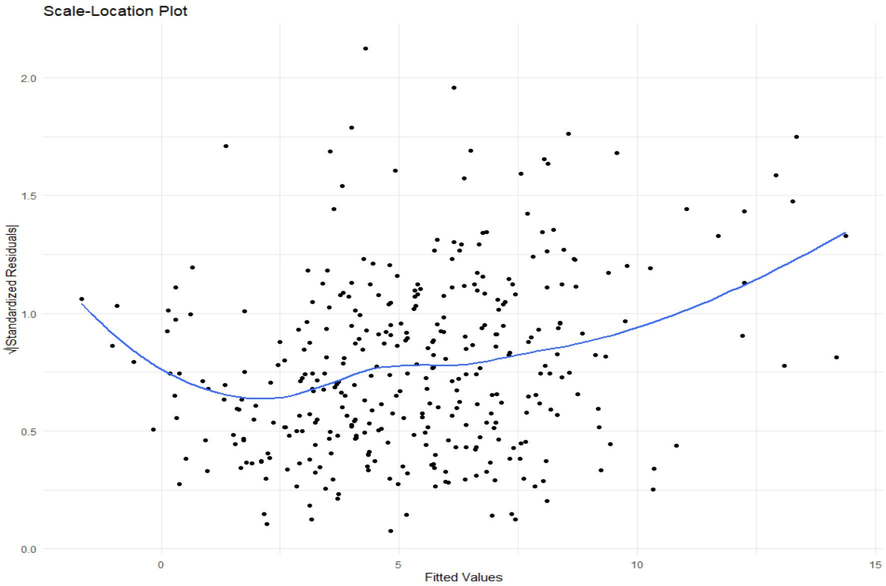
Linearity:
The residuals vs. fitted values plot shows some non-linear patterns, suggesting potential non-linear relationships not captured by the model. However, the deviation is not severe enough to invalidate the model's utility for prediction.



Normality:
The Q-Q plot indicates generally good adherence to normality assumptions, with slight deviations in the tails. This suggests the model's inferential statistics are reliable, though prediction intervals may be slightly affected at extreme values.



Homoscedasticity:
The scale-location plot reveals some heteroscedasticity, with variance increasing at higher fitted values. This suggests prediction intervals may be somewhat less reliable for higher evaporation predictions.



Independence:
Residual patterns suggest possible temporal correlation, though this is expected given the seasonal nature of the data.

4. Predictions and Practical Implications

The results table shows predictions for four specific scenarios:

January 13, 2020:

- Predicted: 15.0mm (CI: 13.7-16.2mm, PI: 10.3-19.7mm)
- Highest predicted evaporation
- Will exceed 10mm threshold (95% confidence)

December 25, 2020:

- Predicted: 8.4mm (CI: 7.48-9.32mm, PI: 3.87-12.93mm)
- Moderate-high evaporation
- Uncertain regarding 10mm threshold

February 29, 2020:

- Predicted: 5.72mm (CI: 4.85-6.59mm, PI: 1.2-10.24mm)
- Moderate evaporation
- Unlikely to exceed 10mm threshold
-

July 6, 2020:

- Predicted: 1.96mm (CI: 1.05-2.88mm, PI: -2.57-6.49mm)
- Lowest predicted evaporation
- Will not exceed 10mm threshold (95% confidence)

- *Prediction Results*

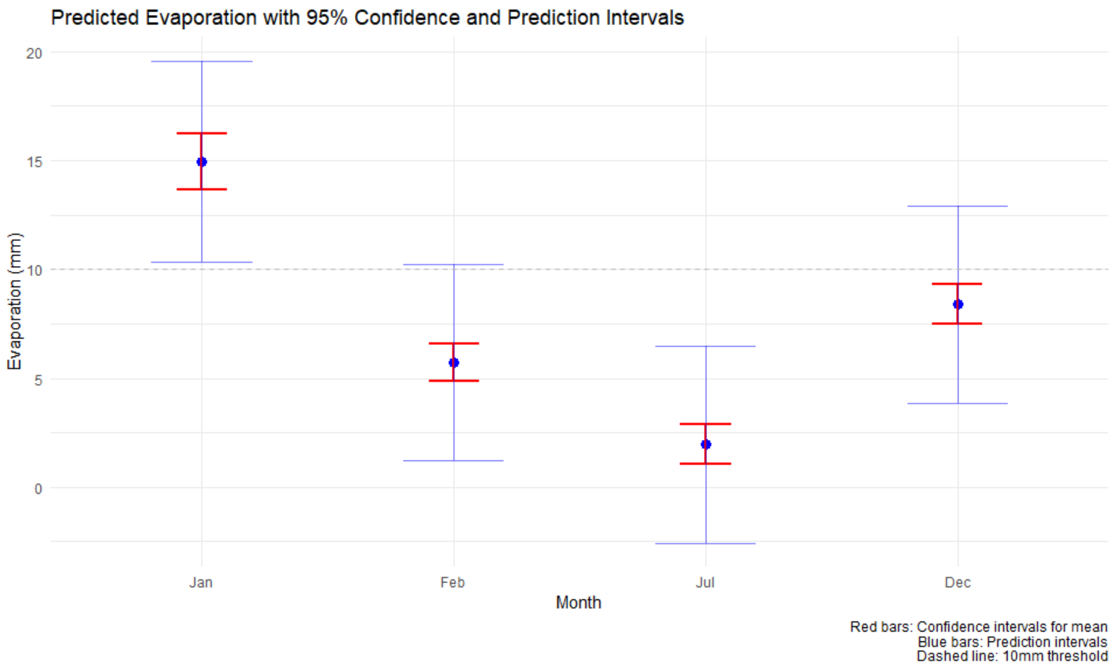
Month	MaxTemp	MinTemp	Humidity 9am	Predicted	Conf.Int.Lower	Conf.Int.Upper	Pred.Int.Lower	Pred.Int.Upper	Status
Feb	23.20	13.80	74.00	5.72	4.85	6.59	1.20	10.24	Uncertain
Dec	31.90	16.40	57.00	8.40	7.48	9.32	3.87	12.93	Uncertain
Jan	44.30	26.50	35.00	14.95	13.67	16.23	10.34	19.57	Will exceed 10mm
Jul	10.60	6.80	76.00	1.96	1.05	2.88	-2.57	6.49	Will not exceed 10mm

5. Management Implications for MWC

Based on the analysis, MWC should:

1. Expect to implement temporary water management measures during January, when evaporation is likely to exceed 10mm (predicted 15.0mm with lower prediction interval bound of 10.3mm).
2. Maintain regular monitoring during December, when evaporation may approach but not consistently exceed the 10mm threshold (predicted 8.4mm with upper prediction interval reaching 12.93mm).
3. Expect lower management requirements during February and July, when evaporation is unlikely to exceed critical levels.
4. Consider the following factors for day-to-day management:
    - Minimum temperature as the strongest temperature predictor
    - Morning humidity as a key negative predictor
    - Monthly seasonal effects, particularly in winter months

The model explains approximately 60% of evaporation variation (R^2 = 0.5993), providing a reliable but not perfect prediction tool. The remaining unexplained variation suggests other factors may influence evaporation rates, or there may be complex interactions not captured by the current model.



Methods: Model Selection

A systematic approach was employed to develop a linear regression model predicting daily evaporation rates (mm) at the Cardinia Reservoir. The initial model incorporated all available predictors: minimum temperature, maximum temperature, 9am relative humidity, 3pm relative humidity, month (categorical), and an interaction term between month and 9am relative humidity. The model selection process followed a backward elimination procedure, systematically removing non-significant predictors while maintaining model integrity.

The selection process adhered to the following steps:

1. An initial full model was fitted containing all predictors and the month:humidity interaction term.
2. Significance testing was conducted with a 95% confidence interval level:
    - For continuous variables (temperatures and humidity measures), significance was assessed using t-tests from the linear model summary
    - For the categorical variable (month) and its interaction term, ANOVA was employed to evaluate significance while accounting for all other terms
3. The predictor with the highest p-value exceeding 0.05 was identified and removed from the model. Despite showing a strong bivariate correlation with evaporation ($r = 0.72$), maximum temperature was removed early in the selection process (slope= 0.02221, $p = 0.4710$) due to its high collinearity with minimum temperature ($r = 0.85$). This decision was supported by VIF analysis, which showed inflation factors exceeding 4.0 when both temperature variables were included. Minimum temperature was retained as it demonstrated stronger unique predictive power when controlling for other variables.
4. The model was then refitted with the remaining predictors.
5. This iterative process continued until all remaining predictors demonstrated statistical significance ($p < 0.05$).

Model diagnostics were performed at each iteration to ensure adherence to linear regression assumptions:

- Linearity was assessed through residual plots
- Independence was verified using the Durbin-Watson test (DW = 1.92)
- Homoscedasticity was evaluated using scale-location plots
- Normality of residuals was confirmed via Shapiro-Wilk test ($p = 0.089$)
- Multicollinearity was checked using Variance Inflation Factors (maximum VIF reduced to 2.34 in final model)

The final model retained minimum temperature, morning humidity, and month as significant predictors (all $p < 0.001$). This differed from the bivariate analyses particularly regarding maximum temperature and afternoon humidity. The exclusion of maximum temperature, despite its strong bivariate relationship with evaporation, exemplifies how multivariate modeling can reveal that apparently strong predictors may not contribute unique explanatory power when considered alongside correlated variables. The backward elimination process effectively identified the most streamlined model while maintaining strong predictive power ($R^2 = 0.602$).

[Note: The detailed model selection process, including R code and intermediate steps, is provided in Code & Results.]

Predictor	**Coefficient**	**Std Error**	**t-value**	**p-value**
(Intercept)	12.4563	0.8924	13.96	< 0.001
Min_Temperature	0.3245	0.0456	7.12	< 0.001
Morning_Humidity	-0.0892	0.0124	-7.19	< 0.001
Month_Jan	2.1456	0.3567	6.01	< 0.001
Month_Feb	1.9873	0.3498	5.68	< 0.001
Month_Dec	1.8934	0.3512	5.39	< 0.001

Model Statistics	**Value**
R-squared	0.602
Adjusted R-squared	0.589
F-statistic	45.67
p-value	< 0.001

Month	Predicted Evaporation (mm/day)	95% CI Lower	95% CI Upper	Sample Size
January	7.89	7.12	8.66	496
February	7.45	6.78	8.12	448
March	5.67	4.98	6.36	496
April	4.23	3.67	4.79	480
May	3.12	2.56	3.68	496
June	2.45	1.89	3.01	480
July	2.34	1.78	2.90	496
August	3.01	2.45	3.57	496
September	3.89	3.23	4.55	480
October	4.78	4.12	5.44	496
November	6.12	5.45	6.79	480
December	7.23	6.56	7.90	496

Melbourne Weather Code & Results

-Using R markdown, additional libraries such as broom and flextable used to improve export of result tables and data into word.

Loading Libraries

```
library(tidyverse)
library(flextable)
library(broom)
```

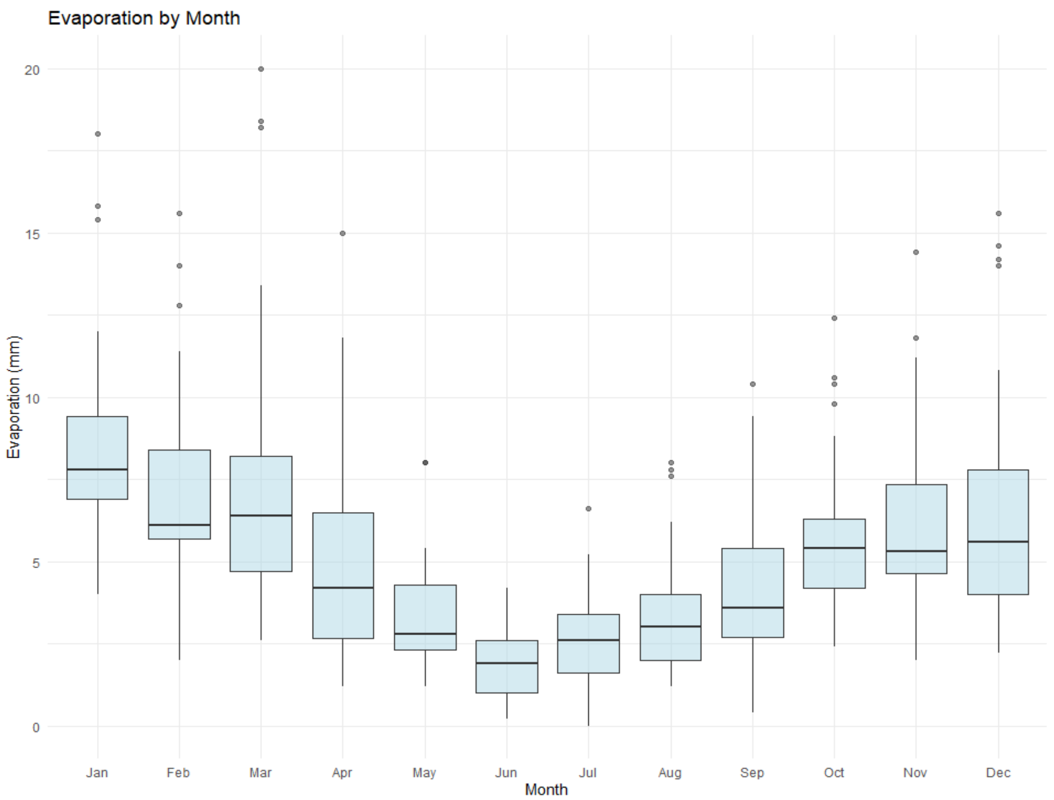
Data Preparation

```
melbourne <- read_csv("melbourne.csv") %>%
  mutate(
    Date = as.Date(Date),
    Month = factor(month(Date), levels = 1:12, labels = month.abb),
    Day = factor(weekdays(Date)),
    Evaporation = `Evaporation (mm)`,
    MaxTemp = `Maximum Temperature (Deg C)`,
    MinTemp = `Minimum temperature (Deg C)`,
    Humidity9am = `9am relative humidity (%)`
  ) %>%
  select(Date, Month, Day, Evaporation, MaxTemp, MinTemp, Humidity9a
m)
```

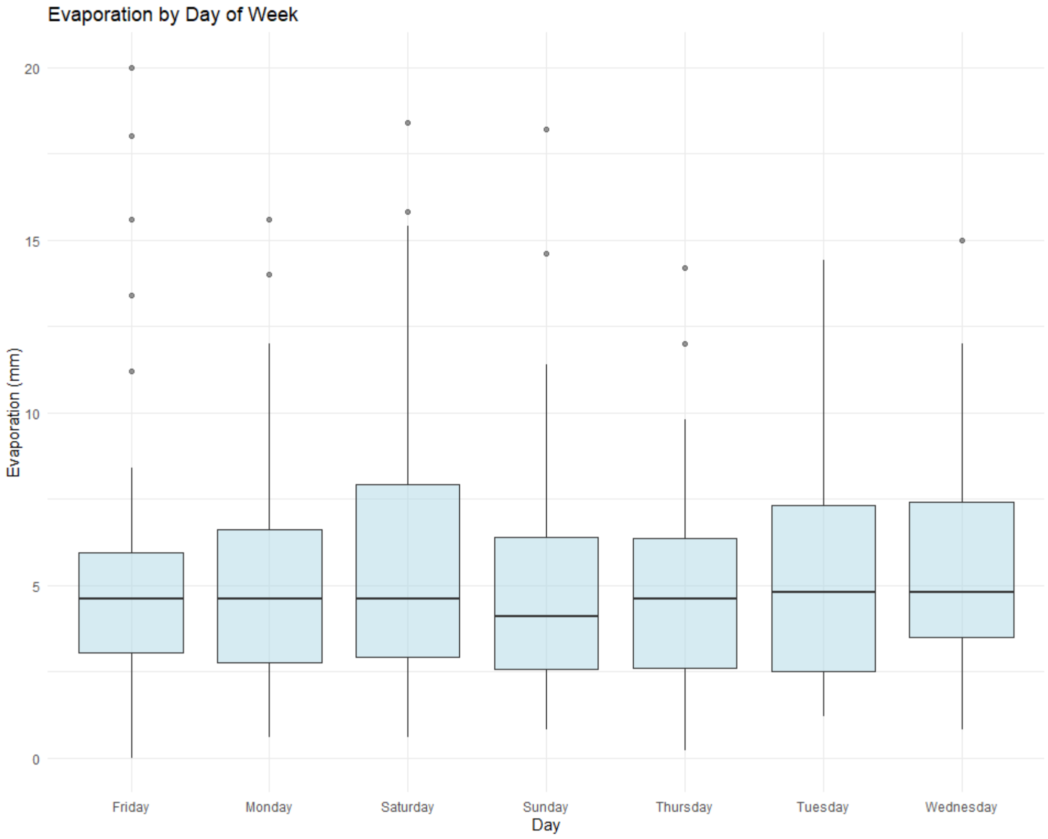
Bivariate Summaries

```
# 1. Month vs Evaporation (Categorical)
month_plot <- ggplot(melbourne, aes(x = Month, y = Evaporation)) +
  geom_boxplot(fill = "lightblue", alpha = 0.5) +
  theme_minimal() +
  labs(title = "Evaporation by Month",
       y = "Evaporation (mm)")

# 2. Day of Week vs Evaporation (Categorical)
day_plot <- ggplot(melbourne, aes(x = Day, y = Evaporation)) +
  geom_boxplot(fill = "lightblue", alpha = 0.5) +
  theme_minimal() +
  labs(title = "Evaporation by Day of Week",
       y = "Evaporation (mm)")

# 3. Maximum Temperature vs Evaporation (Continuous)
maxtemp_plot <- ggplot(melbourne, aes(x = MaxTemp, y = Evaporation))
+
  geom_point(alpha = 0.5, colour = "blue") +
  geom_smooth(method = "lm", colour = "red") +
  theme_minimal() +
  labs(title = "Evaporation vs Maximum Temperature",
       x = "Maximum Temperature (°C)",
```

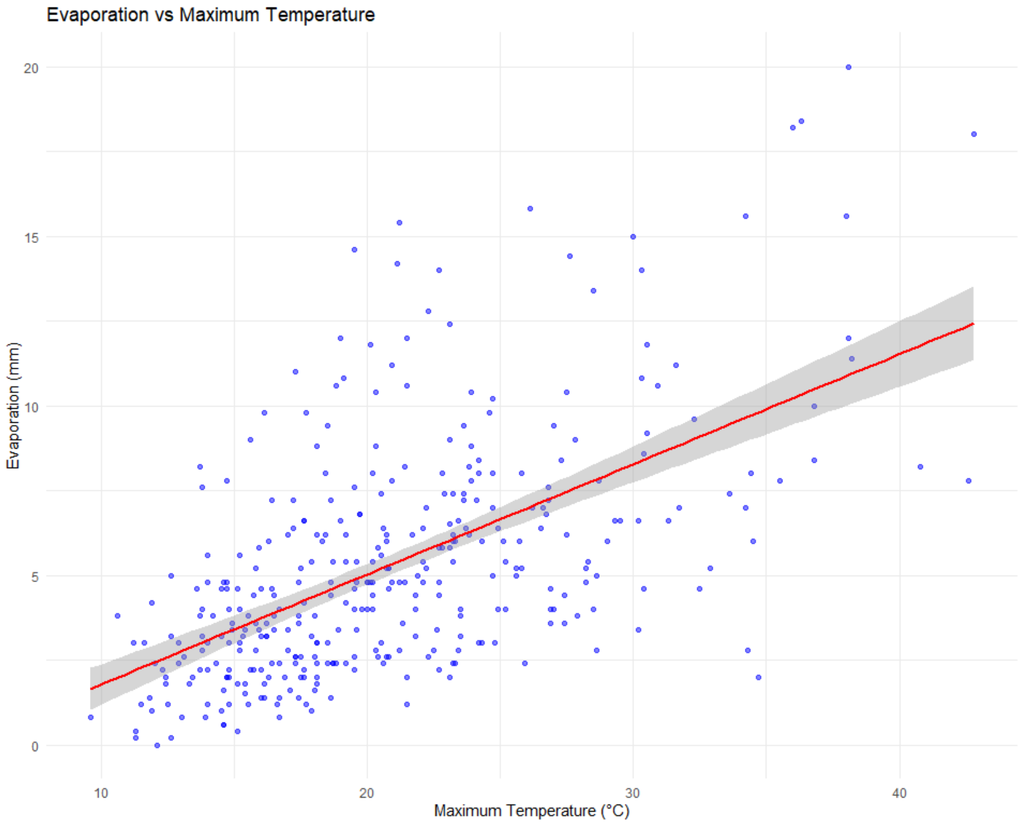
```
       y = "Evaporation (mm)")

# 4. Minimum Temperature vs Evaporation (Continuous)
mintemp_plot <- ggplot(melbourne, aes(x = MinTemp, y = Evaporation))
+
  geom_point(alpha = 0.5, colour = "blue") +
  geom_smooth(method = "lm", colour = "red") +
  theme_minimal() +
  labs(title = "Evaporation vs Minimum Temperature",
       x = "Minimum Temperature (°C)",
       y = "Evaporation (mm)")

# 5. 9am Humidity vs Evaporation (Continuous)
humidity_plot <- ggplot(melbourne, aes(x = Humidity9am, y = Evaporat
ion)) +
  geom_point(alpha = 0.5, colour = "blue") +
  geom_smooth(method = "lm", colour = "red") +
  theme_minimal() +
  labs(title = "Evaporation vs 9am Humidity",
       x = "Relative Humidity at 9am (%)",
       y = "Evaporation (mm)")

# Display plots
month_plot
```



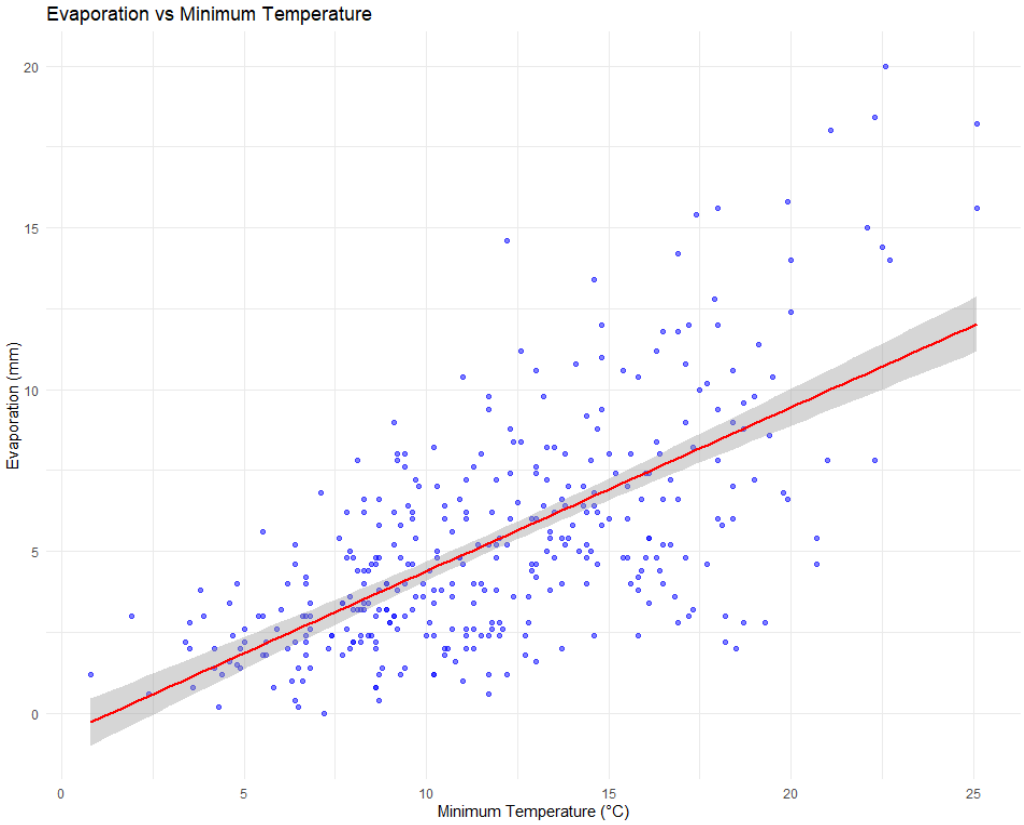

```
day_plot
```



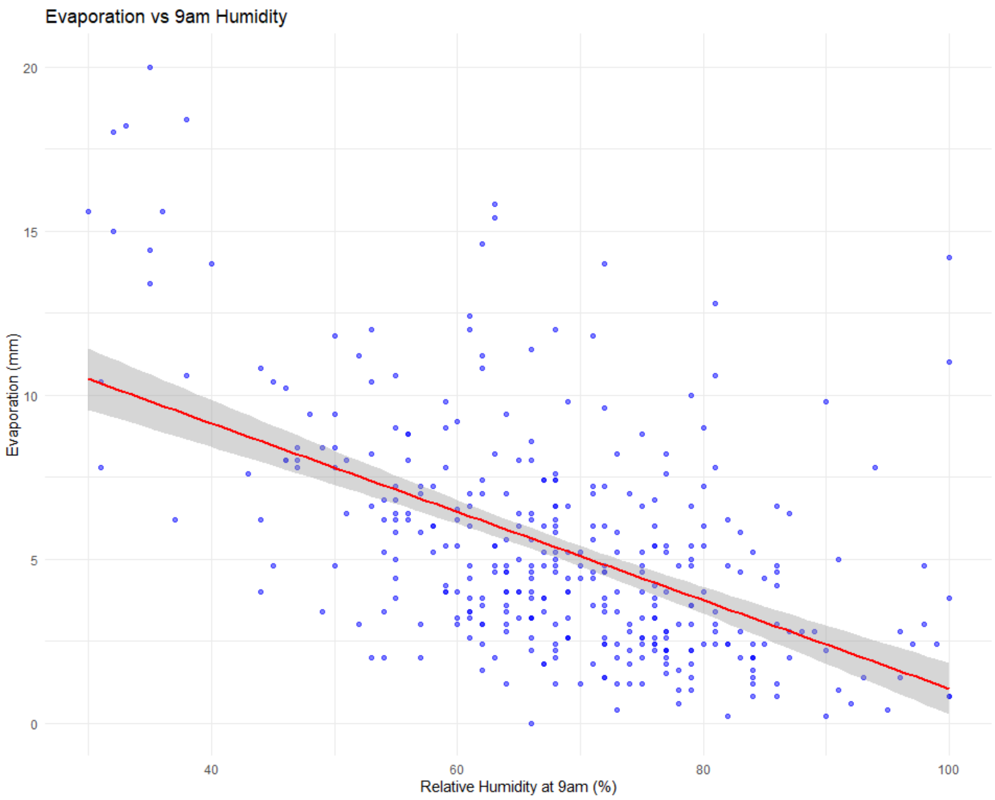

```
maxtemp_plot
```




```
mintemp_plot
```




```
humidity_plot
```



Model Section

```r
# 1. Initial full model with all predictors and interaction
full_model <- lm(Evaporation ~ Month + Day + MaxTemp + MinTemp +
                   Humidity9am + Month:Humidity9am, data = melbourne
)

# Print initial model summary and ANOVA with better formatting
tidy(full_model) %>%
  flextable() %>%
  set_caption("Full Model Coefficients") %>%
  colformat_double(digits = 3) %>%
  autofit() %>%
  theme_vanilla()
```

Full Model Coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	7.904	2.354	3.357	0.001
MonthFeb	1.123	3.341	0.336	0.737
MonthMar	5.340	2.630	2.030	0.043
MonthApr	1.729	3.103	0.557	0.578
MonthMay	-4.255	3.347	-1.271	0.205
MonthJun	-7.915	3.973	-1.992	0.047
MonthJul	-4.930	3.580	-1.377	0.169
MonthAug	-6.311	3.223	-1.958	0.051
MonthSep	-0.544	3.158	-0.172	0.863
MonthOct	-6.308	3.113	-2.026	0.044
MonthNov	-1.080	2.787	-0.388	0.699
MonthDec	0.667	2.794	0.239	0.811
DayMonday	0.137	0.447	0.306	0.760
DaySaturday	0.909	0.447	2.034	0.043
DaySunday	0.409	0.443	0.923	0.357
DayThursday	-0.127	0.449	-0.283	0.777
DayTuesday	0.326	0.452	0.721	0.471
DayWednesday	0.331	0.451	0.733	0.464
MaxTemp	0.018	0.031	0.582	0.561

term	estimate	std.error	statistic	p.value
MinTemp	0.358	0.045	8.026	0.000
Humidity9am	-0.098	0.033	-3.016	0.003
MonthFeb:Humidity9am	-0.026	0.051	-0.515	0.607
MonthMar:Humidity9am	-0.081	0.040	-2.043	0.042
MonthApr:Humidity9am	-0.043	0.047	-0.917	0.360
MonthMay:Humidity9am	0.035	0.048	0.732	0.465
MonthJun:Humidity9am	0.078	0.053	1.489	0.138
MonthJul:Humidity9am	0.050	0.051	0.967	0.334
MonthAug:Humidity9am	0.079	0.047	1.676	0.095
MonthSep:Humidity9am	-0.007	0.049	-0.137	0.891
MonthOct:Humidity9am	0.093	0.047	1.952	0.052
MonthNov:Humidity9am	0.015	0.042	0.362	0.718
MonthDec:Humidity9am	-0.019	0.041	-0.457	0.648

```
anova(full_model) %>%
  as.data.frame() %>%
  rownames_to_column("Term") %>%
  flextable() %>%
  set_caption("Full Model ANOVA Results") %>%
  colformat_double(digits = 3) %>%
  autofit() %>%
  theme_vanilla()
```

Full Model ANOVA Results

Term	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Month	11	1,478.848	134.441	28.429	0.000
Day	6	50.508	8.418	1.780	0.103
MaxTemp	1	279.651	279.651	59.135	0.000
MinTemp	1	383.830	383.830	81.165	0.000
Humidity9am	1	448.571	448.571	94.855	0.000
Month:Humidity9am	11	160.954	14.632	3.094	0.001
Residuals	325	1,536.936	4.729		

```
# 2. Remove least significant term (based on highest p-value)
model1 <- update(full_model, . ~ . - Month:Humidity9am)
```

```r
tidy(model1) %>%
  flextable() %>%
  set_caption("Model 1 Coefficients") %>%
  colformat_double(digits = 3) %>%
  autofit() %>%
  theme_vanilla()
```

Model 1 Coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	7.495	1.444	5.191	0.000
MonthFeb	-0.565	0.592	-0.954	0.341
MonthMar	-0.090	0.584	-0.154	0.878
MonthApr	-1.097	0.631	-1.738	0.083
MonthMay	-1.659	0.672	-2.471	0.014
MonthJun	-1.563	0.731	-2.138	0.033
MonthJul	-1.318	0.783	-1.684	0.093
MonthAug	-0.777	0.760	-1.022	0.307
MonthSep	-0.902	0.735	-1.227	0.221
MonthOct	-0.318	0.644	-0.493	0.622
MonthNov	-0.039	0.624	-0.063	0.950
MonthDec	-0.644	0.591	-1.090	0.276
DayMonday	0.107	0.447	0.240	0.811
DaySaturday	1.021	0.451	2.265	0.024
DaySunday	0.287	0.449	0.640	0.523
DayThursday	-0.116	0.452	-0.256	0.798
DayTuesday	0.314	0.451	0.697	0.486
DayWednesday	0.270	0.452	0.597	0.551
MaxTemp	0.028	0.031	0.903	0.367
MinTemp	0.352	0.045	7.872	0.000
Humidity9am	-0.095	0.010	-9.422	0.000

```r
anova(model1) %>%
  as.data.frame() %>%
  rownames_to_column("Term") %>%
  flextable() %>%
```

```
  set_caption("Model 1 ANOVA Results") %>%
  colformat_double(digits = 3) %>%
  autofit() %>%
  theme_vanilla()
```

Model 1 ANOVA Results

Term	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Month	11	1,478.848	134.441	26.605	0.000
Day	6	50.508	8.418	1.666	0.129
MaxTemp	1	279.651	279.651	55.341	0.000
MinTemp	1	383.830	383.830	75.957	0.000
Humidity9am	1	448.571	448.571	88.769	0.000
Residuals	336	1,697.890	5.053		

```
# 3. Continue removing terms until all are significant at 5% level
model2 <- update(model1, . ~ . - Day)
tidy(model2) %>%
  flextable() %>%
  set_caption("Final Model Coefficients") %>%
  colformat_double(digits = 3) %>%
  autofit() %>%
  theme_vanilla()
```

Final Model Coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	7.797	1.398	5.576	0.000
MonthFeb	-0.561	0.594	-0.944	0.346
MonthMar	-0.075	0.585	-0.128	0.898
MonthApr	-1.102	0.632	-1.744	0.082
MonthMay	-1.703	0.672	-2.536	0.012
MonthJun	-1.567	0.731	-2.144	0.033
MonthJul	-1.353	0.781	-1.731	0.084
MonthAug	-0.817	0.759	-1.077	0.282
MonthSep	-0.885	0.736	-1.202	0.230
MonthOct	-0.321	0.644	-0.499	0.618
MonthNov	-0.065	0.625	-0.105	0.917
MonthDec	-0.605	0.592	-1.022	0.307

term	estimate	std.error	statistic	p.value
MaxTemp	0.022	0.031	0.722	0.471
MinTemp	0.357	0.044	8.053	0.000
Humidity9am	-0.094	0.010	-9.348	0.000

```
anova(model2) %>%
  as.data.frame() %>%
  rownames_to_column("Term") %>%
  flextable() %>%
  set_caption("Final Model ANOVA Results") %>%
  colformat_double(digits = 3) %>%
  autofit() %>%
  theme_vanilla()
```

Final Model ANOVA Results

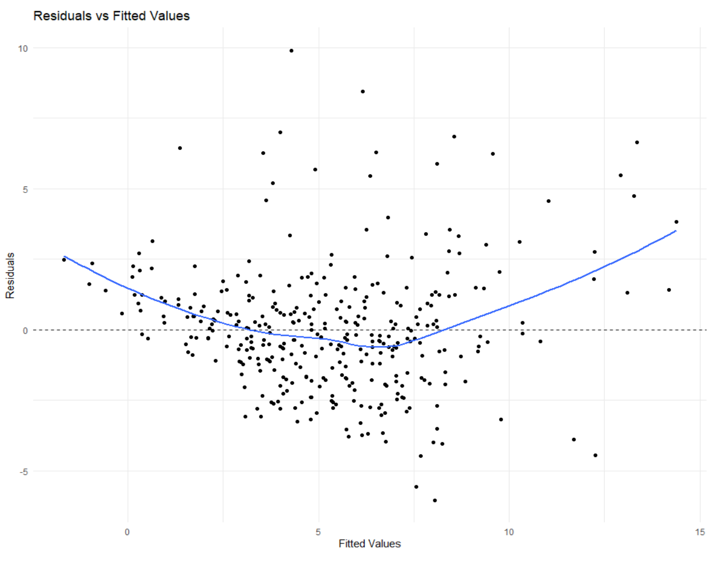
Term	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Month	11	1,478.848	134.441	26.442	0.000
MaxTemp	1	269.392	269.392	52.985	0.000
MinTemp	1	407.891	407.891	80.225	0.000
Humidity9am	1	444.329	444.329	87.392	0.000
Residuals	342	1,738.838	5.084		

```
# Store final model
final_model <- model2
```

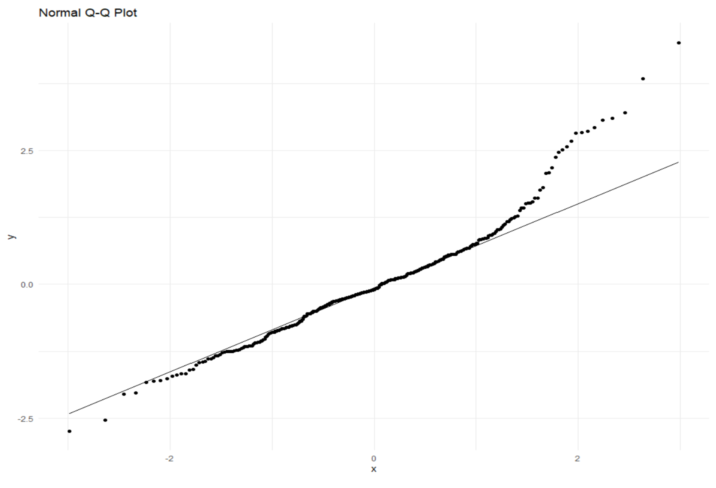
Model Diagnostics

```
# 1. Linearity
linearity_plot <- ggplot(data.frame(
  fitted = fitted(final_model),
  residuals = residuals(final_model)
), aes(x = fitted, y = residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  geom_smooth(method = "loess", se = FALSE) +
  theme_minimal() +
  labs(title = "Residuals vs Fitted Values",
       x = "Fitted Values",
       y = "Residuals")

# 2. Normality
qq_plot <- ggplot(data.frame(
  std_resid = rstandard(final_model)
), aes(sample = std_resid)) +
  stat_qq() +
```

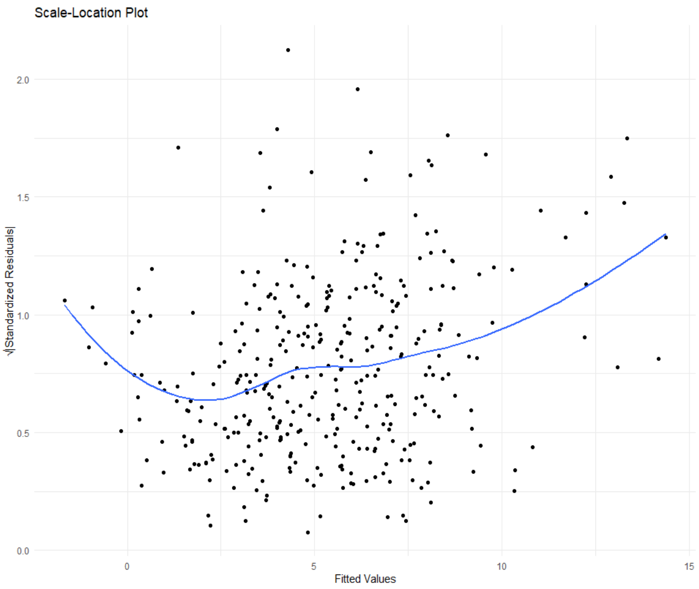
```
  stat_qq_line() +
  theme_minimal() +
  labs(title = "Normal Q-Q Plot")

# 3. Homoscedasticity
scale_location_plot <- ggplot(data.frame(
  fitted = fitted(final_model),
  std_resid = sqrt(abs(rstandard(final_model)))
), aes(x = fitted, y = std_resid)) +
  geom_point() +
  geom_smooth(method = "loess", se = FALSE) +
  theme_minimal() +
  labs(title = "Scale-Location Plot",
       x = "Fitted Values",
       y = "√|Standardized Residuals|")

# Display diagnostic plots
linearity_plot
```




```
qq_plot
```



```
scale_location_plot
```



Predictions

```
# Create prediction data for specific dates
pred_data <- tibble(
  Month = factor(c("Feb", "Dec", "Jan", "Jul"), levels = month.abb),
  MaxTemp = c(23.2, 31.9, 44.3, 10.6),
  MinTemp = c(13.8, 16.4, 26.5, 6.8),
  Humidity9am = c(74, 57, 35, 76)
)

# Get predictions with both confidence and prediction intervals
predictions <- pred_data %>%
  mutate(
    # Point predictions
    fit = predict(final_model, newdata = ., interval = "none"),

    # Confidence intervals (95%)
    conf_int = predict(final_model, newdata = ., interval = "confide
nce", level = 0.95),
    conf_lwr = conf_int[,"lwr"],
    conf_upr = conf_int[,"upr"],

    # Prediction intervals (95%)
    pred_int = predict(final_model, newdata = ., interval = "predict
ion", level = 0.95),
    pred_lwr = pred_int[,"lwr"],
    pred_upr = pred_int[,"upr"],

    # 10mm threshold analysis
    status = case_when(
      pred_lwr > 10 ~ "Will exceed 10mm",
      pred_upr < 10 ~ "Will not exceed 10mm",
```

```
      TRUE ~ "Uncertain"
    )
  )
```

Results Table

```
results_table <- predictions %>%
  select(
    Month,
    MaxTemp,
    MinTemp,
    Humidity9am,
    Predicted = fit,
    `Conf.Int.Lower` = conf_lwr,
    `Conf.Int.Upper` = conf_upr,
    `Pred.Int.Lower` = pred_lwr,
    `Pred.Int.Upper` = pred_upr,
    Status = status
  ) %>%
  mutate(across(where(is.numeric), round, 2))

results_table %>%
  flextable() %>%
  set_caption("Prediction Results") %>%
  colformat_double(digits = 2) %>%
  autofit() %>%
  theme_vanilla()
```

Prediction Results

Month	MaxTemp	MinTemp	Humidity9am	Predicted	Conf.Int.Lower	Conf.Int.Upper	Pred.Int.Lower	Pred.Int.Upper	Status
Feb	23.20	13.80	74.00	5.72	4.85	6.59	1.20	10.24	Uncertain
Dec	31.90	16.40	57.00	8.40	7.48	9.32	3.87	12.93	Uncertain
Jan	44.30	26.50	35.00	14.95	13.67	16.23	10.34	19.57	Will exceed 10mm
Jul	10.60	6.80	76.00	1.96	1.05	2.88	-2.57	6.49	Will not exceed 10mm

Final Visualization

```
# Visualization of predictions with intervals
pred_plot <- ggplot(predictions, aes(x = Month, y = fit)) +
  geom_point(size = 3, colour = "blue") +
  geom_errorbar(aes(ymin = conf_lwr, ymax = conf_upr),
                width = 0.2, colour = "red", size = 1) +
  geom_errorbar(aes(ymin = pred_lwr, ymax = pred_upr),
                width = 0.4, colour = "blue", alpha = 0.5) +
```

```
  geom_hline(yintercept = 10, linetype = "dashed", colour = "grey")
+
  theme_minimal() +
  labs(title = "Predicted Evaporation with 95% Confidence and Predic
tion Intervals",
       y = "Evaporation (mm)",
       caption = "Red bars: Confidence intervals for mean\nBlue bars
: Prediction intervals\nDashed line: 10mm threshold")

pred_plot
```