Assignment-4

Arya Pankaj Ranjan

2023-09-13

Executive Summary

This report addresses the Melbourne Water Corporation's (MWC) need for a reliable model to predict daily evaporation rates at their Cardinia Reservoir. The existing estimates have become unreliable due to recent climate changes. The goal of this report is to build a robust predictive model for evaporation based on various meteorological and temporal factors.

We begin by conducting bivariate summaries to explore the relationships between evaporation and potential predictors, which include month, day of the week, hours of bright sunlight, maximum wind gust speed, minimum temperature, and 9am relative humidity. Our analysis reveals key insights into these relationships.

Subsequently, we proceed with model selection by employing a stepwise approach, guided by the results of the Analysis of Variance (ANOVA),we build a model that incorporates significant predictors. In this process, we also explore the interaction effect between month and 9 am relative humidity. The ANOVA helps us assess the significance of each predictor's contribution to the model, allowing us to iteratively refine the model.

Interpreting the coefficients of our model, we provide practical insights for MWC. We explain how each predictor impacts evaporation, allowing MWC to understand and prepare for changing conditions.

To ensure the model's reliability, we thoroughly test its assumptions, including linearity, independence, and homoscedasticity. This step ensures the model's validity and informs its limitations.

Finally, we present predictions for specific dates in 2024, along with prediction intervals to account for uncertainty. We identify which days are likely to require water management interventions based on a 95% confidence threshold.

Our report adheres to the requested format, providing an executive summary, methods, results, discussion, and conclusions. It also includes an appendix with all R code used for analysis. This comprehensive approach equips MWC with the tools needed to effectively manage water resources in Melbourne.

Methods

Bivariate summaries

Over here, to explore the relationships between the response variable (evaporation) and predictor variables (Month, Day of the week, Sunlight hours, Wind gust speed, Minimum temperature, Relative humidity at 9am), we have used various types of plots depending on the nature of the predictors

1. Month and Evaporation- Over here we have generated box plots for evaporation in every month. We can see that there are outliers present in every month except June and September. The highest spread has occurred in the month of April.

2. Week and Evaporation- Over here we have generated a boxplot to analyse the relationship between the days in the week and evaporation. The highest median evaporation rate occurs on Thursdays with the highest evaporation occurring on friday with the rate of 20mm, Friday's also have the highest number of outliers. Saturday has the maximum spread in terms of evaporation compared to other weeks. Most of the data has a normal distribution except Monday, Thursdays and Wednesday.

3. Bright Sunlight hours and evaporation- Over here we have generated a scatterplot to analyse the relationship between the bright sunlight hours and evaporation. As we can see that there is a very weak association between them but there is a slight positive correlation between the sunlight hours and evaporation.

4. Max Wind Gust speed and evaporation- Over here we have generated a scatterplot to analyse the relationship between the max wind gust speed and evaporation. As shown in the scatterplot, there is a weak association between them and there is a slight positive correlation between the max wind gust speed and evaporation.

5. Minimum Temperature and Evaporation- Over here we have generated a scatterplot to analyse the relationship between the minimum temperature and evaporation. As shown in the scatterplot, there is a strong association between them and their correlation is positive.

6. relative humidity at 9 a.m and Evaporation- Over here we have generated a scatterplot to analyse the relationship between the relative humidity at 9 am and evaporation. As shown in the scatterplot, there is a weak association between them and their correlation is negative.

Model Selection

We conducted an ANOVA test to calculate the significant terms which we found out to be-

Month.L: January- 2.49e-05 Month.Q: February- 0.056404 Monthˆ4: April- 0.001399 DayOfWeekSaturday: Saturday- 0.019181 Speed.of.maximum.wind.gust..km.h.: Maximum Wind gust Speed- 0.000161 Minimum.temperature..Deg.C.: Minimum Temperature in Degrees- 6.58e-06 X9am.relative.humidity. . . .: Relative Humidity at 9 am- 1.81e-12

final model accounts for the simultaneous influence of multiple predictors on the response variable "Evaporation..mm." It considers how each predictor affects the response while controlling for the presence of other predictors.

From the bivariate analysis we conducted above we can see that the month January, February and April have a statistical significance with Evaporation and when it comes to days of week Saturday has a statistical significance, and predictors such as maximum gust wind speed, Minimum temperature in degree Celsius and relative humidity at 9 am have a statistical significance with the response variable evaporation

Model Diagnostics

Linearity Assumption: we have created fitted vs residual plot in order to interpret the assumpution of linearity, It is shown that the assumption of linearity is violated as the smoothing like deviates from the best fit line (zero line). This suggests that there is a non linear relationship between the response variable and the predictor variable.

Independence: We have used to Durbin Watson test for the assumption of independence, as you can see in the test results that DW statistic is close to 2 (around 2.0654), which suggests that there is no strong evidence of positive autocorrelation in the residuals. The p-value of 0.7037 is relatively high, indicating that there is not enough evidence to reject the null hypothesis of no positive autocorrelation. This is generally a good result, as it suggests that the residuals in your model are relatively independent.

Homoscedasticity Assumption: As we can see in the fitted vs residual plot, our assumption of homoscedasticity is violated, the data is scattered across the reference line, This suggests that the variance of the residuals

may not be constant across all levels of the predictors. Heteroscedasticity (non-constant variance) can lead to biased standard errors and affect the validity of hypothesis tests.

Normality Assumption- As we can see in the QQ-plot, The data is normally distributed and most of the points fall under the 45 degree references line, Hence the assumption of normality is met.

In summary, your interpretation indicates that the linearity assumption may be violated due to a non-linear relationship between some predictor variables and the response variable. However, the independence assumption is met, which is good for the validity of the model. The homoscedasticity assumption appears to be violated, suggesting potential issues with constant variance. Finally, the normality assumption is met, indicating that the residuals are approximately normally distributed

Results

Model Interpretation

Over here we will be interpreting the intercept of each coefficient-

Month L- January: This coefficients represent how January affects evaporation compared to a reference month. For example, "Month.L" represents January. the coefficient for "Month.L" is -1.01529, it means that, on average, January tends to have 1.01529 mm less evaporation compared to the reference month.

Month Q- February: This coefficients represent how February affects evaporation compared to a reference month. For example, "Month.Q" represents February. the coefficient for "Month.Q" is 2.66410, it means that, on average, February tends to have 2.66410 mm more evaporation compared to the reference month.

Month C- March: This coefficients represent how March affects evaporation compared to a reference month. For example, "Month.C" represents March. the coefficient for "Month.C" is 0.09673, it means that, on average, March tends to have 0.09673 mm more evaporation compared to the reference month.

Month^4- April: This coefficients represent how April affects evaporation compared to a reference month. For example, "Month^4" represents April. the coefficient for "Month^4" is -1.50899, it means that, on average, April tends to have -1.50899 mm less evaporation compared to the reference month.

Month^5- May: This coefficients represent how May affects evaporation compared to a reference month. For example, "Month^5" represents May. the coefficient for "Month^5" is -0.28446, it means that, on average, May tends to have -0.28446 mm less evaporation compared to the reference month.

Month^6- June: This coefficients represent how June affects evaporation compared to a reference month. For example, "Month^6" represents June. the coefficient for "Month^6" is 0.54251, it means that, on average, June tends to have 0.54251 mm more evaporation compared to the reference month.

Month^7- July: This coefficients represent how July affects evaporation compared to a reference month. For example, "Month^7" represents July. the coefficient for "Month^7" is -0.59993, it means that, on average, July tends to have -0.59993 mm less evaporation compared to the reference month.

Month^8- August: This coefficients represent how August affects evaporation compared to a reference month. For example, "Month^8" represents August. the coefficient for "Month^8" is 0.67723, it means that, on average, August tends to have 0.67723 mm more evaporation compared to the reference month.

Month^9- September: This coefficients represent how September affects evaporation compared to a reference month. For example, "Month^9" represents September. the coefficient for "Month^9" is 0.20853, it means that, on average, September tends to have 0.20853 mm more evaporation compared to the reference month.

Month^10- October: This coefficients represent how October affects evaporation compared to a reference month. For example, "Month^10" represents October. the coefficient for "Month^10" is -0.11797, it means that, on average, October tends to have -0.11797 mm less evaporation compared to the reference month.

Month^11- November: This coefficients represent how November affects evaporation compared to a reference month. For example, “Month^11” represents November. the coefficient for “Month^11” is 0.23809, it means that, on average, November tends to have 0.23809 mm more evaporation compared to the reference month.

DayOfWeekMonday- Monday: This coefficients represent how different days of the week affect evaporation compared to a reference day. For instance, if the coefficient for “DayOfWeekMonday” is 0.42338, it means that, on average, Monday tend to have 0.42338 mm more evaporation compared to the reference day.

DayOfWeekTuesday- Tuesday: This coefficients represent how different days of the week affect evaporation compared to a reference day. For instance, if the coefficient for “DayOfWeekTuesday” is 0.37590, it means that, on average, Tuesday tend to have 0.37590 mm more evaporation compared to the reference day.

DayOfWeekWednesday- Wednesday: This coefficients represent how different days of the week affect evaporation compared to a reference day. For instance, if the coefficient for “DayOfWeekWednesday” is 0.19458, it means that, on average, Wednesday tend to have 0.19458 mm more evaporation compared to the reference day.

DayOfWeekThursday- Thursday: This coefficients represent how different days of the week affect evaporation compared to a reference day. For instance, if the coefficient for “DayOfWeekThursday” is 0.02653, it means that, on average, Thursday tend to have 0.02653 mm more evaporation compared to the reference day.

DayOfWeekSunday- Sunday: This coefficients represent how different days of the week affect evaporation compared to a reference day. For instance, if the coefficient for “DayOfWeekSunday” is 0.41357, it means that, on average, Sunday tend to have 0.41357 mm more evaporation compared to the reference day.

DayOfWeekSaturday- Saturday: This coefficients represent how different days of the week affect evaporation compared to a reference day. For instance, if the coefficient for “DayOfWeekSaturday” is 1.14277, it means that, on average, saturday tend to have 1.14277 mm more evaporation compared to the reference day.

Sunshine..hours.- Sunshine hours: For each additional hour of sunshine, evaporation tends to decrease by 0.03767 mm, on average.

Speed.of.maximum.wind.gust..km.h.- maximum wind gust speed: For each additional km/h in the speed of the maximum wind gust, evaporation tends to increase by 0.05128 mm, on average.

Minimum.temperature..Deg.C.- Minimum Temperature: For each additional degree Celsius in the minimum temperature, evaporation tends to increase by 0.24897 mm, on average

X9am.relative.humidity....- Relative humidity at 9 am: For each additional percentage point in relative humidity at 9 am, evaporation tends to decrease by 0.08820 mm, on average.

As you can look over here we have not mentioned the coefficient for December and friday as they are the reference months and days, and the other coefficients indicate how each of those months and days differs from January and friday in terms of their effect on the response variable.

Discussion

Date: 2019-06-29

Predicted Evaporation: 8.280 mm Lower Prediction Limit: 3.587 mm Upper Prediction Limit: 12.973 Temporary Measures Needed: No Interpretation:

On June 29, 2019, the model predicts a low evaporation of approximately 8.280 mm. The lower prediction limit is 3.587 mm, indicating that with a certain level of confidence, the actual evaporation is expected to be at least 3.587 mm. The upper prediction limit is 12.973 mm, suggesting that with the same level of confidence, the actual evaporation is not expected to exceed 12.973 mm.

the “Temporary Measures Needed” column indicates that temporary measures are not needed on this date, likely due to low predicted evaporation rate.

Date: 2018-12-30

Predicted Evaporation: 17.477 Lower Prediction Limit: 12.996 Upper Prediction Limit: 21.957 Temporary Measures Needed: "Yes" Interpretation:

the model predicts a low evaporation of approximately 17.477 mm. The lower prediction limit is 12.996 mm, indicating that with a certain level of confidence, the actual evaporation is expected to be at least 12.996 mm. The upper prediction limit is 21.957 mm, suggesting that with the same level of confidence, the actual evaporation is not expected to exceed 21.957 mm.

the "Temporary Measures Needed" column indicates that temporary measures are needed on this date, likely due to the high predicted evaporation.

Date: 2018-12-08

Predicted Evaporation: 15.514 mm Lower Prediction Limit: 11.054 mm Upper Prediction Limit: 19.974 mm Temporary Measures Needed: "Yes" Interpretation:

On December 8, 2018, the model predicts a high evaporation of approximately 15.514 mm. The lower prediction limit is 11.054 mm, indicating that with a certain level of confidence, the actual evaporation is expected to be at least 15.51 mm. The upper prediction limit is 19.974 mm, suggesting that with the same level of confidence, the actual evaporation is not expected to exceed 19.974 mm.

the "Temporary Measures Needed" column indicates that temporary measures are needed on this date, likely due to the high predicted evaporation.

Date: 2019-02-20

Predicted Evaporation: 12.707 mm Lower Prediction Limit: 8.126 mm Upper Prediction Limit: 17.288 mm Temporary Measures Needed: "No" Interpretation:

On February 20, 2019, the model predicts an evaporation of approximately 12.707 mm. The lower prediction limit is 8.126 mm, indicating that with a certain level of confidence, the actual evaporation is expected to be at least 8.126 mm. The upper prediction limit is 17.288 mm, suggesting that with the same level of confidence, the actual evaporation is not expected to exceed 17.288 mm.

the "Temporary Measures Needed" column indicates that temporary measures are needed on this date, likely due to the high predicted evaporation.

Conclusion

Bivariate Analysis: We explored the relationships between evaporation and various predictors, including month, day of the week, sunlight hours, wind gust speed, minimum temperature, and relative humidity at 9 am. Notably, we observed that some months, days of the week, and meteorological variables had significant associations with evaporation.

Model Selection: The ANOVA-based model selection process allowed us to identify the most significant predictors for evaporation. We found that the month of January, February, and April, along with specific days of the week, had a statistically significant impact on evaporation. Additionally, maximum wind gust speed, minimum temperature, and relative humidity at 9 am played crucial roles in predicting evaporation.

Model Diagnostics: We assessed the model's assumptions, including linearity, independence, homoscedasticity, and normality. While the linearity assumption showed some deviation, the independence assumption was met, indicating minimal autocorrelation in residuals. However, we observed a violation of the homoscedasticity assumption, suggesting that the variance of residuals may not be constant across predictor levels. Fortunately, the normality assumption was satisfied, indicating normally distributed residuals.

Model Interpretation: The coefficients of the final model provided insights into how each predictor influenced evaporation. For example, we found that January tends to have lower evaporation compared to other months,

while February and April exhibit higher evaporation. Similarly, certain days of the week, such as Saturday, contribute to increased evaporation.

Prediction Intervals: We provided prediction intervals for specific dates in 2024, allowing MWC to gauge the uncertainty around our evaporation predictions. Temporary measures are necessary when evaporation exceeds a certain threshold, ensuring a stable water supply. Our predictions help MWC make informed decisions on when such measures are needed.

Recommendations

Model Maintenance: We recommend that MWC regularly monitor the model's performance and update it as needed. Climate patterns and conditions may change over time, and model recalibration ensures accuracy and predictions.

Non-Linear Relationships: Further investigations into potential non-linear relationships between predictors and evaporation could lead to improved model accuracy. Exploring advanced modeling techniques may be beneficial.

Appendix

Bivariate Analysis

```
library(dplyr)     # For data manipulation
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)   # For data visualization
library(car)       # For Anova function
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
library(lubridate) # For handling date and time data
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```r
colnames(data)
```

```
## NULL
```

```r
# Load the dataset
data <- read.csv("melbourne_A4.csv")

# Check the structure of the dataset
str(data)
```

```
## 'data.frame':    270 obs. of  22 variables:
##  $ ...1                            : int  41 284 299 151 65 184 14 287 16 45 ...
##  $ Date                            : chr  "2019-02-20" "2018-12-8" "2018-12-30" "2019-06-29" ...
##  $ Minimum.temperature..Deg.C.     : num  13.5 22.7 18.2 13 18.7 10.6 16.5 16 18.4 13.8 ...
##  $ Maximum.Temperature..Deg.C.     : num  23.2 30.3 22.7 17.6 28.6 17.6 23.5 33.6 25.1 32.9 ...
##  $ Rainfall..mm.                   : num  0 0 0.6 0.8 0 0 2 0 0 0 ...
##  $ Evaporation..mm.                : num  6.2 14 2.2 4.2 2.8 2 4 7.4 6 5.2 ...
##  $ Sunshine..hours.                : num  9.1 1 0 1.1 6.4 8.4 13.2 9.8 10.7 12.4 ...
##  $ Direction.of.maximum.wind.gust  : chr  "S" "SSW" "SSW" "N" ...
##  $ Speed.of.maximum.wind.gust..km.h.: int  30 35 28 52 31 57 31 44 37 35 ...
##  $ Time.of.maximum.wind.gust       : chr  "15:14:00" "17:54:00" "16:10:00" "12:45:00" ...
##  $ X9am.Temperature..Deg.C.        : num  15.9 26.7 19.6 14.1 20.1 11.3 18.4 22 20 18.1 ...
##  $ X9am.relative.humidity....      : int  60 40 90 86 88 57 66 67 76 69 ...
##  $ X9am.cloud.amount..oktas.       : int  7 7 7 7 8 7 1 5 7 0 ...
##  $ X9am.wind.direction             : chr  "WNW" NA NA "NE" ...
##  $ X9am.wind.speed..km.h.          : chr  "4" "Calm" "Calm" "13" ...
##  $ X9am.MSL.pressure..hPa.         : num  1015 1015 1014 1015 1015 ...
##  $ X3pm.Temperature..Deg.C.        : num  22 29.6 20.9 13.9 25.5 17 22.9 32.7 24.8 31.8 ...
##  $ X3pm.relative.humidity....      : int  46 36 76 85 66 37 42 27 66 25 ...
##  $ X3pm.cloud.amount..oktas.       : int  5 7 8 7 2 1 1 2 1 0 ...
##  $ X3pm.wind.direction             : chr  "ENE" "NE" "SSW" "NNE" ...
##  $ X3pm.wind.speed..km.h.          : int  6 7 13 20 9 33 11 17 13 11 ...
##  $ X3pm.MSL.pressure..hPa.         : num  1014 1014 1013 1006 1012 ...
```

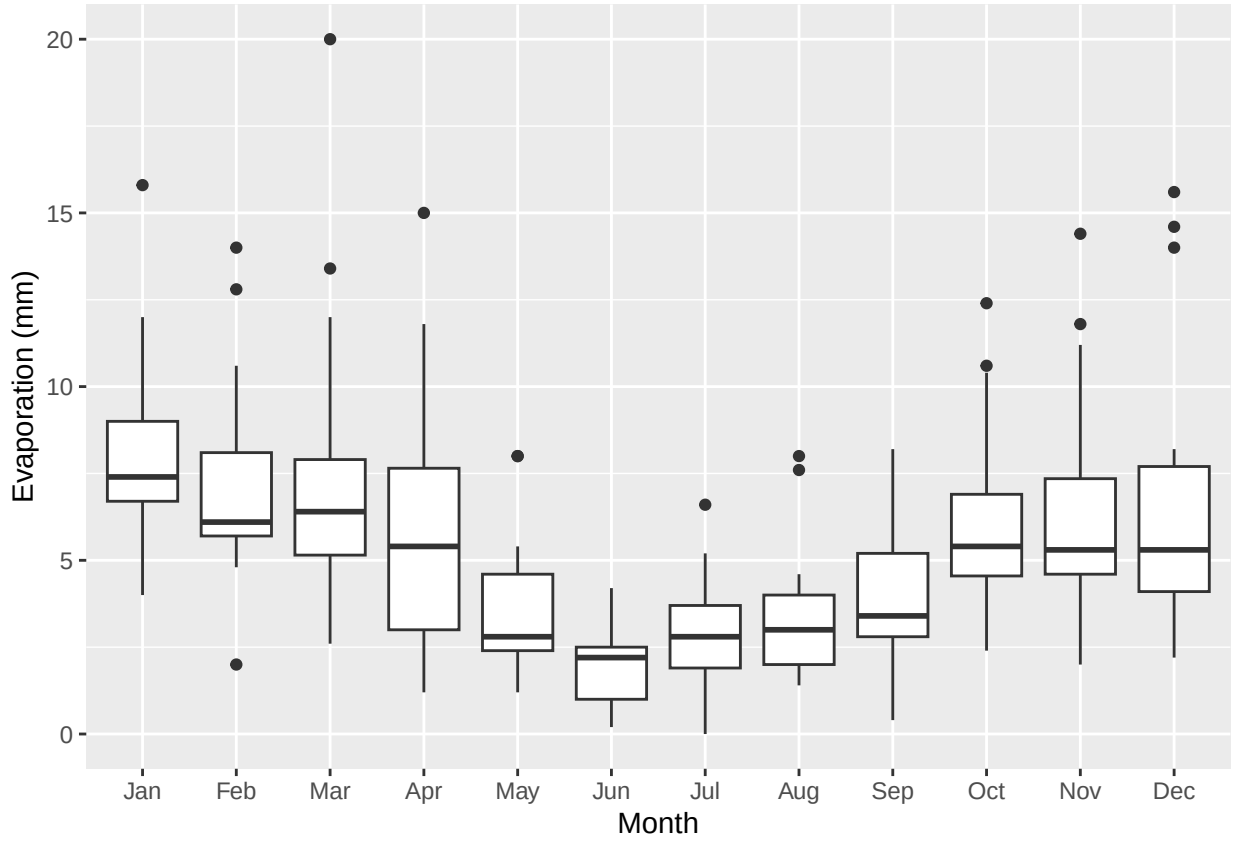
```r
# Convert Date column to a date format
data$Date <- as.Date(data$Date, format = "%Y-%m-%d")

# Extract month and day of the week from the Date column
data$Month <- month(data$Date, label = TRUE)
data$DayOfWeek <- weekdays(data$Date)

# Bivariate summaries
summary(data)
```

```
##       ...1             Date            Minimum.temperature..Deg.C.
##  Min.   :  1.00   Min.   :2018-07-01   Min.   : 1.90
##  1st Qu.: 77.25   1st Qu.:2018-10-01   1st Qu.: 8.60
##  Median :151.50   Median :2019-01-01   Median :11.30
##  Mean   :151.12   Mean   :2018-12-30   Mean   :11.71
##  3rd Qu.:225.75   3rd Qu.:2019-04-05   3rd Qu.:14.70
##  Max.   :300.00   Max.   :2019-06-29   Max.   :25.10
##
##  Maximum.Temperature..Deg.C. Rainfall..mm.    Evaporation..mm. Sunshine..hours.
##  Min.   : 9.60               Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:16.02               1st Qu.: 0.000   1st Qu.: 2.600   1st Qu.: 4.200
##  Median :19.50               Median : 0.000   Median : 4.800   Median : 7.600
##  Mean   :20.50               Mean   : 1.078   Mean   : 5.202   Mean   : 7.019
##  3rd Qu.:23.50               3rd Qu.: 0.600   3rd Qu.: 6.700   3rd Qu.: 9.975
##  Max.   :40.80               Max.   :35.800   Max.   :20.000   Max.   :13.900
##                              NA's   :2        NA's   :7
##  Direction.of.maximum.wind.gust Speed.of.maximum.wind.gust..km.h.
##  Length:270                     Min.   :13.0
##  Class :character               1st Qu.:28.5
##  Mode  :character               Median :35.0
##                                 Mean   :36.9
##                                 3rd Qu.:45.5
##                                 Max.   :80.0
##
##  Time.of.maximum.wind.gust X9am.Temperature..Deg.C. X9am.relative.humidity....
##  Length:270                Min.   : 4.10            Min.   : 30.00
##  Class :character          1st Qu.:11.22            1st Qu.: 61.00
##  Mode  :character          Median :14.30            Median : 68.00
##                            Mean   :14.68            Mean   : 68.36
##                            3rd Qu.:17.77            3rd Qu.: 77.00
##                            Max.   :30.80            Max.   :100.00
##
##  X9am.cloud.amount..oktas. X9am.wind.direction X9am.wind.speed..km.h.
##  Min.   :0.000             Length:270          Length:270
##  1st Qu.:2.000             Class :character    Class :character
##  Median :6.000             Mode  :character    Mode  :character
##  Mean   :4.819
##  3rd Qu.:7.000
##  Max.   :8.000
##
##  X9am.MSL.pressure..hPa. X3pm.Temperature..Deg.C. X3pm.relative.humidity....
##  Min.   : 993.2          Min.   : 9.00            Min.   : 18.00
##  1st Qu.:1012.6          1st Qu.:14.90            1st Qu.: 43.25
##  Median :1018.1          Median :18.60            Median : 53.00
##  Mean   :1017.8          Mean   :19.21            Mean   : 53.01
##  3rd Qu.:1023.5          3rd Qu.:22.20            3rd Qu.: 63.00
##  Max.   :1036.6          Max.   :38.40            Max.   :100.00
##
##  X3pm.cloud.amount..oktas. X3pm.wind.direction X3pm.wind.speed..km.h.
##  Min.   :0.000             Length:270          Min.   : 2.00
##  1st Qu.:2.000             Class :character    1st Qu.:11.00
##  Median :5.500             Mode  :character    Median :13.00
##  Mean   :4.607                                 Mean   :14.61
##  3rd Qu.:7.000                                 3rd Qu.:18.50
```

```
##  Max.   :8.000                                  Max.   :35.00
##
##  X3pm.MSL.pressure..hPa.      Month      DayOfWeek
##  Min.   : 997.2          Jul    : 28   Length:270
##  1st Qu.:1010.1          May    : 27   Class :character
##  Median :1016.0          Feb    : 24   Mode  :character
##  Mean   :1015.8          Oct    : 24
##  3rd Qu.:1021.1          Nov    : 24
##  Max.   :1032.8          Jun    : 23
##                          (Other):120
```

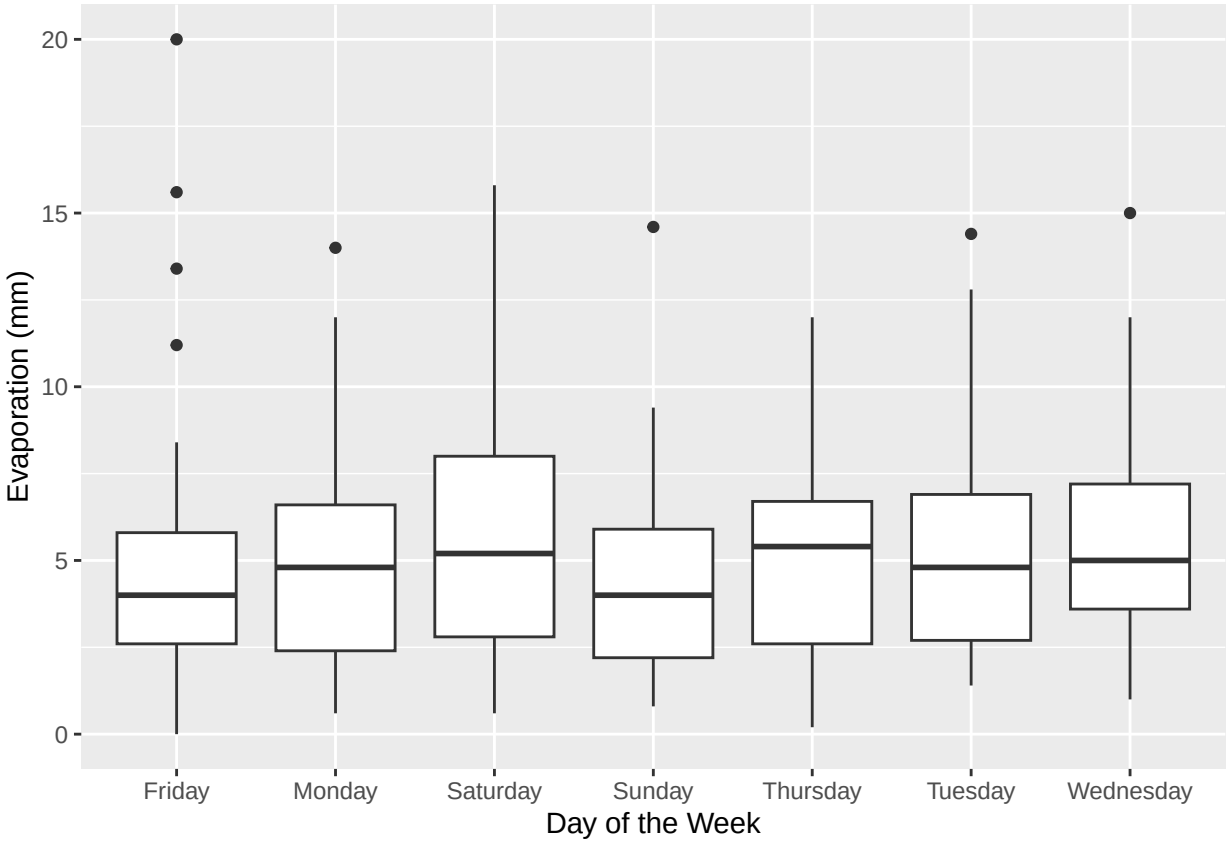
```
# Boxplot for evaporation by Month
ggplot(data, aes(x = Month, y = Evaporation..mm.)) +
  geom_boxplot() +
  labs(x = "Month", y = "Evaporation (mm)")
```

```
## Warning: Removed 7 rows containing non-finite values (‘stat_boxplot()‘).
```




```
# Boxplot for evaporation by Day of the Week
ggplot(data, aes(x = DayOfWeek, y = Evaporation..mm.)) +
  geom_boxplot() +
  labs(x = "Day of the Week", y = "Evaporation (mm)")
```

```
## Warning: Removed 7 rows containing non-finite values (‘stat_boxplot()‘).
```

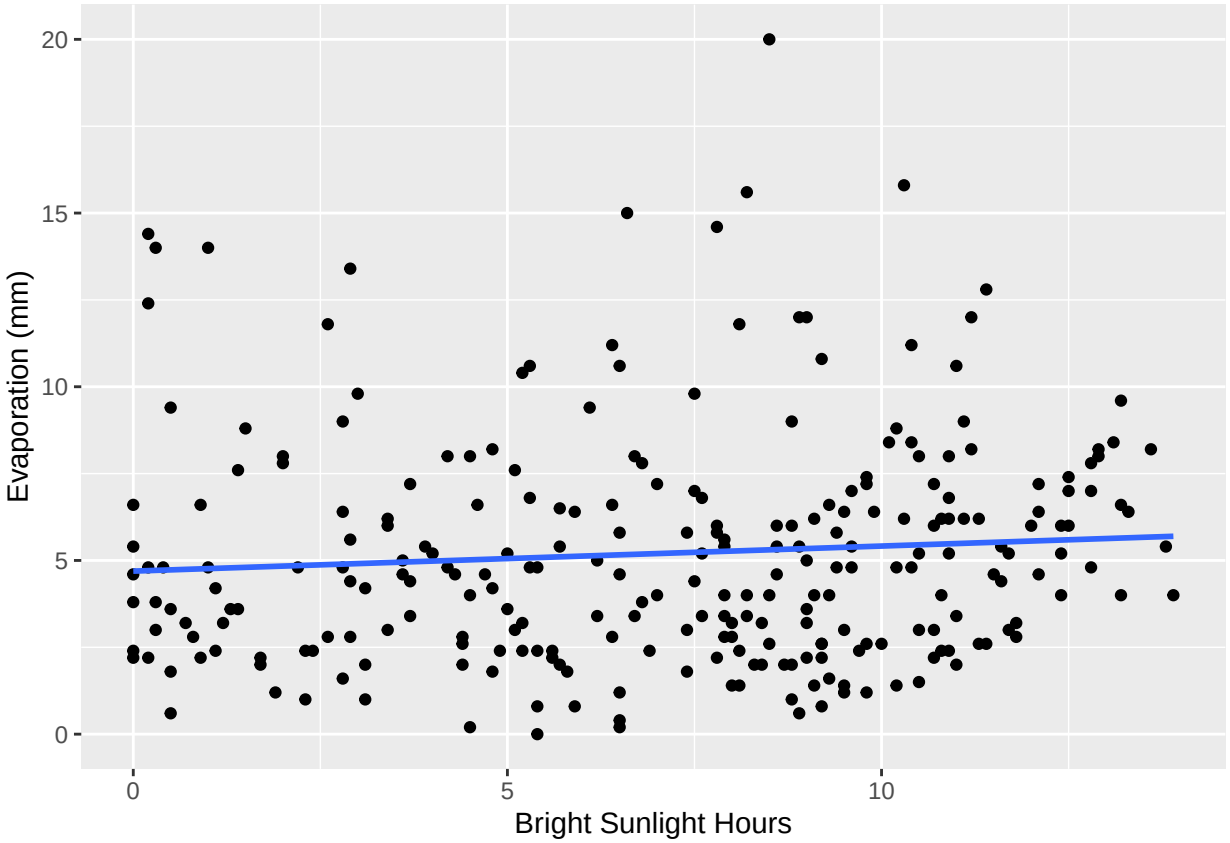



```
ggplot(data, aes(x = Sunshine..hours., y = Evaporation..mm.)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Bright Sunlight Hours", y = "Evaporation (mm)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 7 rows containing non-finite values (`stat_smooth()`).
```

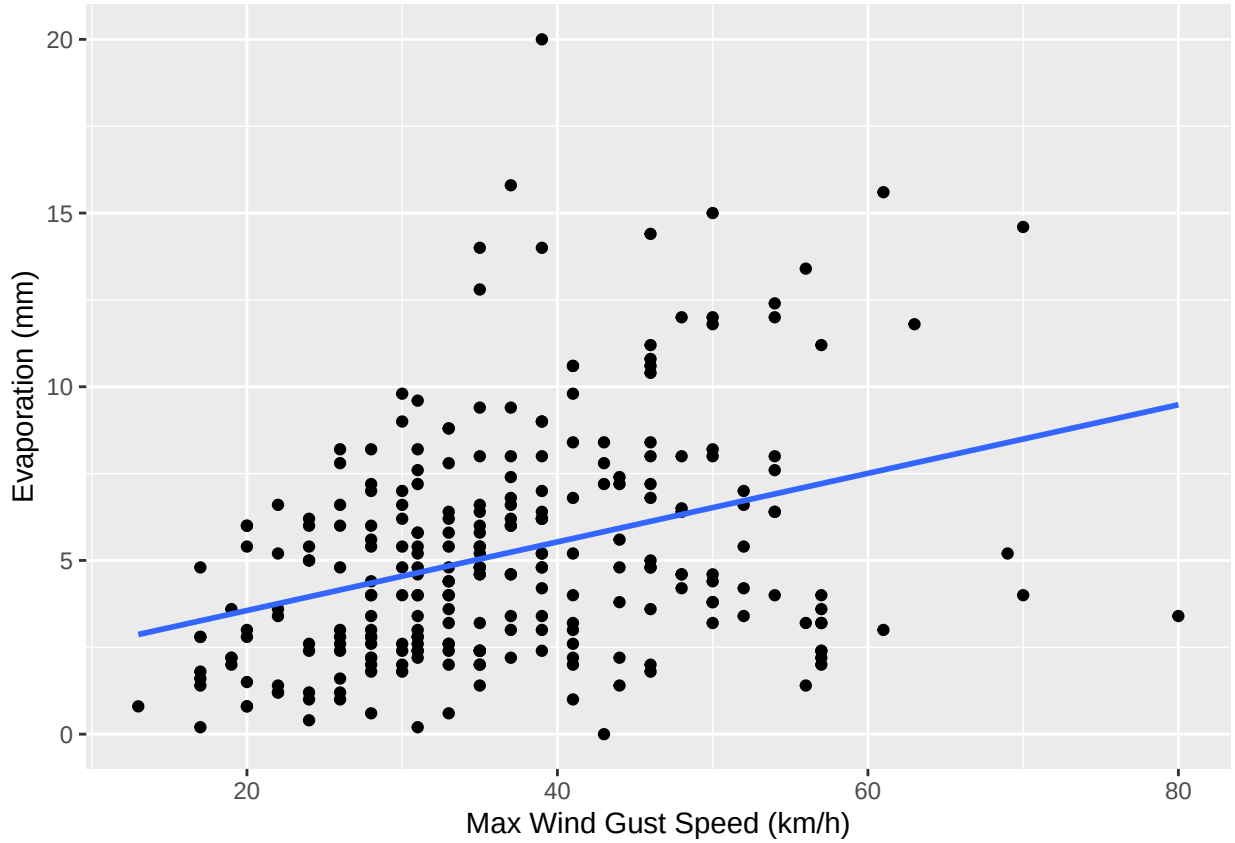
```
## Warning: Removed 7 rows containing missing values (`geom_point()`).
```




```
ggplot(data, aes(x = `Speed.of.maximum.wind.gust..km.h.`, y = `Evaporation..mm.`)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Max Wind Gust Speed (km/h)", y = "Evaporation (mm)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

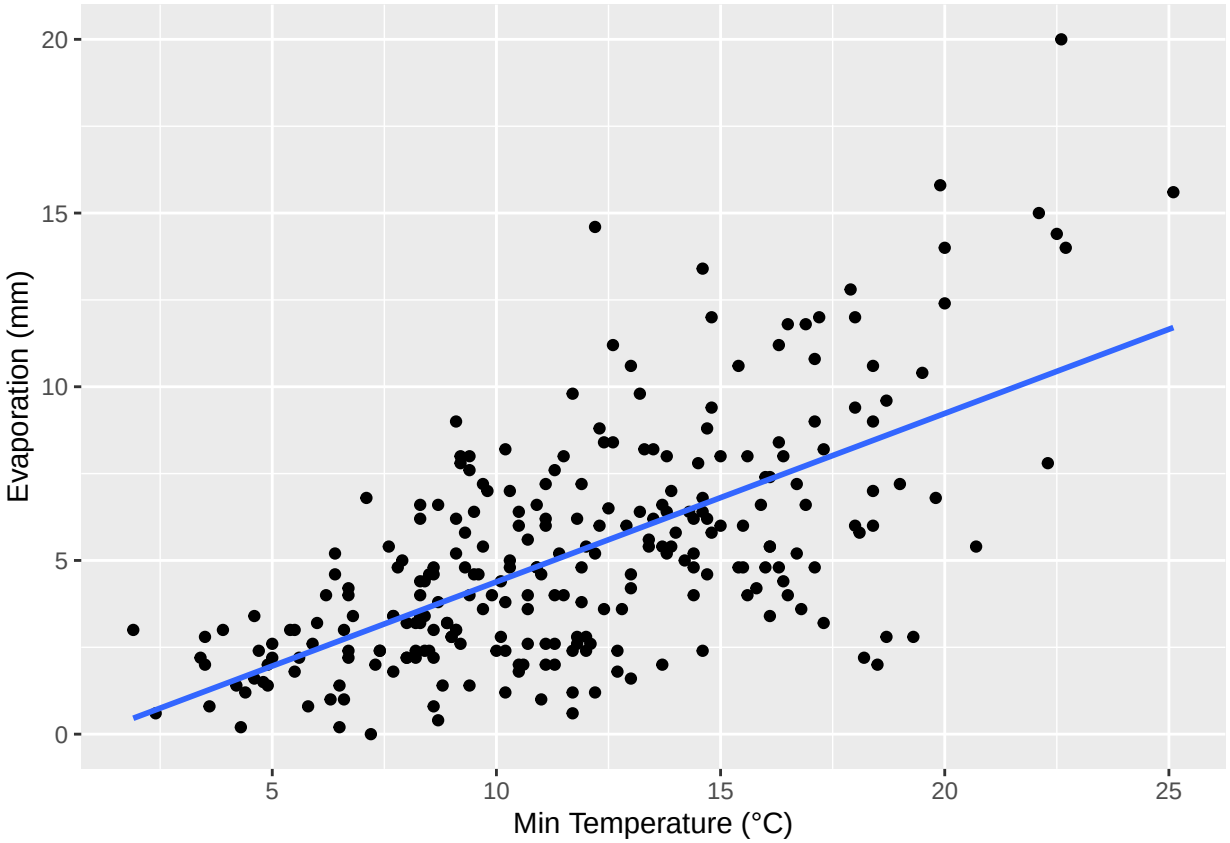
```
## Warning: Removed 7 rows containing non-finite values (`stat_smooth()`).
## Removed 7 rows containing missing values (`geom_point()`).
```




```
ggplot(data, aes(x = Minimum.temperature..Deg.C., y = Evaporation..mm.)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Min Temperature (°C)", y = "Evaporation (mm)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

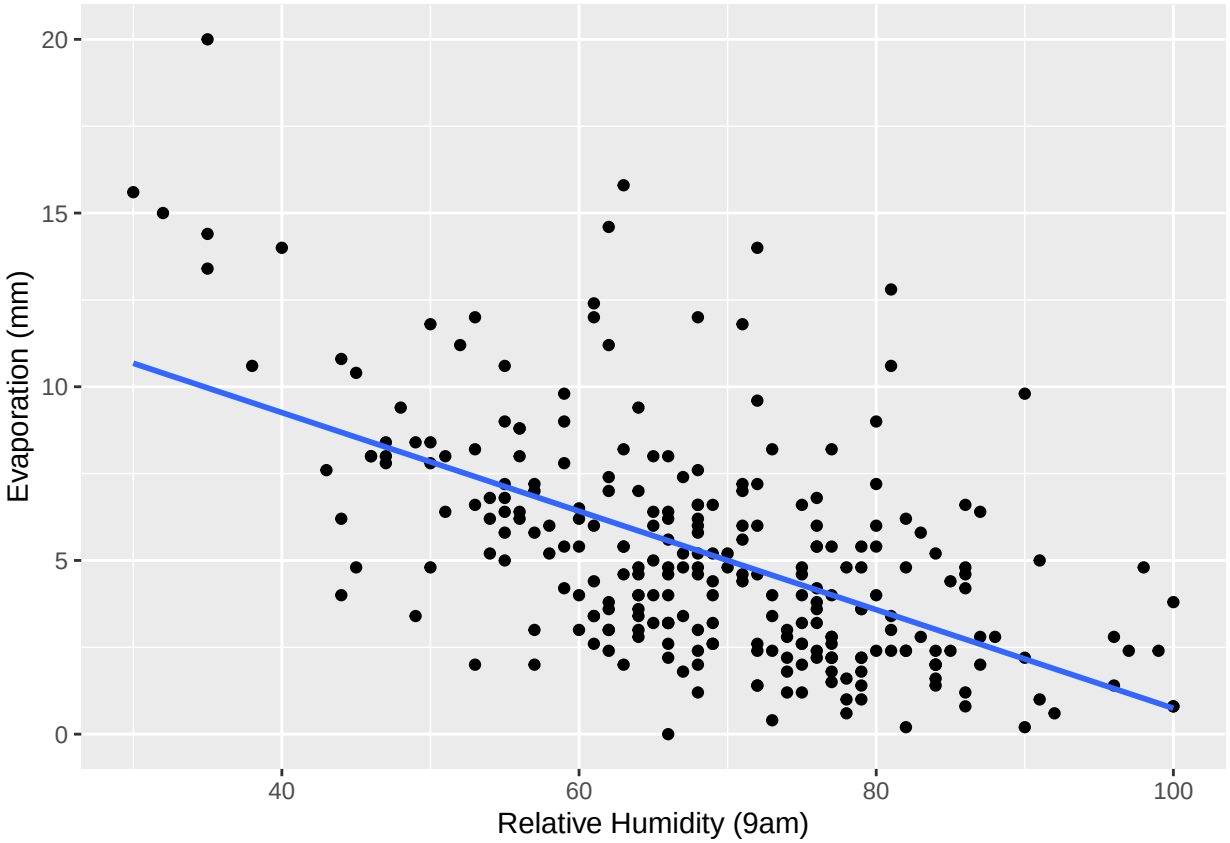
```
## Warning: Removed 7 rows containing non-finite values (`stat_smooth()`).
## Removed 7 rows containing missing values (`geom_point()`).
```




```
ggplot(data, aes(x = X9am.relative.humidity...., y = Evaporation..mm.)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Relative Humidity (9am)", y = "Evaporation (mm)")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 7 rows containing non-finite values (`stat_smooth()`).
## Removed 7 rows containing missing values (`geom_point()`).
```



Model Selection- ANOVA

```
data$Date <- as.Date(data$Date, format = "%Y-%m-%d")
data$Month <- month(data$Date, label = TRUE)
data$DayOfWeek <- weekdays(data$Date)


numeric_vars <- c("Sunshine..hours.", "Speed.of.maximum.wind.gust..km.h.", "Minimum.temperature..Deg.C.
for (var in numeric_vars) {
  data[is.na(data[[var]]), var] <- mean(data[[var]], na.rm = TRUE)
}


full_model <- lm(Evaporation..mm. ~ Month + DayOfWeek + Sunshine..hours. + Speed.of.maximum.wind.gust..


max_iterations <- 100
for (iteration in 1:max_iterations) {
  anova_result <- Anova(full_model, type = "III")
  max_pvalue <- max(anova_result$"Pr(>F)", na.rm = TRUE)
  if (!is.na(max_pvalue) && max_pvalue > 0.05) {
    non_significant_predictor <- rownames(anova_result)[which.max(anova_result$"Pr(>F)")]
    full_model <- update(full_model, . ~ . - eval(parse(text = non_significant_predictor)))
  } else {
    break
  }
}
```

```r
significant_predictors <- names(coef(full_model)[-1]) # Exclude intercept
cat("Significant predictors in the final model:", paste(significant_predictors, collapse = ", "), "\n")
```

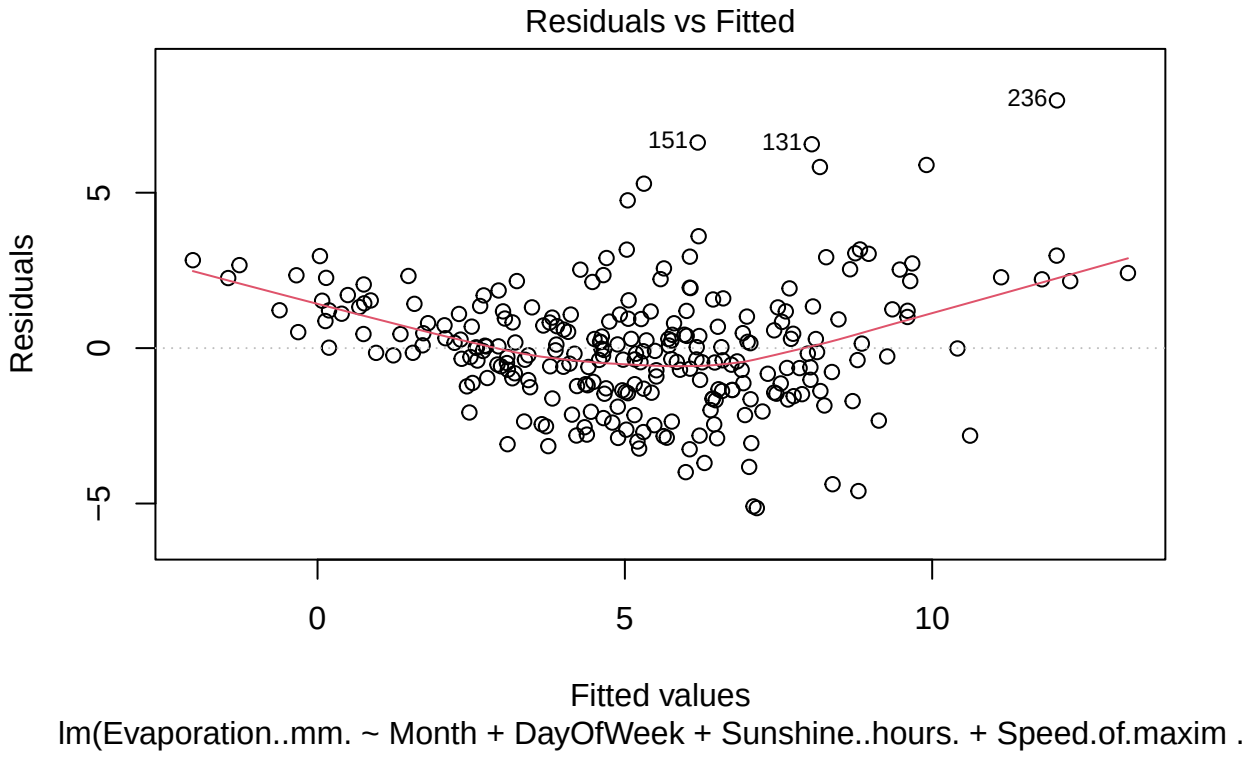
```
## Significant predictors in the final model: Month.L, Month.Q, Month.C, Month^4, Month^5, Month^6, Mon
```

```r
summary(full_model)
```

```
##
## Call:
## lm(formula = Evaporation..mm. ~ Month + DayOfWeek + Sunshine..hours. +
##     Speed.of.maximum.wind.gust..km.h. + Minimum.temperature..Deg.C. +
##     X9am.relative.humidity...., data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.147 -1.245 -0.060  1.090  7.970
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       6.41187    1.49146   4.299 2.49e-05 ***
## Month.L                          -1.01529    0.52959  -1.917 0.056404 .
## Month.Q                           2.66410    0.80104   3.326 0.001019 **
## Month.C                           0.09673    0.48517   0.199 0.842146
## Month^4                          -1.50899    0.46684  -3.232 0.001399 **
## Month^5                          -0.28446    0.44191  -0.644 0.520375
## Month^6                           0.54251    0.44337   1.224 0.222296
## Month^7                          -0.59993    0.43546  -1.378 0.169570
## Month^8                           0.67723    0.43883   1.543 0.124075
## Month^9                           0.20853    0.44712   0.466 0.641362
## Month^10                         -0.11797    0.44267  -0.266 0.790085
## Month^11                          0.23809    0.43330   0.549 0.583181
## DayOfWeekMonday                   0.42338    0.47499   0.891 0.373631
## DayOfWeekSaturday                 1.14277    0.48467   2.358 0.019181 *
## DayOfWeekSunday                   0.41357    0.46936   0.881 0.379120
## DayOfWeekThursday                 0.02653    0.48292   0.055 0.956229
## DayOfWeekTuesday                  0.37590    0.46173   0.814 0.416382
## DayOfWeekWednesday                0.19458    0.46055   0.423 0.673035
## Sunshine..hours.                 -0.03767    0.04311  -0.874 0.383146
## Speed.of.maximum.wind.gust..km.h.  0.05128    0.01337   3.835 0.000161 ***
## Minimum.temperature..Deg.C.       0.24897    0.05402   4.609 6.58e-06 ***
## X9am.relative.humidity....       -0.08820    0.01186  -7.437 1.81e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.051 on 241 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.6473, Adjusted R-squared:  0.6166
## F-statistic: 21.06 on 21 and 241 DF,  p-value: < 2.2e-16
```

Model Diagnostics

```
# Scatterplot of residuals vs. predicted values
plot(full_model, which = 1)
```




```
# durbin watson test for independence
library(lmtest)
```

```
## Loading required package: zoo
```

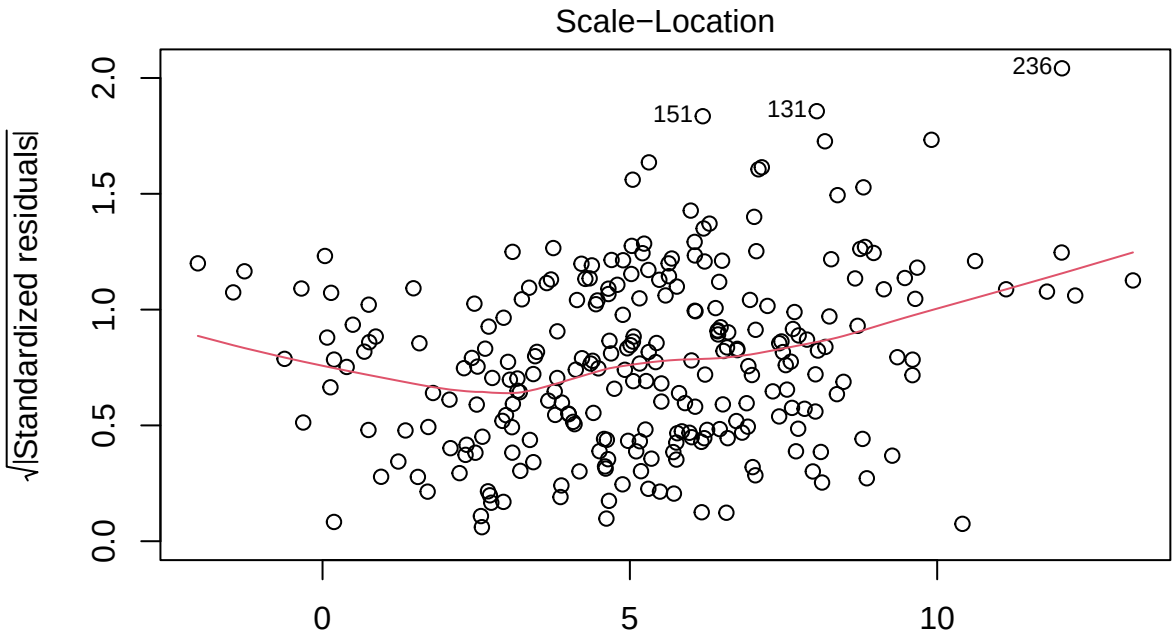
```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

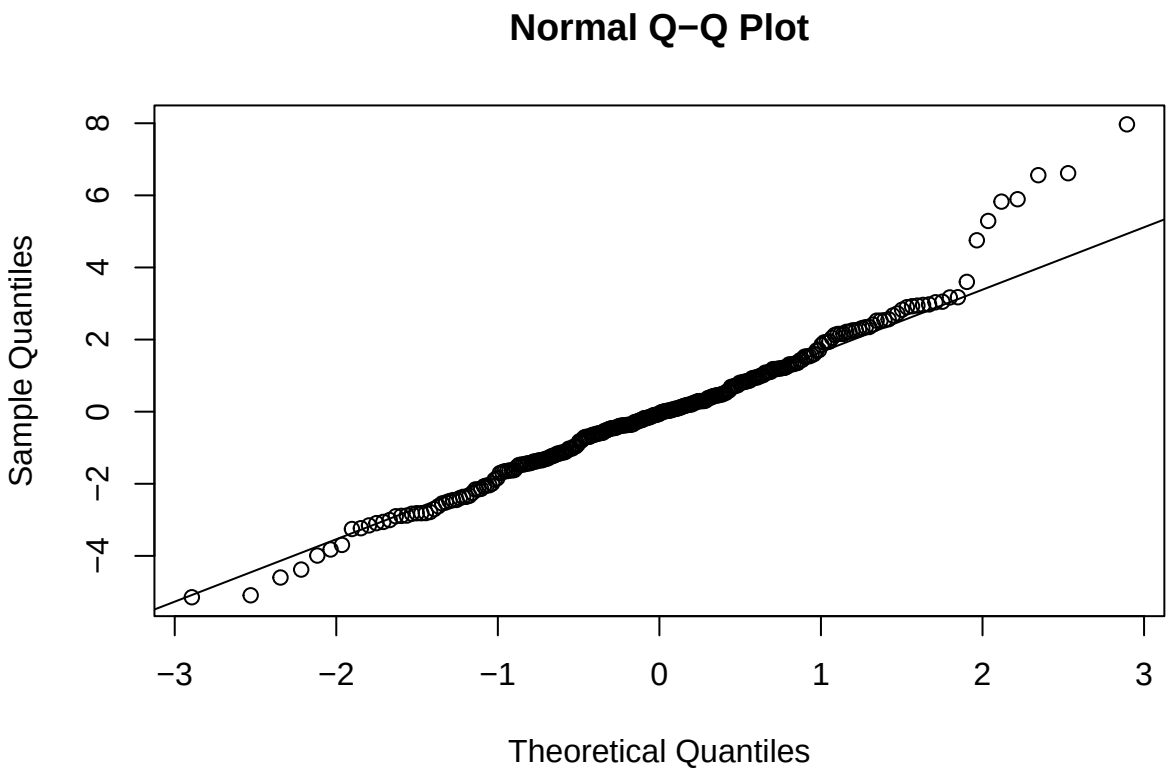
```
dwtest(full_model)
```

```
##
##  Durbin-Watson test
##
## data:  full_model
## DW = 2.0654, p-value = 0.7037
## alternative hypothesis: true autocorrelation is greater than 0
```

```r
# Scatterplot of residuals vs. predicted values
plot(full_model, which = 3)
```




```r
# QQ plot of residuals
qqnorm(residuals(full_model))
qqline(residuals(full_model))
```



Predictions

```
library(car)
library(lubridate)

data <- read.csv("melbourne_A4.csv")

data$Date <- as.Date(data$Date, format = "%Y-%m-%d")
data$Month <- month(data$Date, label = TRUE)
data$DayOfWeek <- weekdays(data$Date)

new_data <- data.frame(
  Month = factor(c("Feb", "Dec", "Jan", "Jul"), levels = levels(data$Month)),
  DayOfWeek = factor(rep("Saturday", 4), levels = c("Monday", "Tuesday", "Wednesday", "Thursday", "Frida
  Sunshine..hours. = c(3.4, 9, 9.6, 4),
  Speed.of.maximum.wind.gust..km.h. = c(23.2, 76, 44.3, 10.6),
  Minimum.temperature..Deg.C. = c(13.8, 16.4, 26.5, 6.8),
  X9am.relative.humidity.... = c(0.74, 0.37, 0.35, 0.76)
)

predictions <- predict(full_model, newdata = new_data, interval = "prediction", level = 0.95)

results_df <- data.frame(
  Date = data$Date[1:4],
```

```r
  Predicted_Evaporation = round(predictions[, 1], 3),
  Lower_Prediction_Limit = round(predictions[, 2], 3),
  Upper_Prediction_Limit = round(predictions[, 3], 3),
  Temporary_Measures_Needed = ifelse(round(predictions[, 1], 2) > 9, "Yes", "No")
)

print(results_df)
```

```
##         Date Predicted_Evaporation Lower_Prediction_Limit
## 1 2019-02-20                12.707                  8.126
## 2 2018-12-08                15.514                 11.054
## 3 2018-12-30                17.477                 12.996
## 4 2019-06-29                 8.280                  3.587
##   Upper_Prediction_Limit Temporary_Measures_Needed
## 1                 17.288                       Yes
## 2                 19.974                       Yes
## 3                 21.957                       Yes
## 4                 12.973                        No
```