Executive Summary

Specific temporal and meteorological factors directly impact the amount of evaporation (mm) in a given day. As a matter of interest, three numeric and two categorical factors were analysed to determine their potential influence on daily Melbourne evaporation values. Our analysis revealed that warmer conditions associated with higher minimum temperatures, seasonal influences related to different months and less saturated air associated with a lower relative humidity (%) creates an environment for higher evaporation rates. Therefore, the effect/relationship of Melbourne's day-to-day weather on evaporation (response variable) was modelled upon these three predictors (month, Minimum temperature, and 9am relative humidity (%)). According to our predictions, we have identified, with 95% confidence that one (January) out of the four extreme scenario's meeting specific characteristics will exceed the 10mm evaporation limit at MWC's Cardinia Reservoir.

Methodology

Having been engaged by the Melbourne Water Corporation to model the effect of Melbourne's day-to-day weather on evaporation, the crux of the data analysed is based upon the specific temporal and meteorological factors potentially influencing the amount of daily evaporation. Notably, this analysis was conducted in RStudio, as the prescribed IDE, using the R programming language which is renowned for statistical computing and graphics.

As a matter of interest, three quantitative and two categorical variables (transformed into factors) required analysis to determine their influence on daily evaporation.

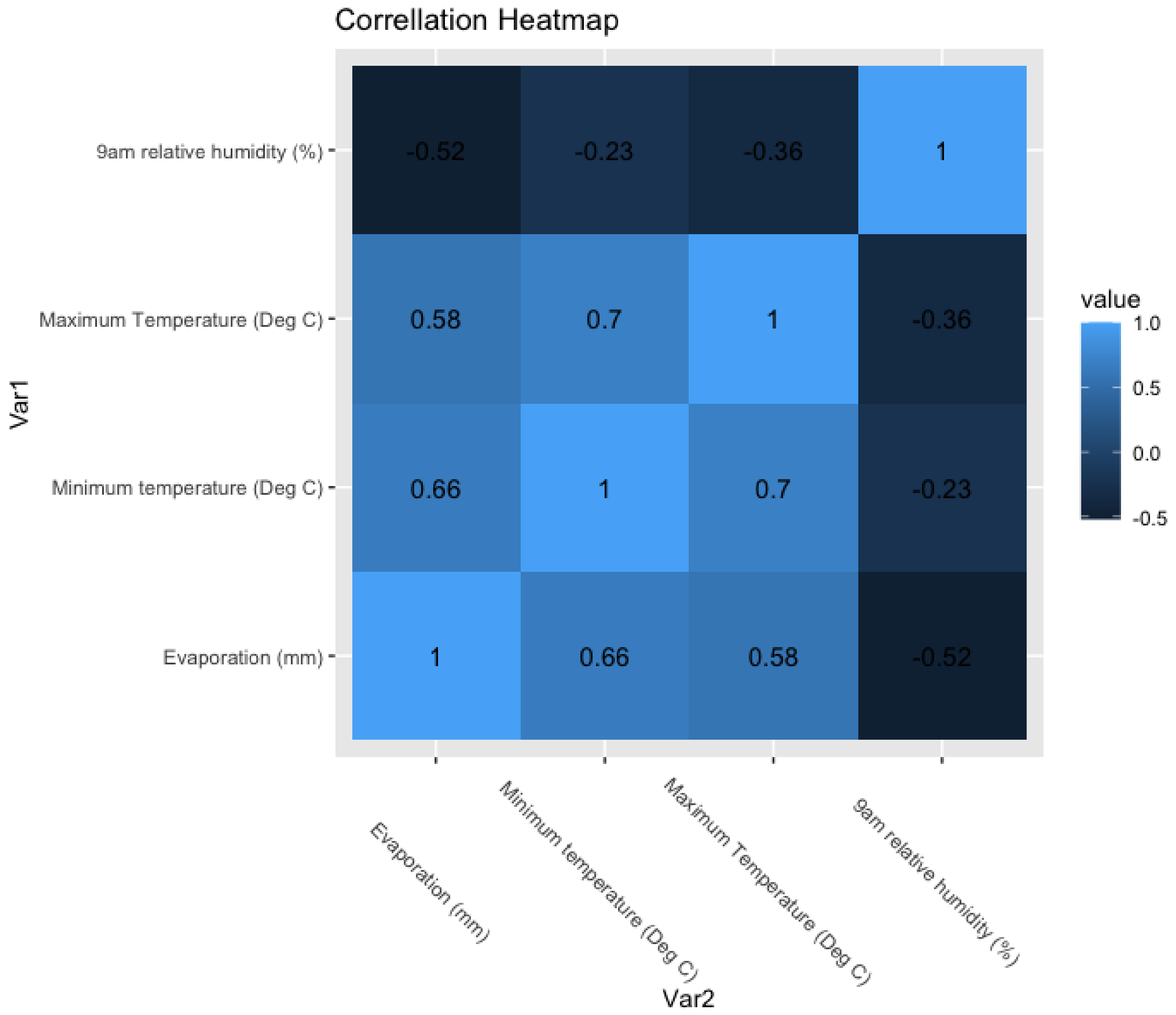
Firstly, conducting a univariate analysis upon our quantitative variables (Histogram) allows us to ascertain whether their distributions are fairly normally distributed or whether they need to undergo transformations. The shape, location, spread and outliers of our predictor variables are described within the appendix. Notably, Evaporation (highly right skewed) and Maximum temperature (moderate/high right skew) warrant transformations due to their skewness. Resultantly, square root transforming Evaporation (Appendix: Histogram 1) and log transforming Maximum temperature (Appendix: Histogram 4) significantly improves their normality.

V1	Variable	Skewness	Transformation	Skewness
1	Evaporation (mm)	1.33	Square Root	0.36
2	Minimum temperature (Deg C)	0.31	None	0.31
3	Maximum Temperature (Deg C)	0.94	Log	0.24

V1	Variable	Skewness	Transformation	Skewness
4	9am relative humidity (%)	-0.27	None	-0.27

Subsequent to our univariate analysis and distribution transformations, a bivariate analysis depicts the relationship of our five variables with daily evaporation. Notably, the linear relationship of quantitative vs quantitative variables requires scatterplots/correlation matrices while quantitative vs categorical variables require boxplots.

Relationship between Evaporation and predictor Vairables



Minimum temperature

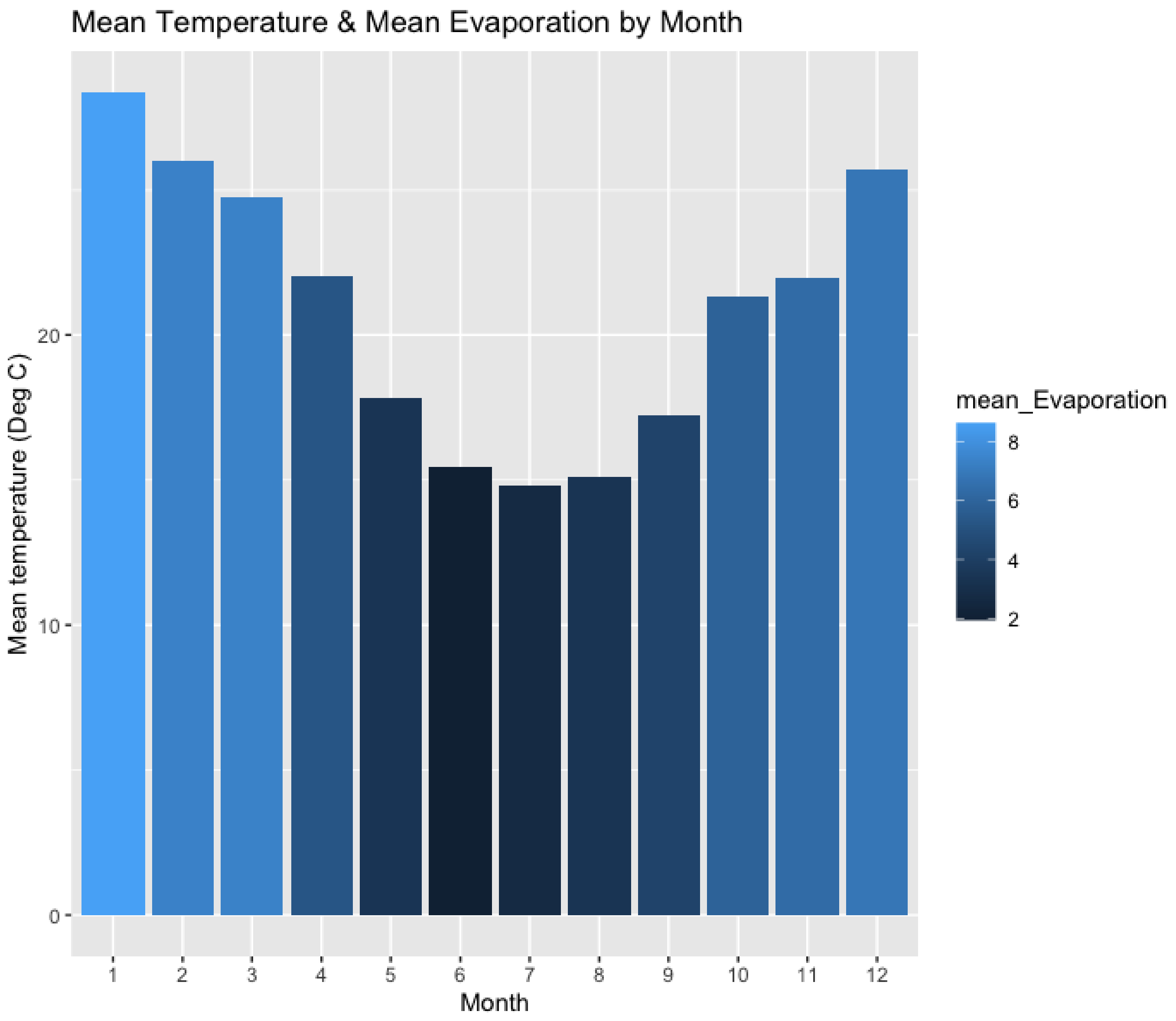
With a Pearson correlation of 0.66 we have sufficient evidence that Evaporation and Minimum temperature have a strong positive correlation (Correlation Heatmap). This relationship is evident from the clear upward trend in the scatterplot (Appendix: Scatterplot 1). Intuitively this makes sense since as temperature increases, the amount of energy necessary for evaporation decreases. Therefore, warmer conditions (increase in minimum temperatures) produce higher evaporation rates.

Max temperature

Evaporation and Maximum temperature have a strong positive correlation (Pearson correlation: 0.58). This relationship is evident from the clear upward trend in the scatterplot (Scatterplot 2). The intuition that warmer conditions produce higher evaporation rates is evidenced by this strong positive correlation and our scatterplot.

Month

Evaporation and months clearly indicate a quadratic relationship. We see a clear relationship between hotter months and higher evaporation (Boxplot 5). It is evident that winter months (cooler temperature) produce less evaporation indicated by the convexity of our plot. Intuitively this makes sense since warmer conditions produce higher evaporation rates as discussed above.



9am relative humidity

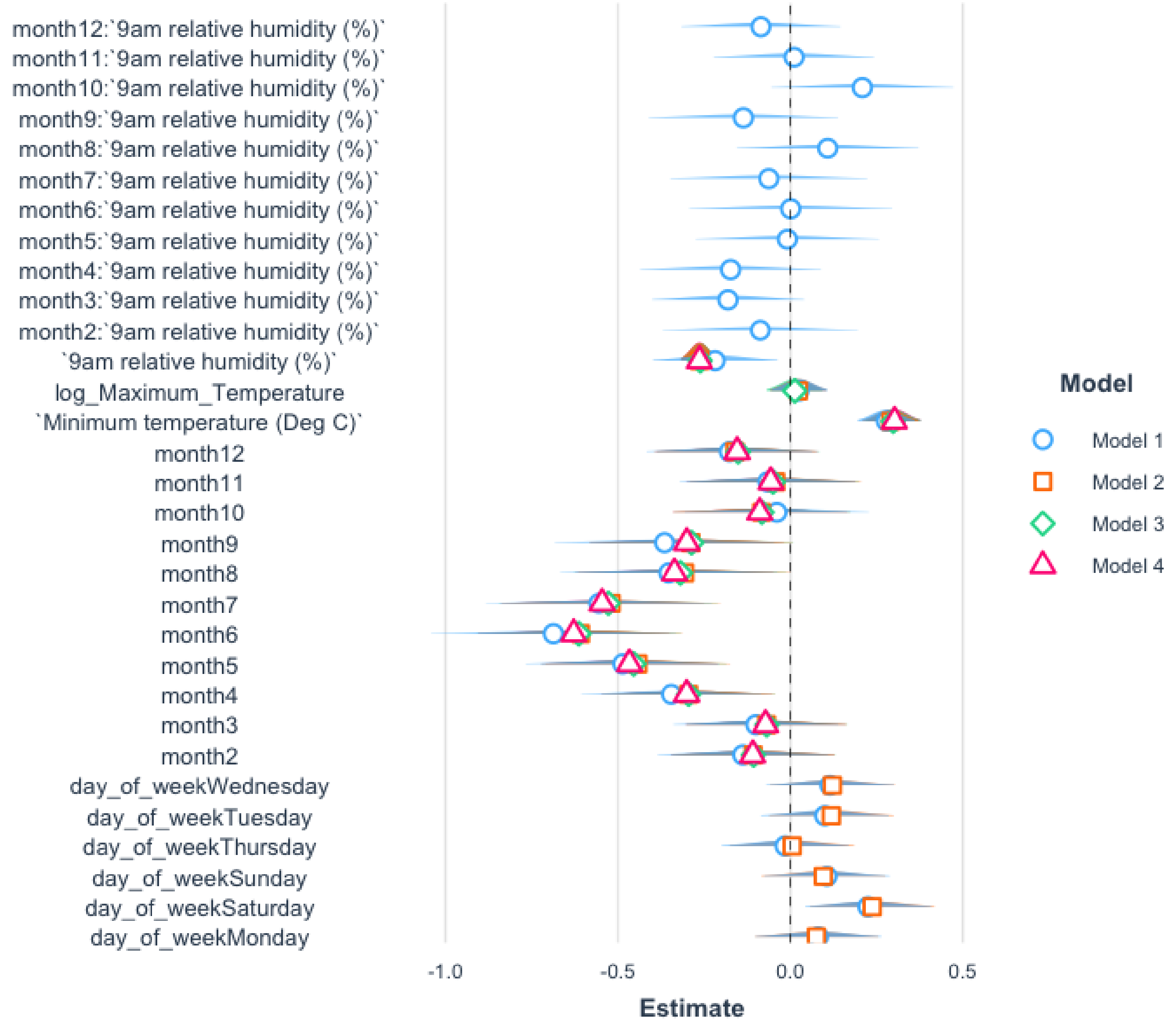
Evaporation and 9am relative humidity is strongly negatively correlated (Pearson correlation: -0.52). This relationship is evident from the clear downward trend in Scatterplot 3. This negative relationship is grounded upon the fact that the more humid the air, the closer the air is to saturation, and less evaporation can occur as reflected in our scatterplot.

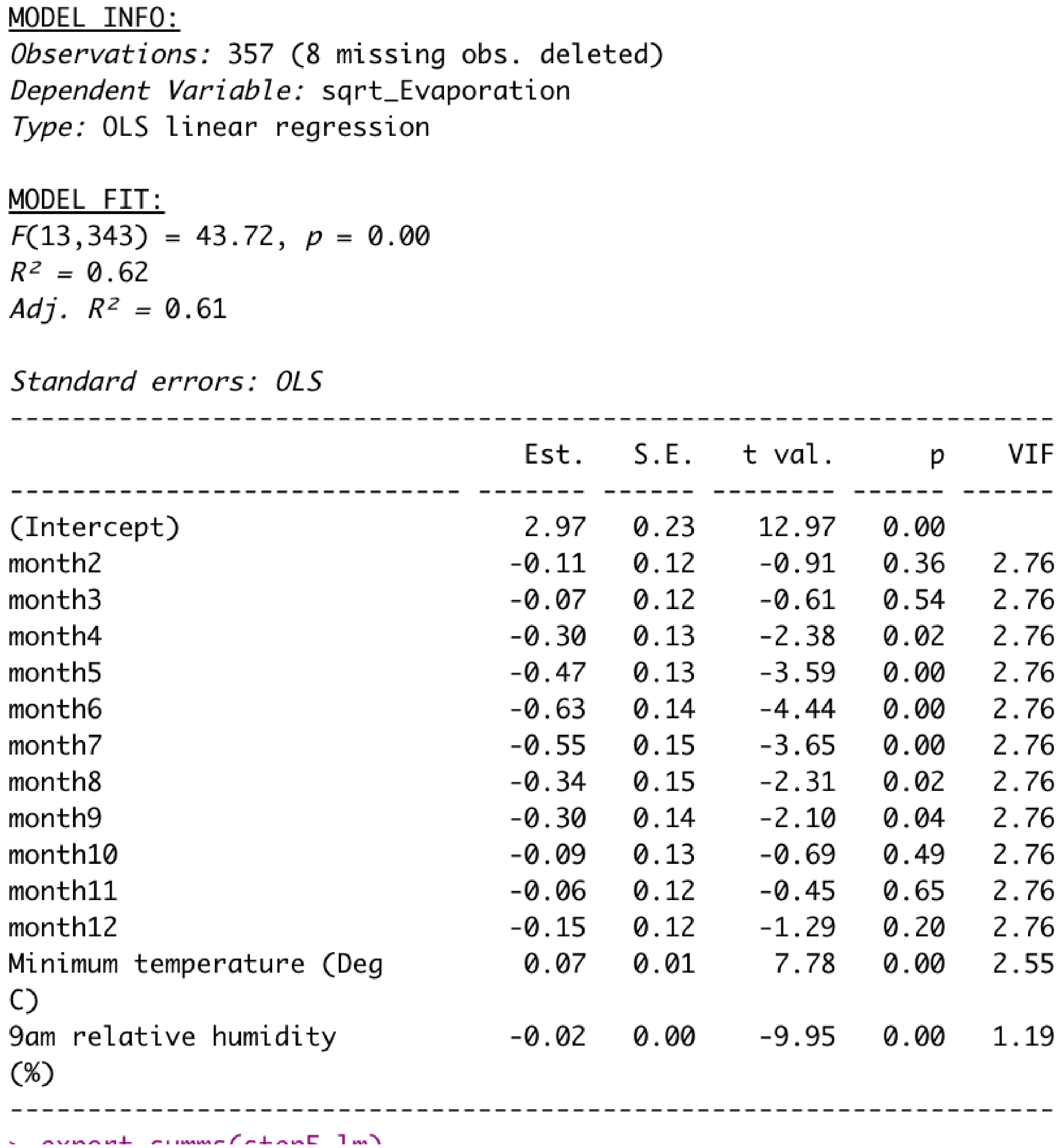
Day-of_Week

Evaporation and day-of-week have no relationship (i.e., there is no indication that any one day correlates with evaporation rates) (Boxplot 6). Intuitively we would expect evaporation rates to be influenced by seasonal conditions as opposed to daily conditions on average.

Significant terms in our final model

Following our bivariate analysis, four linear models were fitted and iteratively updated culminating in a final model that contained only the predictor variables that statistically influenced Evaporation.


month12:`9am relative humidity (%)`
month11:`9am relative humidity (%)`
month10:`9am relative humidity (%)`
month9:`9am relative humidity (%)`
month8:`9am relative humidity (%)`
month7:`9am relative humidity (%)`
month6:`9am relative humidity (%)`
month5:`9am relative humidity (%)`
month4:`9am relative humidity (%)`
month3:`9am relative humidity (%)`
month2:`9am relative humidity (%)`
`9am relative humidity (%)`
log_Maximum_Temperature
`Minimum temperature (Deg C)`
month12
month11
month10
month9
month8
month7
month6
month5
month4
month3
month2
day_of_weekWednesday
day_of_weekTuesday
day_of_weekThursday
day_of_weekSunday
day_of_weekSaturday
day_of_weekMonday
-1.0
-0.5
0.0
0.5
Estimate
Model
○ Model 1
□ Model 2
◇ Model 3
△ Model 4



```
MODEL INFO:
Observations: 357 (8 missing obs. deleted)
Dependent Variable: sqrt_Evaporation
Type: OLS linear regression

MODEL FIT:
F(13,343) = 43.72, p = 0.00
R² = 0.62
Adj. R² = 0.61

Standard errors: OLS
-----------------------------------------------------------------
                               Est.   S.E.   t val.      p    VIF
----------------------------- ------- ------ -------- ------ ------
(Intercept)                    2.97   0.23    12.97   0.00
month2                        -0.11   0.12    -0.91   0.36   2.76
month3                        -0.07   0.12    -0.61   0.54   2.76
month4                        -0.30   0.13    -2.38   0.02   2.76
month5                        -0.47   0.13    -3.59   0.00   2.76
month6                        -0.63   0.14    -4.44   0.00   2.76
month7                        -0.55   0.15    -3.65   0.00   2.76
month8                        -0.34   0.15    -2.31   0.02   2.76
month9                        -0.30   0.14    -2.10   0.04   2.76
month10                       -0.09   0.13    -0.69   0.49   2.76
month11                       -0.06   0.12    -0.45   0.65   2.76
month12                       -0.15   0.12    -1.29   0.20   2.76
Minimum temperature (Deg       0.07   0.01     7.78   0.00   2.55
C)
9am relative humidity         -0.02   0.00    -9.95   0.00   1.19
(%)
-----------------------------------------------------------------
> export_summs(stepF_lm)
```

```
##                                    V1  Df   Sum Sq    Mean Sq  F value       Pr(>F)
## 1                               month  11 77.71297  7.0648152 33.59465 4.363433e-48
## 2 `Minimum temperature (Deg C)`   1 21.00650 21.0065042 99.89025 8.247075e-21
## 3    `9am relative humidity (%)`   1 20.80217 20.8021692 98.91859 1.206394e-20
## 4                           Residuals 343 72.13148  0.2102958       NA           NA
```

Our final model has three statistically significant terms at the 0.1% (0.001) significance level, two of which are quantitative, Minimum temperature (8.77e-14) & 9am relative humidity (2e-16), and one categorical term, Month (2.2e-16).

Our bivariate analysis indicated days-of-week had no linear relationship with evaporation, and this was confirmed through our anova test with a p-value greater than 0.05 (not statistically significant).

Our bivariate analysis illustrated minimum temperature and Evaporation are strongly positively correlated. Additionally, the intuition that higher minimum temperature produce higher rates of evaporation has been statistically confirmed through our model’s summary statistics (p-value of 8.77e-14) (statistically significant at the significance level of 0.1%).

Our bivariate analysis illustrated 9am relative humidity and Evaporation are strongly negatively correlated. Drier air is more conducive to higher evaporation rates and this relationship has been confirmed through our model’s summary statistics (p-value of 2e-16) (significant at the significance level of 0.1%).

Our bivariate analysis illustrated a clear quadratic trend between months and Evaporation (Boxplot 5). Intuitively we expected colder months to produce lower evaporation rates i.e., seasonal changes have a bearing on evaporation rates. This relationship has been confirmed through our model’s anova test with a p-value of 2.2e-16.

Despite maximum temperature and Evaporation being highly positively correlated, this relationship was rejected through our model’s summary statistics (p-value > 0.05). Notably, correlations can simply appear because of a trend and not necessarily through a cause-and-effect relationship. This intuitively seems unlikely. Notably, signs of multicollinearity between minimum and maximum temperature (Heatmap) potentially resulted in our model placing more weight on minimum temperature (VIF in model table).

Model Diagnostics

A more robust discussion on Model diagnostics and Autocorellation appears in the appendix. Despite the slight convexity, the residual vs Fitted plot depicts a roughly straight red line about zero on the y-axis, depicting the presence of a **linear relationship**; the bulk proportion of our observations lie on the line with minimal deviations illustrating the **residuals are normally distributed**; putting more weight towards the ncvTest as opposed to the slight increase in the red line the assumption of **homoscedasticity** (constant spread) is justified.; finally, information about the sampling process is not provided so it is difficult to decide **independence** as we do not know how the data was obtained.

Results

Intercept Term

An astute observer would notice that month1 (January) is missing from the list of predictors in our model summary. Notably, month1 (January) is our reference category of comparison with the other categories (month2-month12) being compared to this reference month. Therefore, the coefficient for the intercept term is actually the coefficient of the reference category of the month1 predictor. The intercept provides us with the value of our response variable when our predictor variables are all held constant. Therefore, if all our predictor terms were to be held constant, we would expect evaporation (mm) to be 8.8477mm on average (2.974509mm on the square root scale).

Categorical predictor

In terms of our categorical predictor, if all other terms were held constant, Evaporation in the month of June (month6) would result in a decrease in evaporation by -0.3948815mm (-0.628396mm on the square root scale) as compared to our January (reference category). Since our analysis illustrated January contains the greatest quantity of evaporation (hottest month), in comparison to this reference category we would expect other months to produce reduced amounts of evaporation on average.

9am relative humidity

An increase in 9am relative humidity by one unit (1% in this context), results in evaporation (mm) decreasing on average by -0.000377mm (-0.0194mm on the square root scale). Therefore, holding all other terms constant, the difference between a day having 2% and 3% relative humidity at 9am is -0.000377mm of evaporation on average. This makes sense since the lower the relative humidity, the drier the air, and the higher the evaporation rate.

Minimum temperature

An increase in Minimum temperature by one unit (1 degree), results in evaporation (mm) increasing on average by 0.00446mm (0.0667mm on the square root scale). Therefore, holding all other terms constant, the difference between a day having a minimum temperature of 2 degrees and 3 degrees is 0.00446mm of evaporation on average. This makes sense since warmer conditions (as min temp increases) produce higher evaporation rates.

Discussion

Prediction intervals contain more uncertainty than confidence intervals. As seen in the tables, the range of prediction intervals are wider because they provide a range for a single observation that meets the extreme scenarios, while the range of confidence intervals provide a likely range of values over the whole dataset. Notably, the amount of evaporation predicted for our specific data lies between the lower and upper bounds in our table. For example, the amount of evaporation predicted for July lies between 1.50mm to 2.52mm with there being a 5% probability that the amount of evaporation will not be fall between these values. We can see that the higher the predicted rate of evaporation the greater the range of these prediction interval with a mean range of 9.83mm. The same pattern is found in our confidence interval predictions but with a significantly smaller mean range of 2.1mm. With these wider ranges three of our days (February, December, and January) have a evaporation prediction interval upper bound greater than 10mm with July falling safely under that requirement in terms of both confidence and prediction intervals. In contrast, the smaller ranges of our confidence intervals result in only January having a complete interval in excess of our 10mm requirement. Therefore, based on our prediction intervals we can say with 95% confidence that January, February and December will exceed the 10mm evaporation limit at MWC’s Cardinia Reservoir while July will not. However, based on our confidence intervals only January will exceed 10mm with this not occurring for July, February and December.

Confidence Interval

V1	Month	Min Temperature (Deg C)	9am relative Humidity (%)	lower	fit	upper
1	February	13.8	74	4.722296	5.518040	6.375713
2	December	16.4	57	6.935714	7.883199	8.891326
3	January	26.5	35	14.661481	16.513793	18.476260
4	July	6.8	76	1.499938	1.976880	2.519559

Prediction Interval

V1	Month	Min Temperature (Deg C)	9am relative Humidity (%)	lower	fit	upper
1	February	13.8	74	2.0450807	5.518040	10.680076
2	December	16.4	57	3.5685697	7.883199	13.885618
3	January	26.5	35	9.8075572	16.513793	24.957332
4	July	6.8	76	0.2361888	1.976880	5.410456

It comes as no surprise that the warmest average month (January) with the highest minimum temperature and lowest relative humidity (%) resulted in our highest predicted evaporation rate of 16.5mm. In contrast, our coldest average month (July) with the lowest minimum temperature and highest relative humidity unsurprisingly resulted in our lowest predicted evaporation rate of 1.98mm. Notably, February and December both exhibit warmer conditions and mean evaporation rates. Additionally, our February and December predictions differ by only 2.6 degrees in terms of their minimum temperature which is substantially less than January conditions and substantially more than our June conditions. However, February possesses a relatively high 9am humidity (%) compared to December and January as is more akin to July humidity. Therefore, we see a clear divergence between the two predicted values of February (5.52mm) and December (7.88mm).

Conslusion

With the primary focus of our analysis pertaining the impact of specific temporal and meteorological factors on Melbourne daily evaporation rates we have conclusively indicated that warmer conditions associated with higher minimum temperature, seasonal influences related to different months and less saturated air associated with a lower relative humidity (%) creates an environment for higher evaporation rates. Provided our linear assumptions hold true, we can say with 95% confidence that our predicted value for January will exceed the 10mm evaporation limit at MWC's Cardinia Reservoir and requires temporary measures to ensure a continuous supply of water.

Appendix

load packages

```
pacman::p_load(pacman, tidyverse, dplyr, stringr, ggplot2, rmarkdown,modelr, inspectd
f,moments ,rio, lubridate, knitr, gridExtra, ggcorrplot, reshape2,olsrr,lmtest,car,im
ager,EBImage)
```

```
##
## The downloaded binary packages are in
##  /var/folders/fl/rgw952ps3z11_sd2xjpg36p80000gn/T//RtmpXAlVDF/downloaded_packages
```

```
##
## imager installed
```

```
## Warning in pacman::p_load(pacman, tidyverse, dplyr, stringr, ggplot2, rmarkdown, :
Failed to install/load:
## imager
```

Import csv.file

```
df <- rio::import('/Users/garethbayvel/Desktop/melbourne.csv',setclass  = 'tibble')
```

Isolate day of week from yyyy-mm-dd format and set as factor

```
df$day_of_week <- weekdays(df$Date)%>%
  as.factor()
```

Isolate month from yyyy-mm-dd format and set as factor

```
df$month <- month(df$Date) %>%
  as.factor()
```

View output ITO our bivariate analysis columns

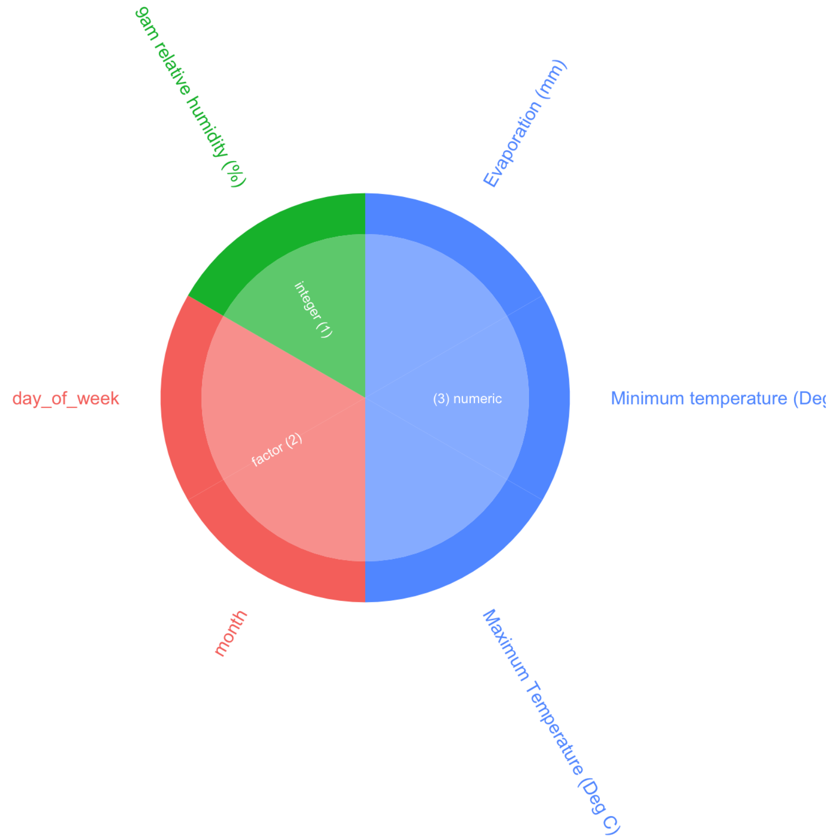
```
df <- df%>%
  select('Evaporation (mm)', month, day_of_week,`Minimum temperature (Deg C)`,
         `Maximum Temperature (Deg C)`,`9am relative humidity (%)`,everything())

head(df)
```

```
## # A tibble: 6 x 23
##   `Evaporation (mm… month day_of_week `Minimum temperature… `Maximum Temperatur…
##               <dbl> <fct> <fct>                      <dbl>                <dbl>
## 1               7   1     Tuesday                     15.5                 26.2
## 2               7   1     Wednesday                   18.4                 22.2
## 3               6.6 1     Thursday                    15.9                 29.5
## 4               7.8 1     Friday                      18                   42.6
## 5              15.4 1     Saturday                    17.4                 21.2
## 6               6.4 1     Sunday                      14.6                 22.1
## # … with 18 more variables: 9am relative humidity (%) <int>, Date <date>,
## #   Rainfall (mm) <dbl>, Sunshine (hours) <dbl>,
## #   Direction of maximum wind gust <chr>,
## #   Speed of maximum wind gust (km/h) <int>, Time of maximum wind gust <chr>,
## #   9am Temperature (Deg C) <dbl>, 9am cloud amount (oktas) <int>,
## #   9am wind direction <chr>, 9am wind speed (km/h) <chr>,
## #   9am MSL pressure (hPa) <dbl>, 3pm Temperature (Deg C) <dbl>,
## #   3pm relative humidity (%) <int>, 3pm cloud amount (oktas) <int>,
## #   3pm wind direction <chr>, 3pm wind speed (km/h) <int>,
## #   3pm MSL pressure (hPa) <dbl>
```

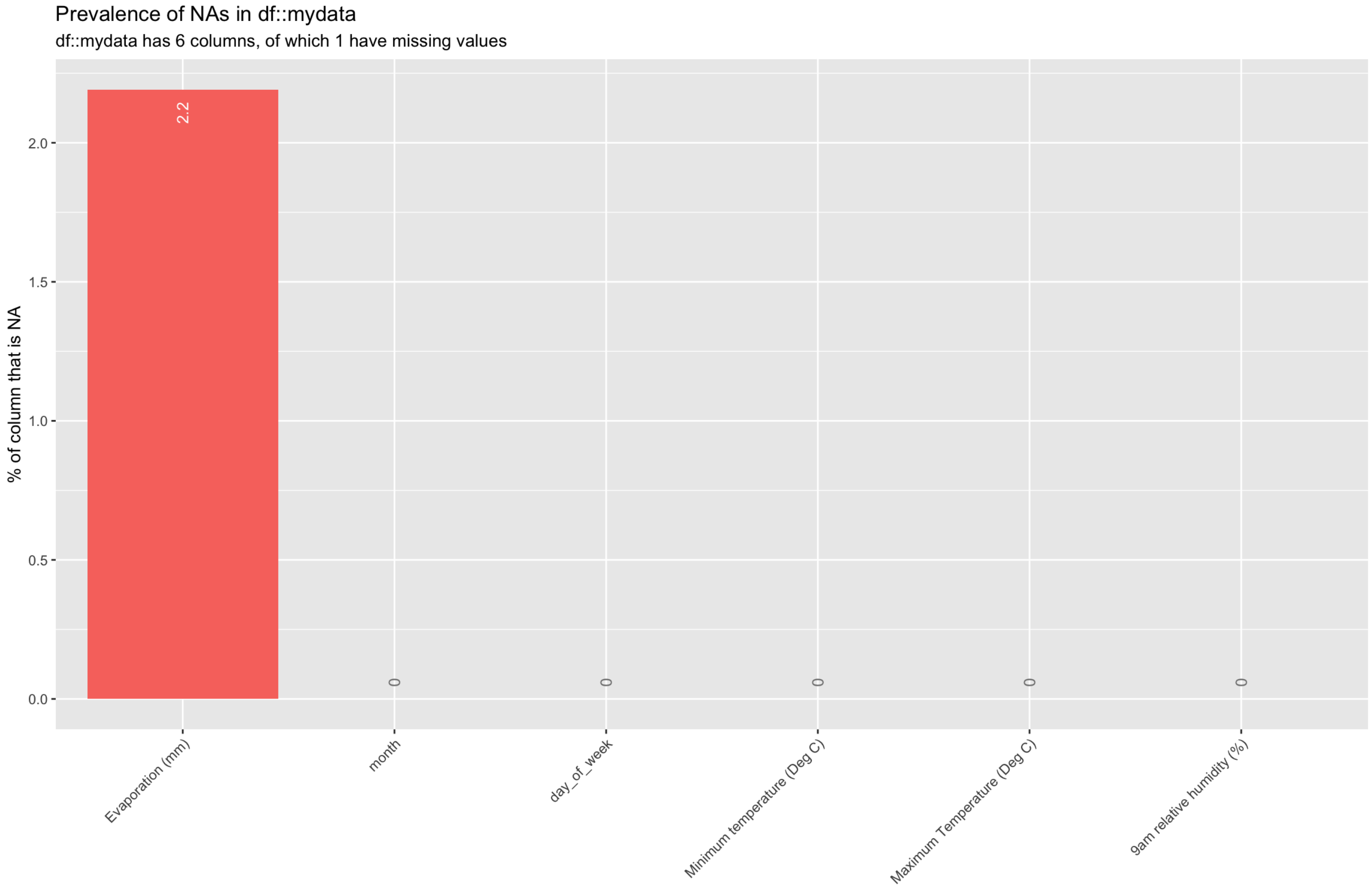
Data Types

```
mydata <- df[, c(1,2,3,4,5,6)]
inspect_types(mydata)%>%
  show_plot()
```



Missing Values

```
inspect_na(mydata)%>%
  show_plot()
```



Transforming Evaporation

```
df <- df%>%
  mutate(sqrt_Evaporation = sqrt(`Evaporation (mm)`))
```

Transforming Maximum Temp

```
df <- df%>%
  mutate(log_Maximum_Temperature = log(`Maximum Temperature (Deg C)`))
```

Table 1: Skewness of quanitative variables

```
skewness_table <- tribble(
  ~Variable,~Skewness, ~ Transformation, ~Skewness,
  'Evaporation (mm)',round(skewness(df$`Evaporation (mm)`,na.rm = T),digits = 2),'Squ
are Root',round(skewness(df$sqrt_Evaporation,na.rm = T),digits = 2),
  'Minimum temperature (Deg C)', round(skewness(df$`Minimum temperature (Deg C)`,na.r
m = T),digits = 2), 'None',round(skewness(df$`Minimum temperature (Deg C)`,na.rm = T
),digits = 2),
  'Maximum Temperature (Deg C)', round(skewness(df$`Maximum Temperature (Deg C)`,na.r
m = T),digits = 2),"Log",round(skewness(df$log_Maximum_Temperature,na.rm = T),digits
 = 2),
  '9am relative humidity (%)', round(skewness(df$`9am relative humidity (%)`,na.rm =
T),digits = 2), 'None', round(skewness(df$`9am relative humidity (%)`,na.rm = T),digi
ts = 2)
)
kable(skewness_table)
```

Variable	Skewness	Transformation	Skewness
Evaporation (mm)	1.33	Square Root	0.36
Minimum temperature (Deg C)	0.31	None	0.31
Maximum Temperature (Deg C)	0.94	Log	0.24
9am relative humidity (%)	-0.27	None	-0.27

The skewness table indicates Evaporation is highly right skewed (skewness statistic >1)

The skewness table indicates Minimum temperature is fairly symmetric (skewness statistic of 0.31)

The skewness table indicates 9am relative humidity is fairly symmetric (skewness statistic of -0.27)

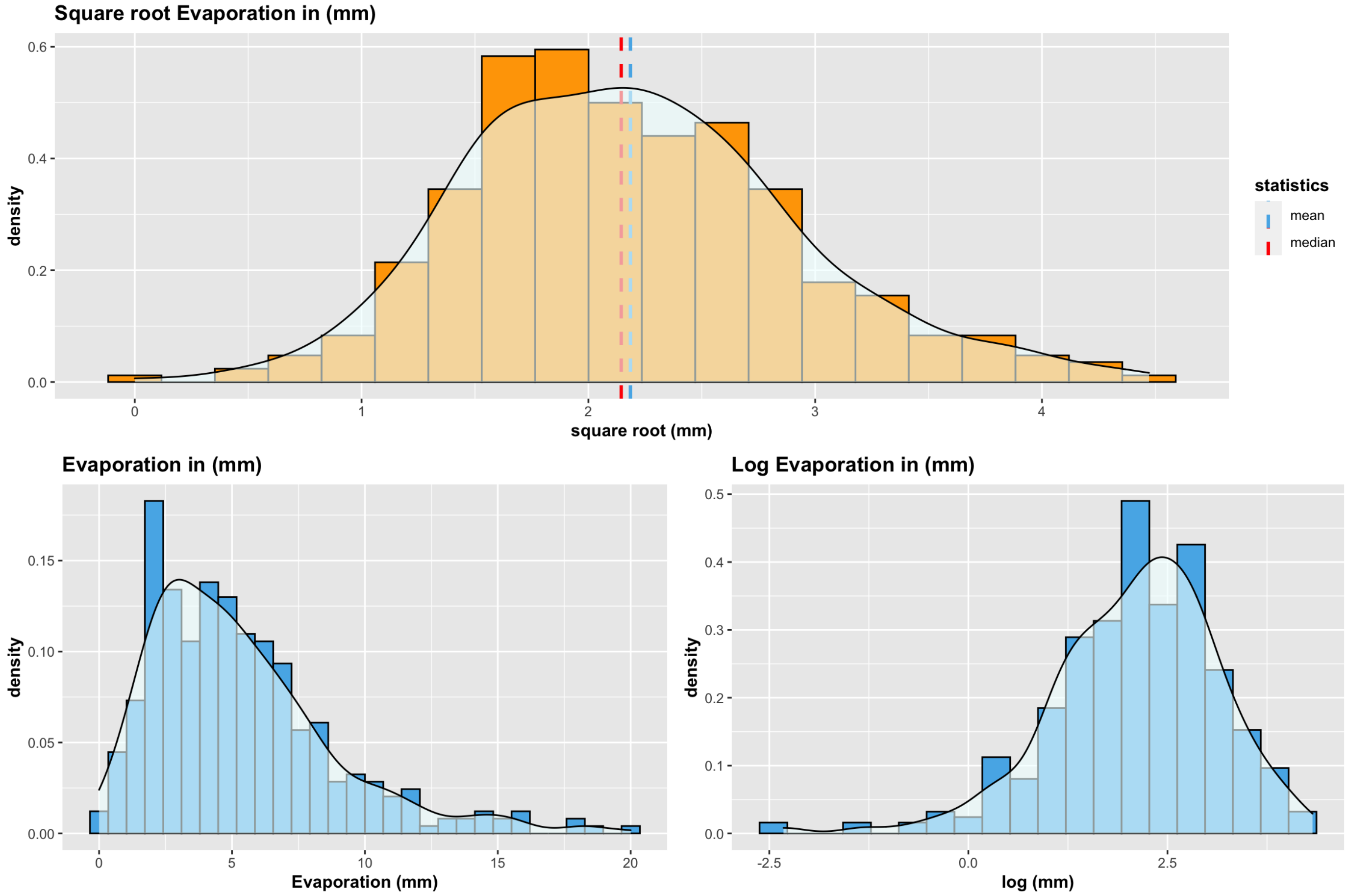
The skewness table indicates Maximum Temperature is moderately to highly right skewed (skewness statistic 0.94)

Histogram 1: Ploting Histrograms and boxplots to analyse shape, location, spread, and outlier

```
hist_evaporation <- df%>%
  ggplot(aes(x = `Evaporation (mm)`, y = ..density..)) +
  geom_histogram(col = 'black', bins = 30, fill= '#56B4E9') +
  labs(title = "Evaporation in (mm)") +
  xlab("Evaporation (mm)") +
  theme(title = element_text(face = 'bold')) +
  geom_density(alpha = 0.6,fill="azure") +
  scale_color_manual(name = "statistics", values = c(median = "red", mean = "#E69F00"
))

sqrt_hist_evaporation <- df%>%
  ggplot(aes(x = sqrt(`Evaporation (mm)`), y = ..density..)) +
  geom_histogram(col = 'black', bins = 20, fill= 'orange') +
  labs(title = "Square root Evaporation in (mm)") +
  xlab("square root (mm)") +
  theme(title = element_text(face = 'bold')) +
  geom_vline(aes(xintercept=mean(sqrt(`Evaporation (mm)`), na.rm =T),
                color="mean"), linetype="dashed", size=1) +
  geom_vline(aes(xintercept=median(sqrt(`Evaporation (mm)`), na.rm =T),
                color="median"), linetype="dashed", size=1) +
  geom_density(alpha = 0.6,fill="azure") +
  scale_color_manual(name = "statistics", values = c(median = "red", mean = "#56B4E9"
))

log10_hist_evaporation <- df%>%
  ggplot(aes(x = log2(`Evaporation (mm)`), y = ..density..)) +
  geom_histogram(col = 'black', bins = 20, fill= '#56B4E9') +
  labs(title = "Log Evaporation in (mm)") +
  xlab("log (mm)") +
  theme(title = element_text(face = 'bold')) +
  geom_density(alpha = 0.6,fill="azure") +
  scale_color_manual(name = "statistics", values = c(median = "red", mean = "#E69F00"
))
grid.arrange(sqrt_hist_evaporation,
             arrangeGrob(hist_evaporation, log10_hist_evaporation, ncol = 2),
             nrow = 2)
```



Shape

The skewness table indicates Evaporation is highly right skewed (Appendix: Table 1) (skewness statistic >1), this is evident in our histogram which depicts a right skewed unimodal distribution (Appendix: Histogram 1). Notably, our distribution has a mean (5.31) greater than its median (4.6) indicating a right skew.

Transformation

The distribution of Evaporation clearly warrants a transformation since it is highly skewed to the right. Notably, taking the square root of this distribution significantly improves its normality reducing its skewness statistic to 0.36 which is fairly symmetrical (Appendix: Table 1). Histogram 1 indicates taking the square root of this distribution is the appropriate transformation.

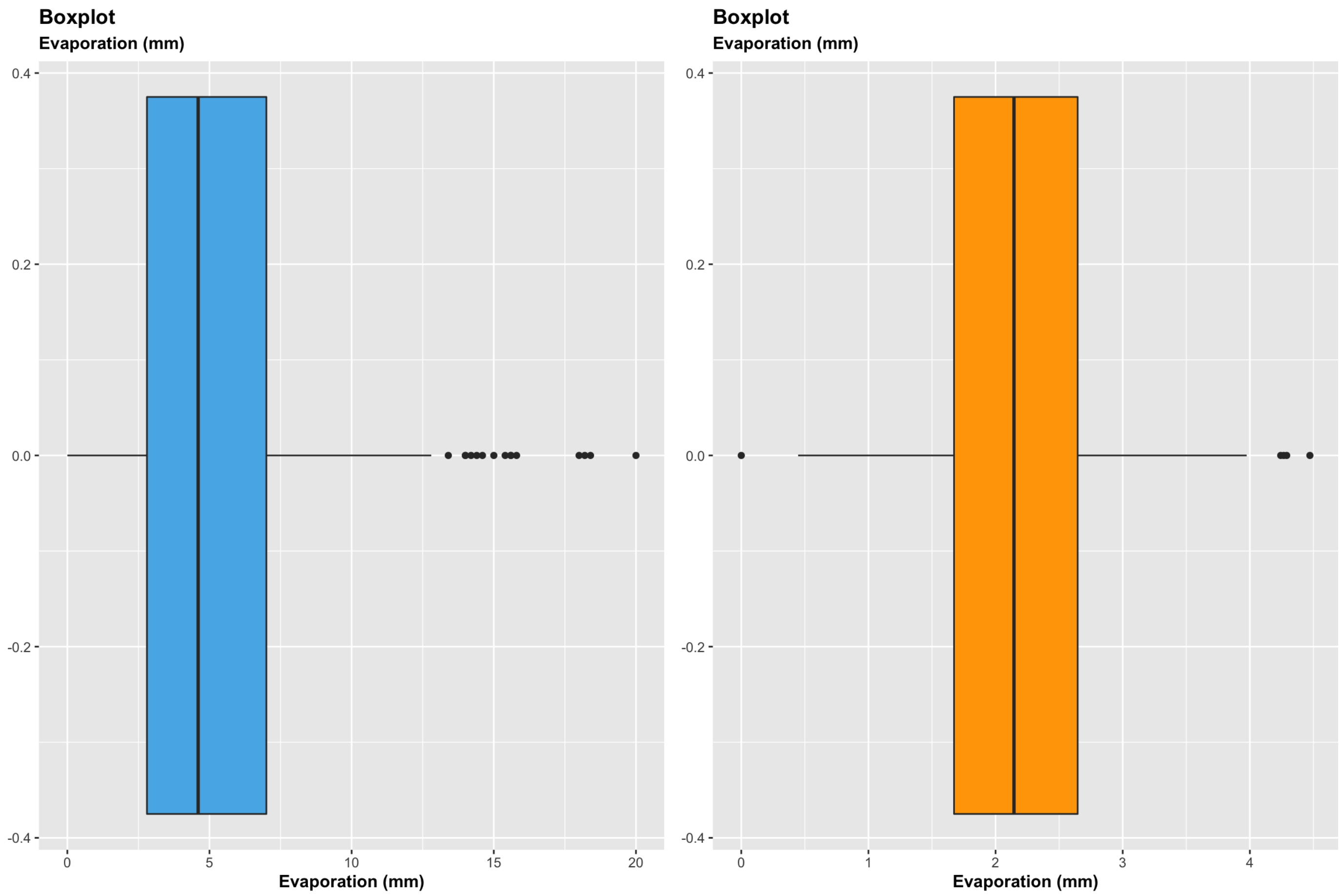
Location

As our distribution is highly right skewed the median is the preferred metric for central tendency as our mean seems to be distorted by 15 outliers. The location (median) of our distribution is 4.6.

Boxplot 1: Boxplot to describe Spread of Evaporation

```
boxplot_evaporation <- df %>%
  ggplot(aes(`Evaporation (mm)`)) +
  geom_boxplot(fill = '#56B4E9') +
  labs(title = "Boxplot", subtitle = "Evaporation (mm)") +
  xlab('Evaporation (mm)')+
  theme(title = element_text(face = 'bold'))

sqrt_boxplot_evaporation <- df %>%
  ggplot(aes(sqrt(`Evaporation (mm)`))) +
  geom_boxplot(fill = 'orange') +
  labs(title = "Boxplot", subtitle = "Evaporation (mm)") +
  xlab('Evaporation (mm)')+
  theme(title = element_text(face = 'bold'))
grid.arrange(boxplot_evaporation, sqrt_boxplot_evaporation, ncol=2)
```



Outliers

Outliers are generally indicated by laying more than 1.5 times the IQR the upper (Q3) and lower (Q1) quartiles (IQR 1). Notably, we have 15 outliers laying more than 1.5 times the upper quartile (Appendix: Boxplot 1). Specific notice should be paid to 4 extreme outliers all laying more than 2.6 times greater than our upper Q3. Interestingly, the 3 highest outliers occurred in March. Additionally, our distribution contains 8 null values.

Summary 1: Spread Using IQR

```
summary(df$`Evaporation (mm)`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   2.800   4.600   5.312   7.000  20.000       8
```

IQR 1: IQR

```
IQR(df$`Evaporation (mm)`, na.rm=T)
```

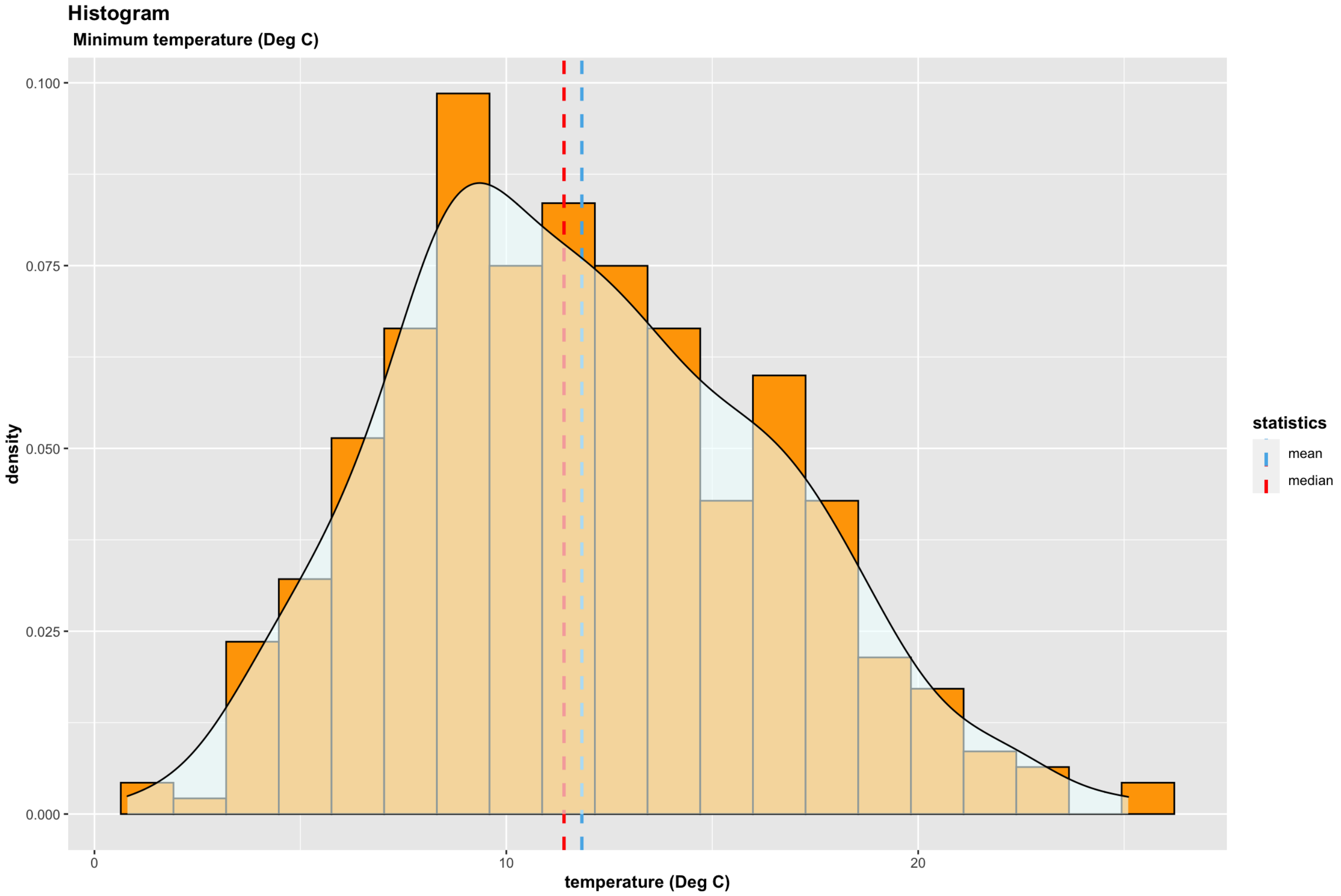
```
## [1] 4.2
```

Spread

Our distribution has an IQR of 4.2 indicating that the spread of the middle 50% of our set of Evaporation values lies within 2.8 to 7 mm with a range of 20 (min of 0 - max of 20) (Appendix: Summary 1).

Histogram 2: Distribution of Minimum temperature (Deg C)

```
df%>%
  ggplot(aes(x = `Minimum temperature (Deg C)`, y = ..density..)) +
  geom_histogram(col = 'black', bins = 20, fill= 'orange') +
  labs(title = "Histogram", subtitle = " Minimum temperature (Deg C)") +
  xlab("temperature (Deg C)") +
  theme(title = element_text(face = 'bold')) +
  geom_vline(aes(xintercept=mean(`Minimum temperature (Deg C)`, na.rm =T),
               color="mean"), linetype="dashed", size=1) +
  geom_vline(aes(xintercept=median(`Minimum temperature (Deg C)`, na.rm =T),
               color="median"), linetype="dashed", size=1) +
  geom_density(alpha = 0.6,fill="azure") +
  scale_color_manual(name = "statistics", values = c(median = "red", mean = "#56B4E9"
))
```



Shape

The skewness table indicates Minimum temperature is fairly symmetric (Appendix: Table 1) (skewness statistic of 0.31), this is evident in our histogram which depicts a fairly symmetric unimodal distribution with an appropriate right skew (Appendix: Histogram 2). Notably, our distribution has a mean (11.8) slightly greater than its median (11.4) indicating a slight right skew.

Transformation

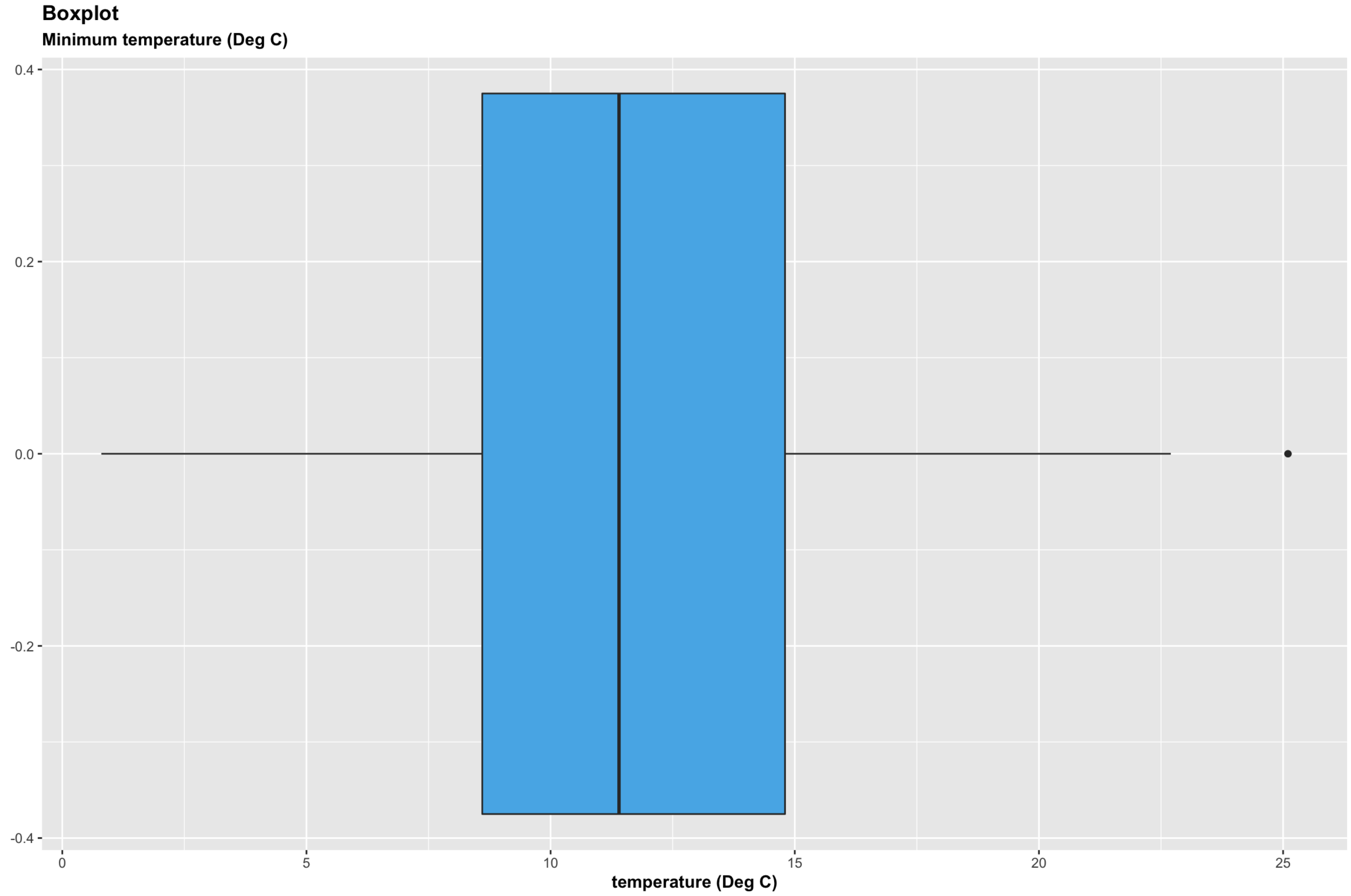
The distribution of Minimum temperature does not warrant a transformation since its slight right skew is acceptable.

Location

As our distribution is slightly skewed to the right the median is the preferred metric for central tendency as our mean seems to be distorted by 2 outliers. The location (median) of our distribution is 11.4.

Boxplot 2: Boxplot to describe Spread of Minimum temperature (Deg C)

```
df %>%
  ggplot(aes(`Minimum temperature (Deg C)`)) +
  geom_boxplot(fill = '#56B4E9') +
  labs(title = "Boxplot", subtitle = "Minimum temperature (Deg C)") +
  xlab('temperature (Deg C)')+
  theme(title = element_text(face = 'bold'))
```



Outliers

Notably, we have 2 outliers laying more than 1.5 times the upper quartile with no indication that they are correlated as they occurred 4 months apart (Appendix: Boxplot 2).

Summary 2: Spread Using IQR

```
summary(df$`Minimum temperature (Deg C)`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.80    8.60   11.40   11.83   14.80   25.10
```

IQR

```
IQR(df$`Minimum temperature (Deg C)`, na.rm=T)
```

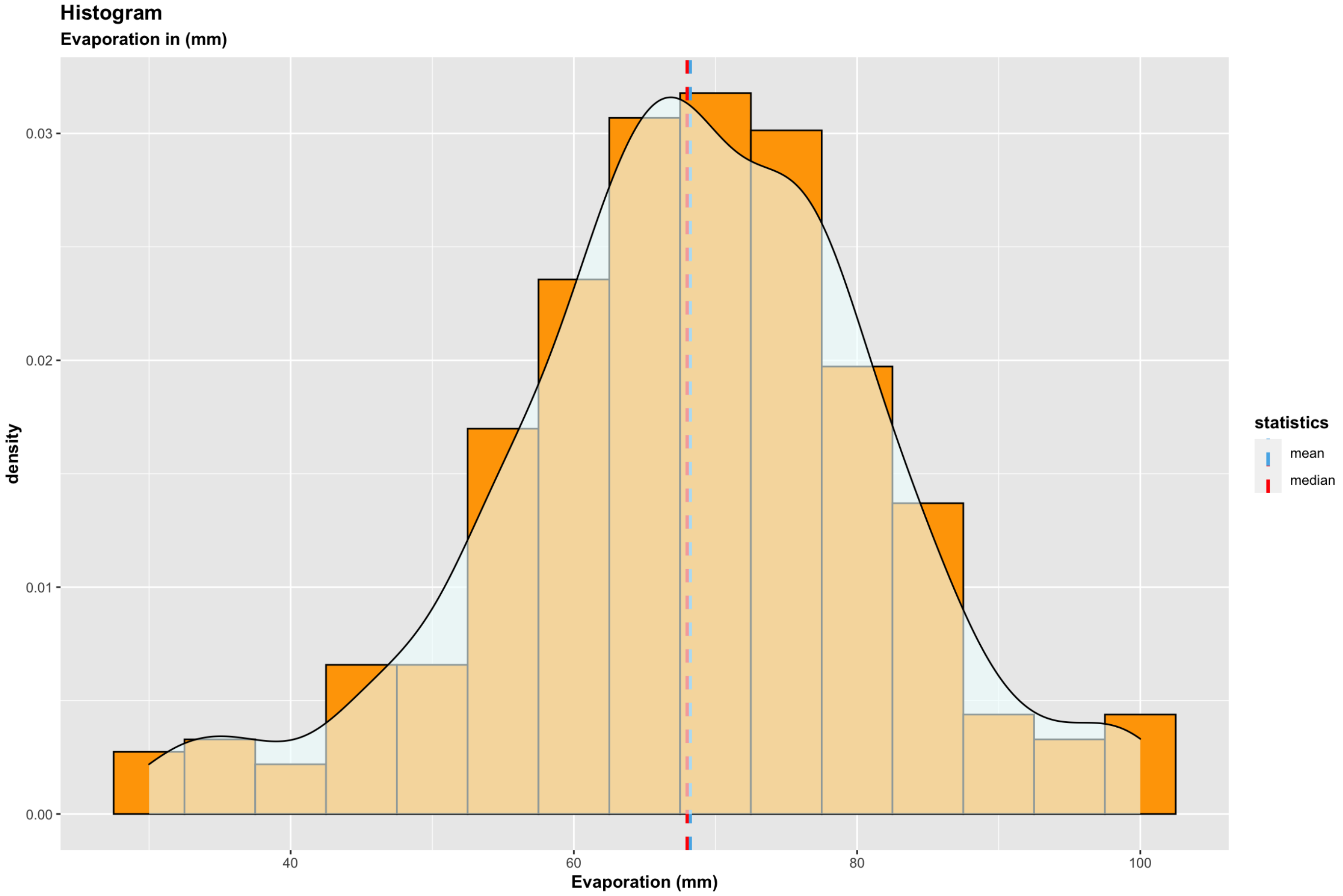
```
## [1] 6.2
```

Spread

Our distribution has an IQR of 6.2 indicating that the spread of the middle 50% of our set of Minimum temperature values lies within 8.6 to 14.8 degrees with a range of 24.3 (min of 0.8 - max of 25.1) (Appendix: Boxplot 2, Summary 2).

Histogram 3: Distribution of 9am relative humidity

```
df%>%
  ggplot(aes(x = `9am relative humidity (%)`, y = ..density..)) +
  geom_histogram(col = 'black', bins = 15, fill= 'orange') +
  labs(title = "Histogram", subtitle = "Evaporation in (mm)") +
  xlab("Evaporation (mm)") +
  theme(title = element_text(face = 'bold')) +
  geom_vline(aes(xintercept=mean(`9am relative humidity (%)`, na.rm =T),
                color="mean"), linetype="dashed", size=1) +
  geom_vline(aes(xintercept=median(`9am relative humidity (%)`, na.rm =T),
                color="median"), linetype="dashed", size=1) +
  geom_density(alpha = 0.6,fill="azure") +
  scale_color_manual(name = "statistics", values = c(median = "red", mean = "#56B4E9"
))
```



### Shape The skewness table indicates 9am relative humidity is fairly symmetric (Appendix: Table 1) (skewness statistic of -0.27), this is evident in our histogram which depicts a fairly symmetric unimodal distribution with an appropriate left skew (Appendix: Histogram 3).

Transformation

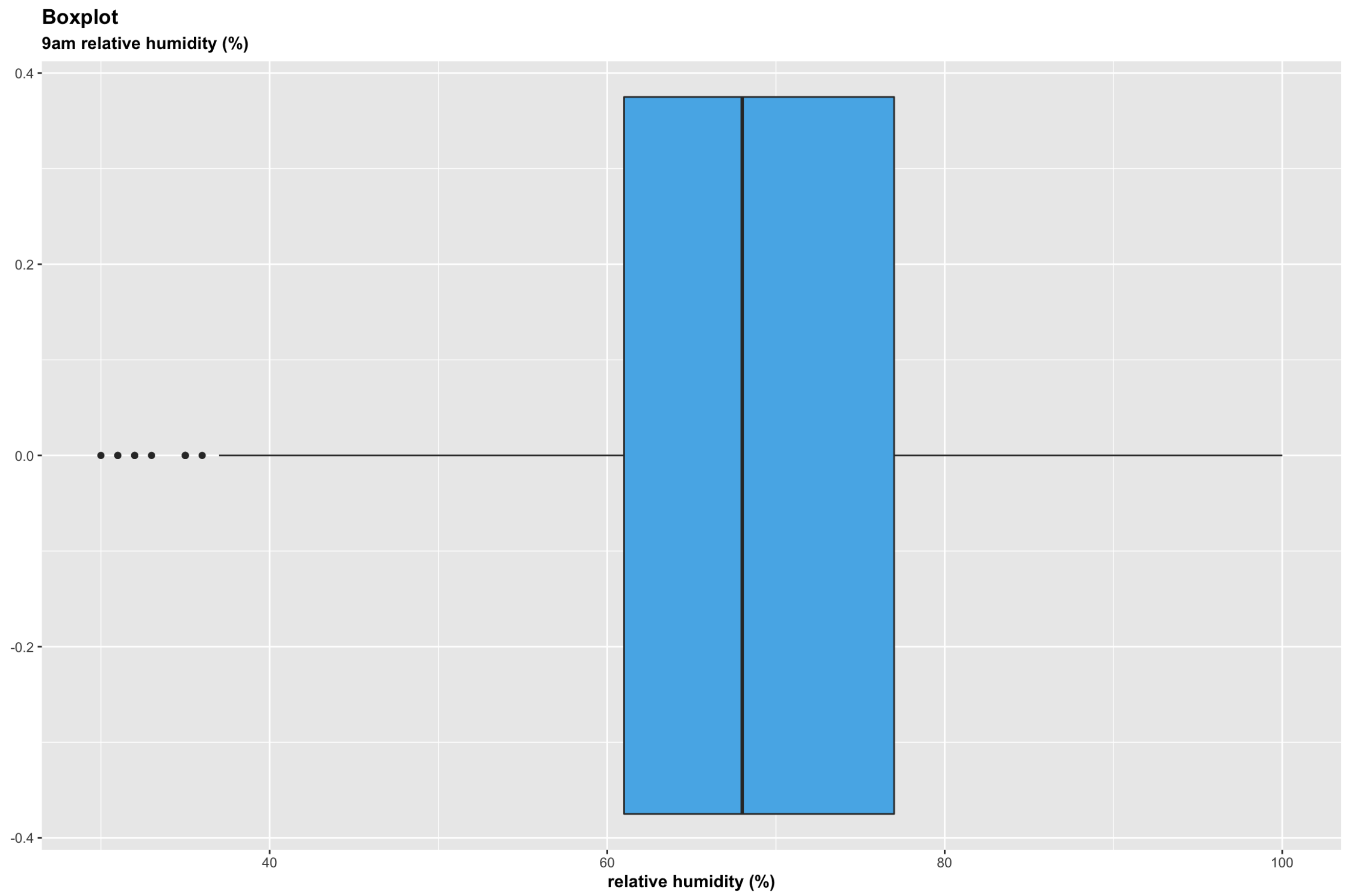
The distribution of 9am relative humidity does not warrant a transformation since its slight left skew is acceptable.

Location

As our distribution is slightly skewed to the left the median is the preferred metric for central tendency as our mean seems to be distorted by 10 outliers. The location (median) of our distribution is 68.

Boxplot 3: Boxplot to describe Spread of 9am relative humidity

```
df %>%
  ggplot(aes(`9am relative humidity (%)`)) +
  geom_boxplot(fill = '#56B4E9') +
  labs(title = "Boxplot", subtitle = "9am relative humidity (%)") +
  xlab('relative humidity (%)')+
  theme(title = element_text(face = 'bold'))
```



Outliers

Notably, we have 10 outliers laying less than 1.5 times the lower quartile. Interestingly, these outliers occurred in months that we would consider the hotter months (Jan-Apr & Oct-Dec) (Appendix: Boxplot 3).

Summary 3: Spread Using IQR

```
summary(df$`9am relative humidity (%)`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   30.00   61.00   68.00   68.22   77.00  100.00
```

IQR

```
IQR(df$`9am relative humidity (%)`, na.rm=T)
```

```
## [1] 16
```

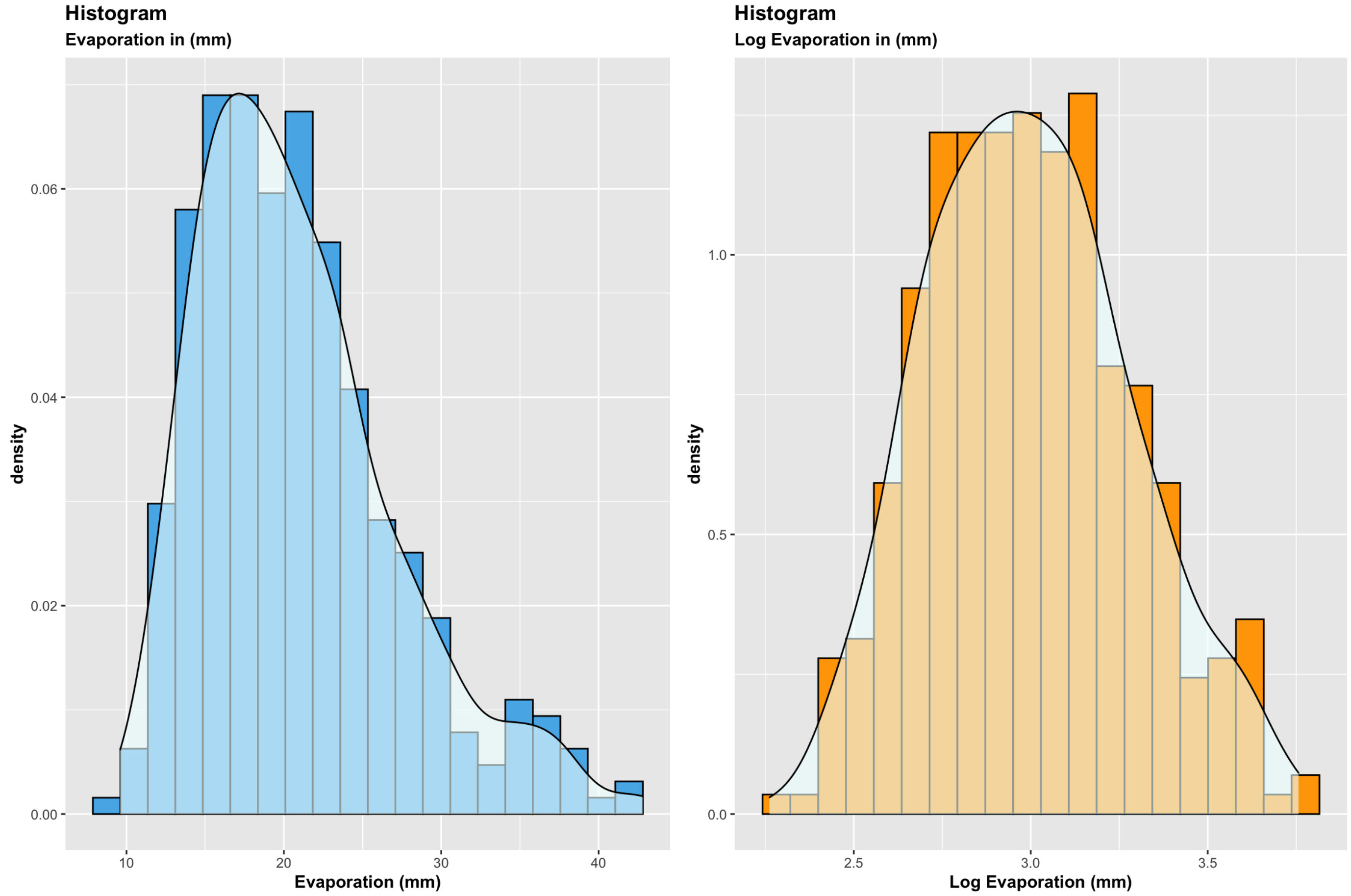
Spread

Our distribution has an IQR of 16 indicating that the spread of the middle 50% of our set of 9am relative humidity values lie within 61 to 77 percent with a range of 70 (min of 30 - max of 100) (Appendix: Boxplot 3, Summary 3).

Histogram 4: Distribution of Maximum Temperature (Deg C)

```
max_temp <- df%>%
  ggplot(aes(x = `Maximum Temperature (Deg C)`, y = ..density..)) +
  geom_histogram(col = 'black', bins = 20, fill= '#56B4E9') +
  labs(title = "Histogram", subtitle = "Evaporation in (mm)") +
  xlab("Evaporation (mm)") +
  theme(title = element_text(face = 'bold')) +
  geom_density(alpha = 0.6,fill="azure") +
  scale_color_manual(name = "statistics", values = c(median = "red", mean = "#E69F00"
))

log_max_temp <- df%>%
  ggplot(aes(x = log(`Maximum Temperature (Deg C)`), y = ..density..)) +
  geom_histogram(col = 'black', bins = 20, fill= 'orange') +
  labs(title = "Histogram", subtitle = "Log Evaporation in (mm)") +
  xlab("Log Evaporation (mm)") +
  theme(title = element_text(face = 'bold')) +
  geom_density(alpha = 0.6,fill="azure") +
  scale_color_manual(name = "statistics", values = c(median = "red", mean = "#56B4E9"
))
grid.arrange(max_temp, log_max_temp, ncol=2)
```



Shape

The skewness table indicates Maximum Temperature is moderately to highly right skewed (Appendix: Table 1) (skewness statistic 0.94), this is evident in our histogram which depicts a right skewed unimodal distribution (Appendix: Histogram 4). Notably, our distribution has a mean (20.87) greater than its median (19.7) indicating a right skew.

Transformation

The distribution of Maximum Temperature warrants a transformation since it is on the higher end of moderately skewed towards being highly skewed to the right. Notably, taking the log of this distribution significantly improves its normality reducing its skewness statistic to 0.24 which is fairly symmetrical (Appendix: Table 1). Histogram 4 indicates taking the log of this distribution is the appropriate transformation.

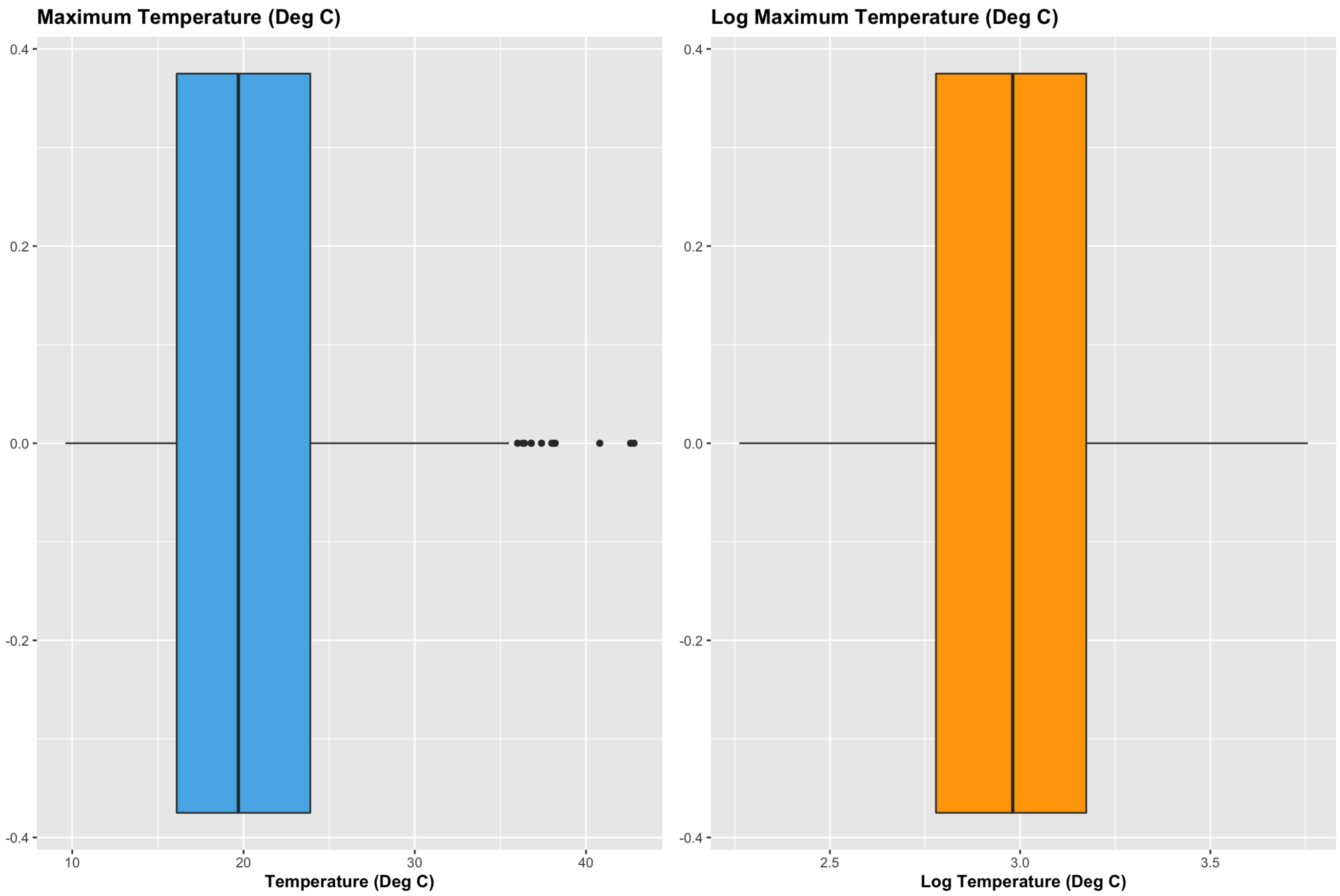
Location

As our distribution is highly right skewed the median is the preferred metric for central tendency as our mean seems to be distorted by 13 outliers. The location (median) of our distribution is 19.7.

Boxplot 4: Boxplot to describe Spread of Maximum Temperature (Deg C)

```
boxplot_maxTemp <- df %>%
  ggplot(aes(`Maximum Temperature (Deg C)`)) +
  geom_boxplot(fill = '#56B4E9') +
  labs(title = "Maximum Temperature (Deg C)") +
  xlab('Temperature (Deg C)')+
  theme(title = element_text(face = 'bold'))

log_boxplot_maxTemp <- df %>%
  ggplot(aes(log(`Maximum Temperature (Deg C)`))) +
  geom_boxplot(fill = 'orange') +
  labs(title = "Log Maximum Temperature (Deg C)") +
  xlab('Log Temperature (Deg C)')+
  theme(title = element_text(face = 'bold'))
grid.arrange(boxplot_maxTemp, log_boxplot_maxTemp, ncol=2)
```



Outliers

Notably, we have 13 outliers laying more than 1.5 times the upper quartile. Specific notice should be paid to 2 outliers that contain null values for evaporation. Additionally, 3 outliers are laying more than 2.1 times greater than our upper Q3. Notably, these outliers all occurred during summer periods (Dec-March)

Summary 4: Spread Using IQR

```
summary(df$`Maximum Temperature (Deg C)`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.60   16.10   19.70   20.87   23.90   42.80
```

IQR

```
IQR(df$`Maximum Temperature (Deg C)`, na.rm=T)
```

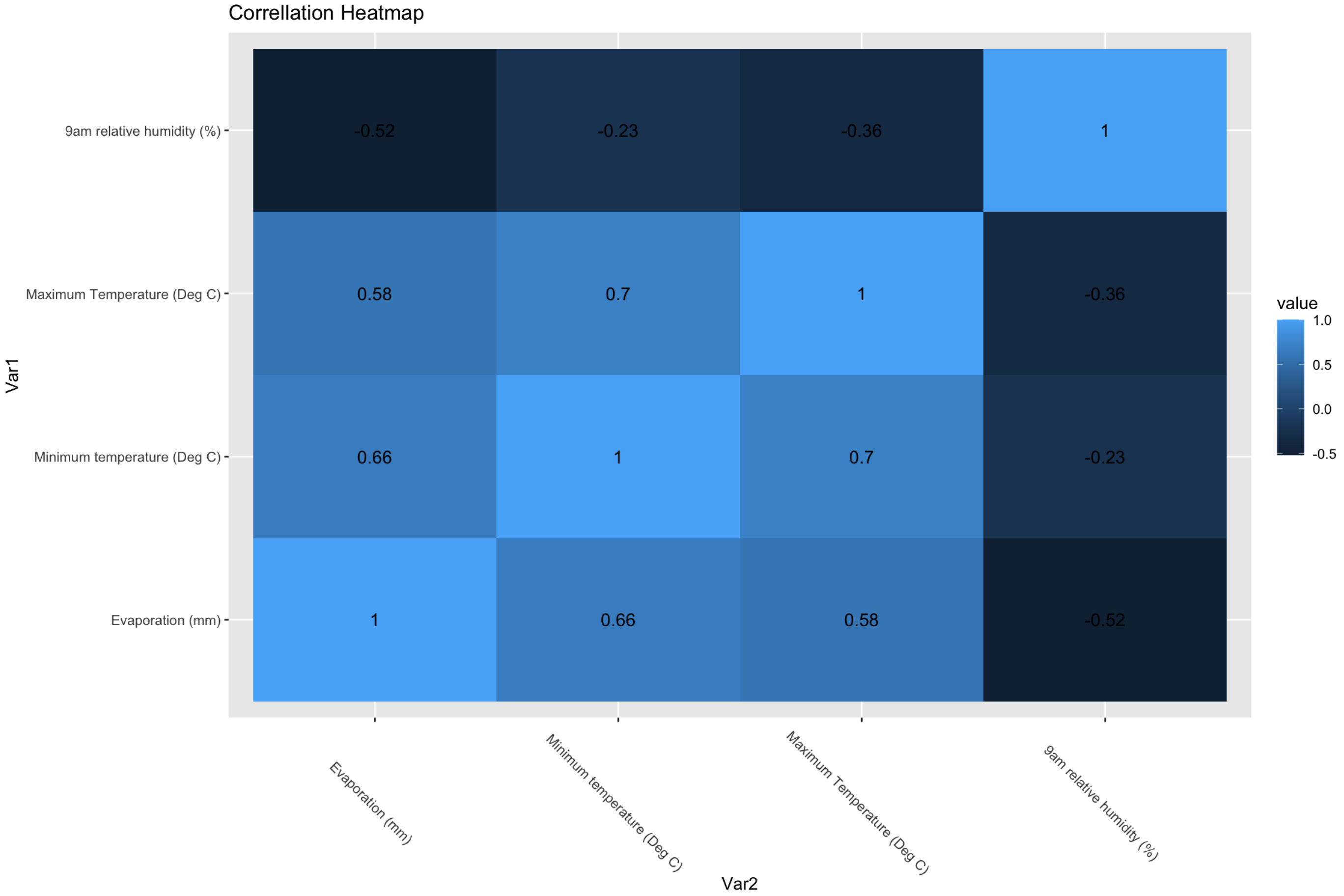
```
## [1] 7.8
```

Spread

Our distribution has an IQR of 7.8 indicating that the spread of the middle 50% of our set of maximum temperature values lies within 16.1 to 23.9 degrees with a range of 33.2 (min of 9.6 - max of 42.8) (Appendix: Boxplot 4, Summary 4).

Correlation Heatmap: Relationship

```
mydata <- df[, c(1,4,5,6)]
cormat <- round(cor(mydata,use="complete.obs"),2)
melted_cormat <- melt(cormat)
ggplot(data = melted_cormat, aes(x=Var2, y=Var1, fill=value)) +
  geom_tile()+
  geom_text(aes(Var2, Var1, label = value), color = "black", size = 4)+
  theme(axis.text.x = element_text(angle = -45))+
  labs(title = "Correllation Heatmap")
```

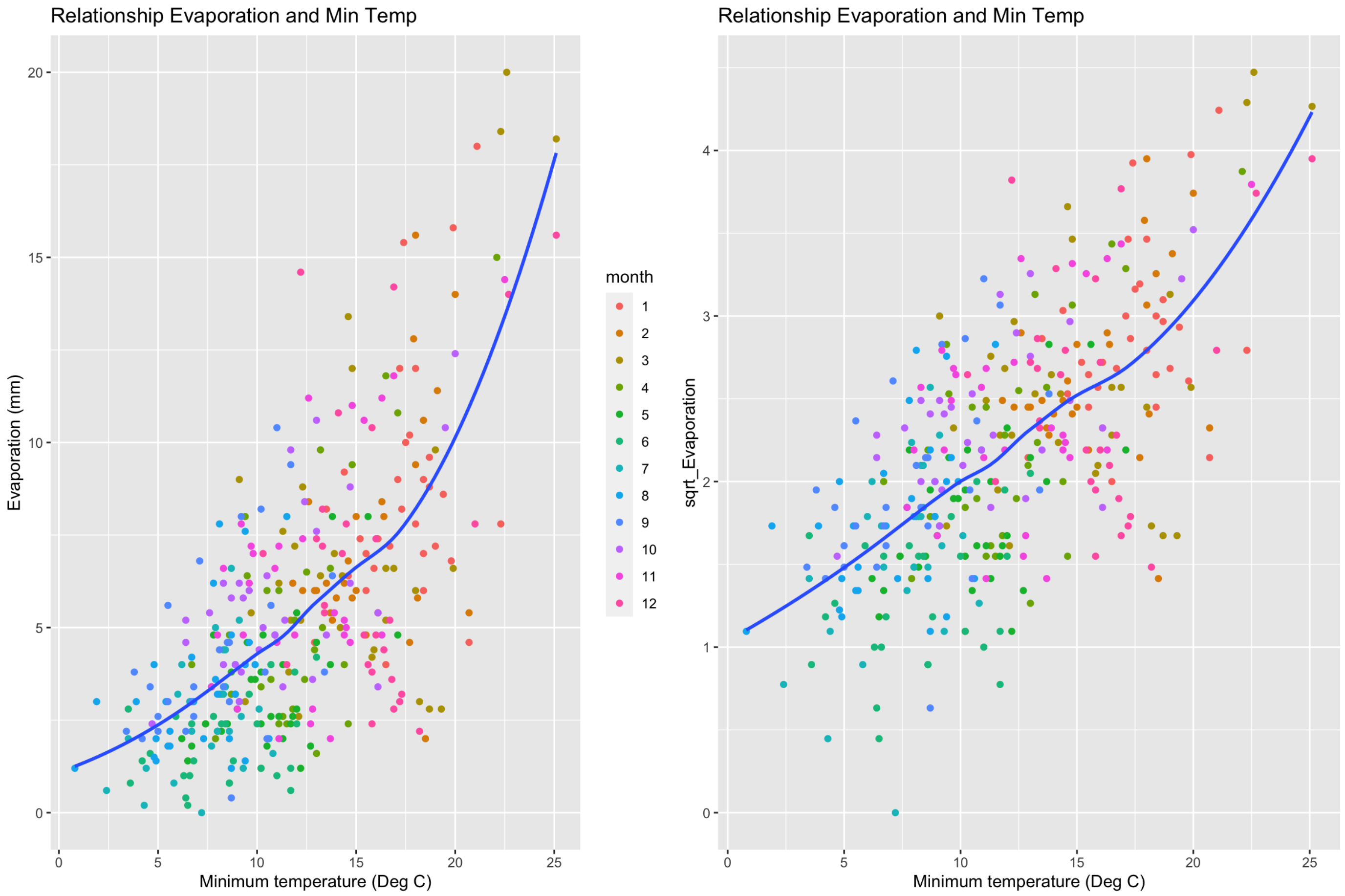


Scatterplot 1: Evaporation Min temp

```
linear_min<- df%>%
  ggplot(aes(y =`Evaporation (mm)` ,x =`Minimum temperature (Deg C)` )) +
  geom_point(aes(color = month))+
  geom_smooth(se=F)+
  labs(title = "Relationship Evaporation and Min Temp")

linear_min2 <- df%>%
  ggplot(aes(y =sqrt_Evaporation ,x =`Minimum temperature (Deg C)` )) +
  geom_point(aes(color = month),show.legend = F)+
  geom_smooth(se=F)+
  labs(title = "Relationship Evaporation and Min Temp")

grid.arrange(linear_min, linear_min2, ncol=2)
```

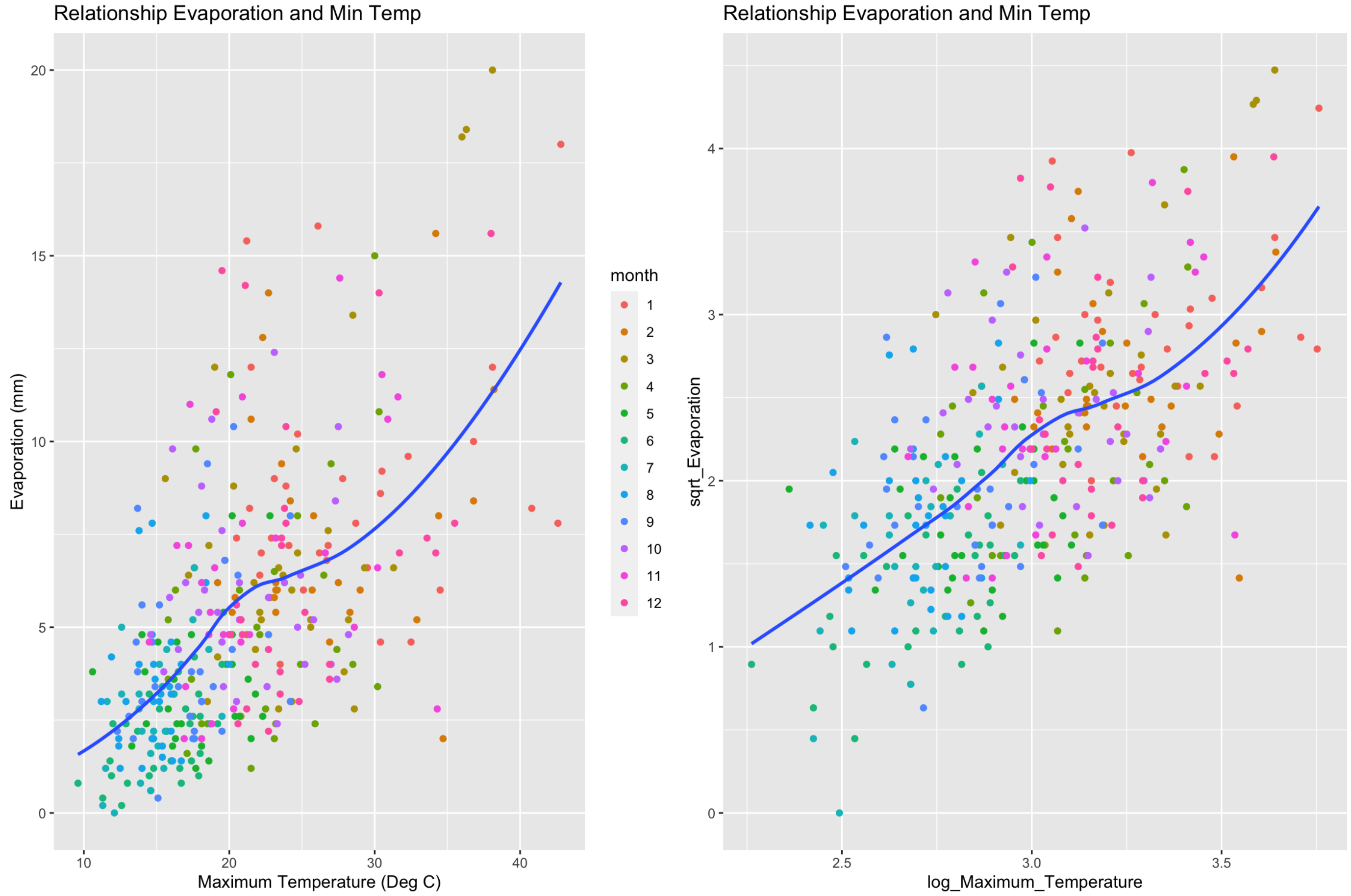


Scatterplot 2: Evaporation Max temp

```
linear_max<- df%>%
  ggplot(aes(y =`Evaporation (mm)` ,x =`Maximum Temperature (Deg C)` )) +
  geom_point(aes(color = month))+
  geom_smooth(se=F)+
  labs(title = "Relationship Evaporation and Min Temp")

linear_max2 <- df%>%
  ggplot(aes(y =sqrt_Evaporation ,x =log_Maximum_Temperature )) +
  geom_point(aes(color = month),show.legend = F)+
  geom_smooth(se=F)+
  labs(title = "Relationship Evaporation and Min Temp")

grid.arrange(linear_max, linear_max2, ncol=2)
```

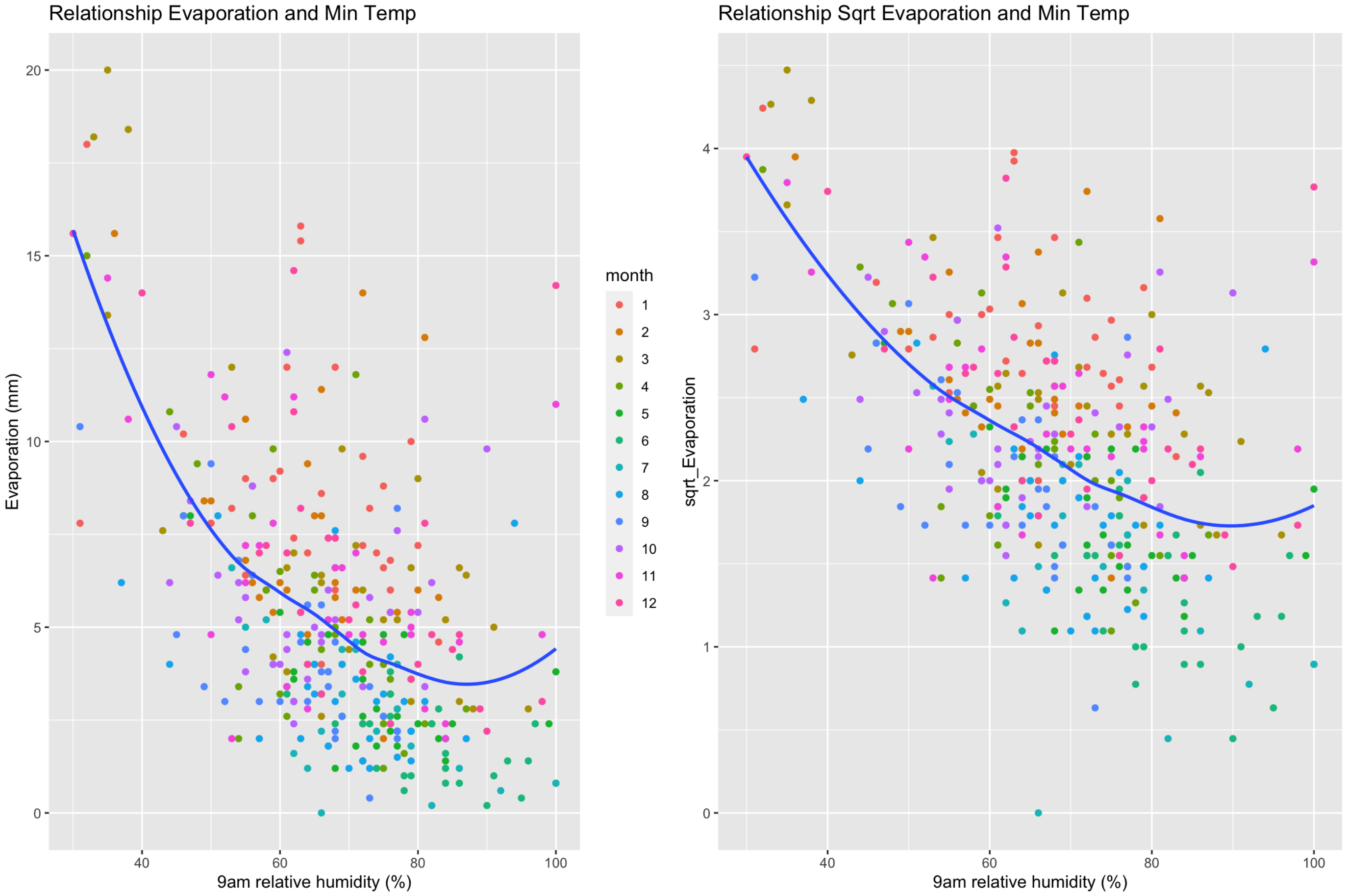


Scatterplot 3: Evaporation 9am relative humidity

```
linear_relative<- df%>%
  ggplot(aes(y =`Evaporation (mm)` ,x =`9am relative humidity (%)` )) +
  geom_point(aes(color = month))+
  geom_smooth(se=F)+
  labs(title = "Relationship Evaporation and Min Temp")

linear_relative2 <- df%>%
  ggplot(aes(y =sqrt_Evaporation ,x = `9am relative humidity (%)`  )) +
  geom_point(aes(color = month),show.legend = F)+
  geom_smooth(se=F)+
  labs(title = "Relationship Sqrt Evaporation and Min Temp")

grid.arrange(linear_relative, linear_relative2, ncol=2)
```



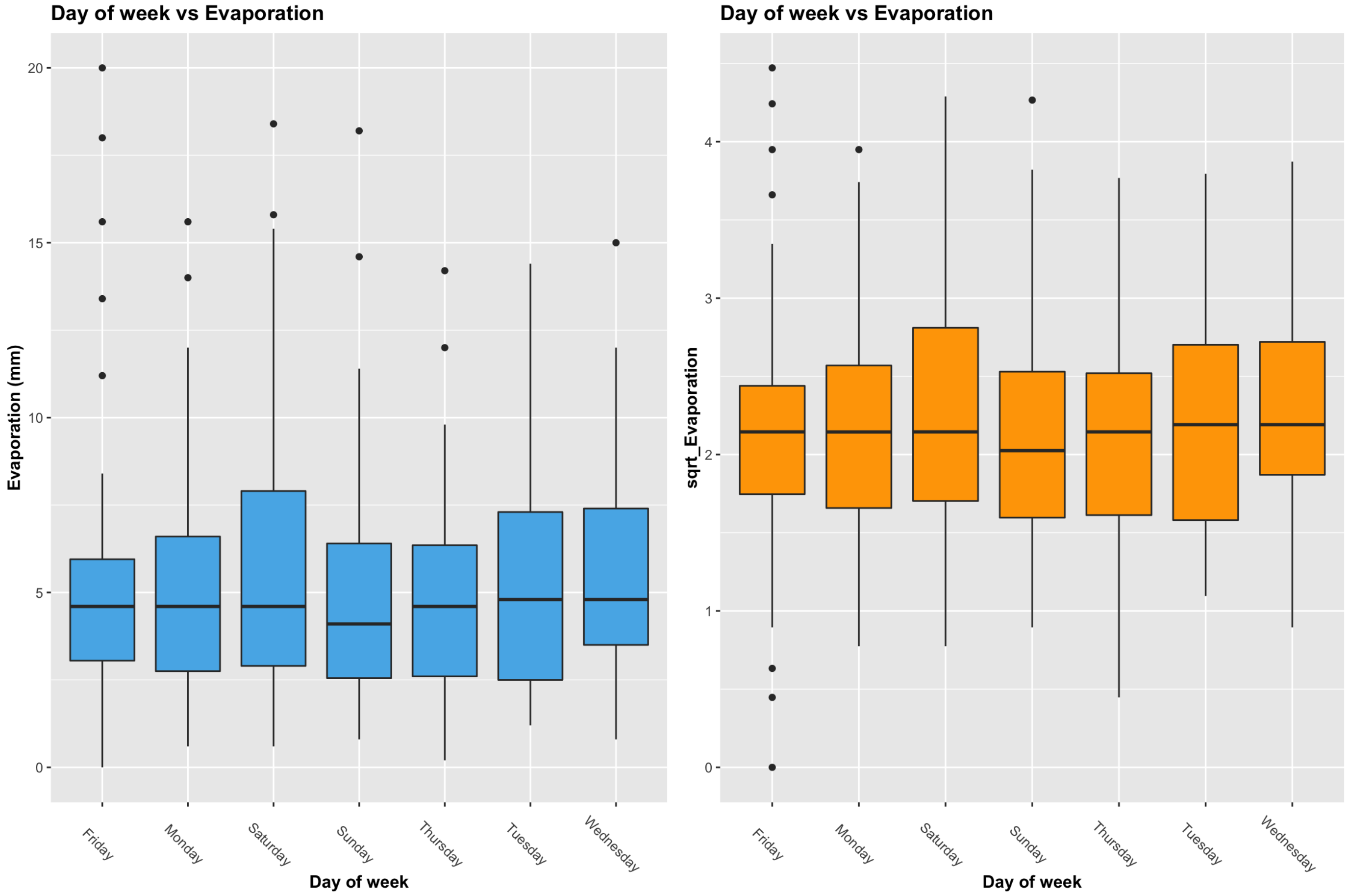
Relationship Evaporation and Min Temp

Relationship Sqrt Evaporation and Min Temp

Boxplot 5: Boxplot day vs evaporation

```
day_evaporation <- df %>%
  ggplot(aes(x = day_of_week, y= `Evaporation (mm)`)) +
  geom_boxplot(fill = '#56B4E9') +
  labs(title = "Day of week vs Evaporation") +
  xlab("Day of week")+
  theme(title = element_text(face = 'bold'))+
  theme(axis.text.x = element_text(angle = -45))

day_sqrt_evaporation <- df %>%
  ggplot(aes(x = day_of_week, y= sqrt_Evaporation)) +
  geom_boxplot(fill = 'orange') +
  labs(title = "Day of week vs Evaporation") +
  xlab("Day of week")+
  theme(title = element_text(face = 'bold'))+
  theme(axis.text.x = element_text(angle = -45))
grid.arrange(day_evaporation, day_sqrt_evaporation, ncol=2)
```



Shape

Roughly symmetrical with evaporation on the square root scale.

Location

The medians (horizontal bar in the rectangles) are roughly the same, with a noticeable decrease around Sunday and Thursday.

Spread

The IQR differs for each month. Notably, Saturday and Tuesday contain the greatest spread displaying the widest IQR. Additionally, Friday illustrates the smallest spread.

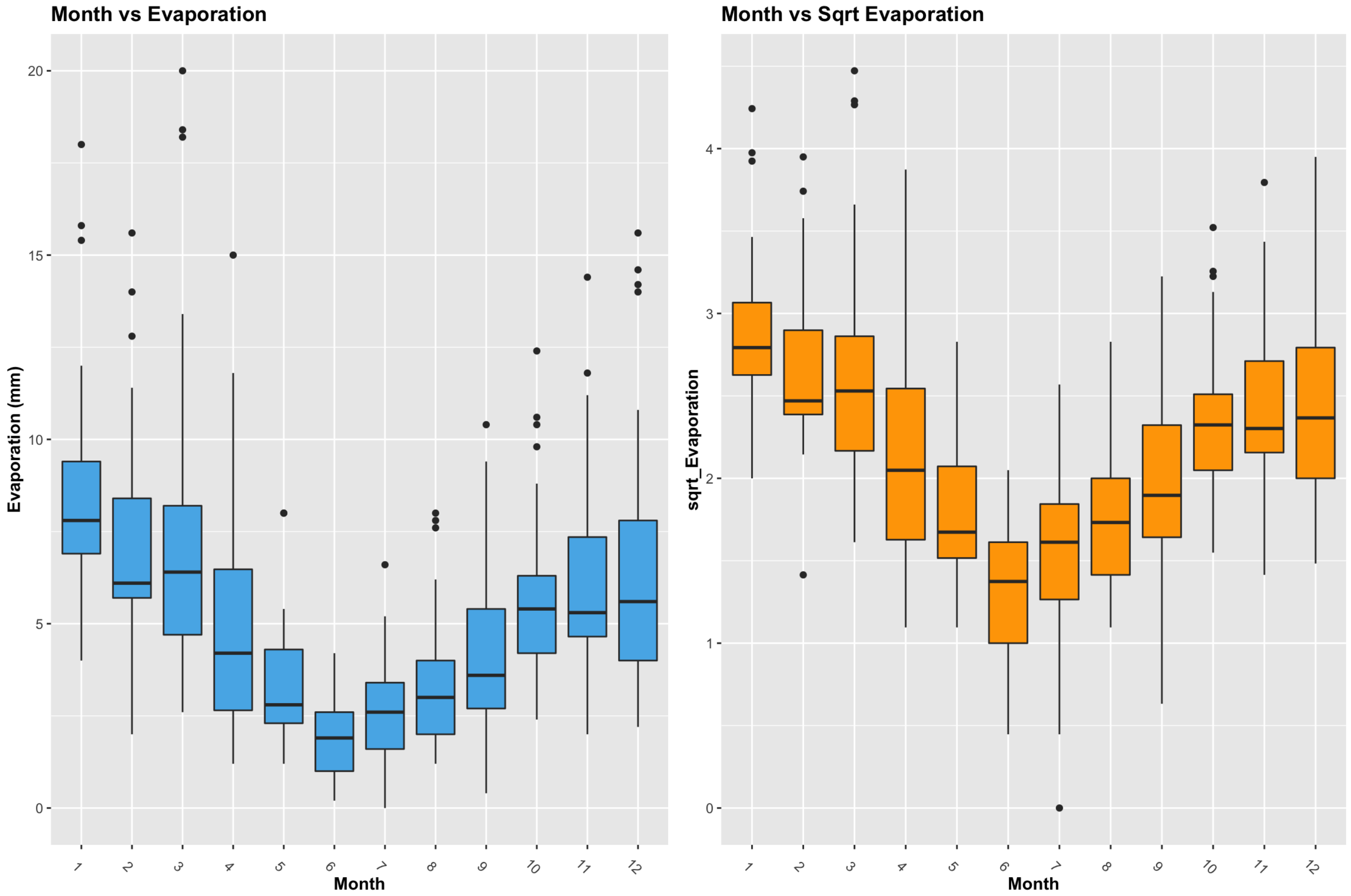
Outliers

Notably, transforming the distribution has reduced the number of outliers within our dataset. The only potential outliers are for Friday, Monday and Sunday with evaporation on the square root scale.

Boxplot 6: Boxplot month vs evaporation

```
month_evaporation <- df %>%
  ggplot(aes(x = month, y= `Evaporation (mm)`)) +
  geom_boxplot(fill = '#56B4E9') +
  labs(title = "Month vs Evaporation") +
  xlab("Month")+
  theme(title = element_text(face = 'bold'))+
  theme(axis.text.x = element_text(angle = -45))

month_sqrt_evaporation <- df %>%
  ggplot(aes(x = month, y= sqrt_Evaporation)) +
  geom_boxplot(fill = 'orange') +
  labs(title = "Month vs Sqrt Evaporation") +
  xlab("Month")+
  theme(title = element_text(face = 'bold'))+
  theme(axis.text.x = element_text(angle = -45))
grid.arrange(month_evaporation, month_sqrt_evaporation, ncol=2)
```



Shape

The shape of our distribution is quadratic with a clear convexity around May to August.

Location

The medians (horizontal bar in the rectangles) are all at different levels which is expected due to the convexity of the relationship with Evaporation. We have medians from April to September that clearly illustrate 6 months of lower evaporation levels and October to March that clearly depict 6 months of higher evaporation levels.

Spread

The IQR differs for each month. Notably, March and December contain the greatest spread displaying the widest IQR. Additionally, January and October contain the smallest spread.

Outliers

Notably, transforming the distribution has reduced the number of outliers within our dataset. The only potential outliers are for January, February, March, July, October, and November with evaporation on the square root scale.

Linear Models

Step 1

```
step1_lm <- lm(sqrt(`Evaporation (mm)`)~ day_of_week + month + `Minimum temperature
(Deg C)`+
             log(`Maximum Temperature (Deg C)`)+ `9am relative humidity (%)` +
             `9am relative humidity (%)`:month ,data = df)
```

Step 1: P-values for Quantitative Variables

```
summary(step1_lm)
```

```
## 
## Call:
## lm(formula = sqrt(`Evaporation (mm)`) ~ day_of_week + month + 
##     `Minimum temperature (Deg C)` + log(`Maximum Temperature (Deg C)`) + 
##     `9am relative humidity (%)` + `9am relative humidity (%)`:month, 
##     data = df)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.52989 -0.27198  0.02943  0.25458  1.98388 
## 
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                          2.4698848  0.6427763   3.843 0.000146 ***
## day_of_weekMonday                    0.0810739  0.0929458   0.872 0.383705    
## day_of_weekSaturday                  0.2253079  0.0928339   2.427 0.015767 *  
## day_of_weekSunday                    0.1060134  0.0922357   1.149 0.251247    
## day_of_weekThursday                 -0.0151301  0.0934991  -0.162 0.871548    
## day_of_weekTuesday                   0.0999861  0.0939085   1.065 0.287794    
## day_of_weekWednesday                 0.1149561  0.0938971   1.224 0.221735    
## month2                               0.3011003  0.6949546   0.433 0.665109    
## month3                               0.8091040  0.5473624   1.478 0.140326    
## month4                               0.5266418  0.6458451   0.815 0.415423    
## month5                              -0.4414309  0.6949048  -0.635 0.525719    
## month6                              -0.6906895  0.8254309  -0.837 0.403341    
## month7                              -0.2414661  0.7433270  -0.325 0.745507    
## month8                              -0.8973844  0.6684679  -1.342 0.180387    
## month9                               0.3217211  0.6566481   0.490 0.624504    
## month10                             -1.0912093  0.6482137  -1.683 0.093256 .  
## month11                             -0.1181664  0.5794662  -0.204 0.838541    
## month12                              0.2534829  0.5816560   0.436 0.663275    
## `Minimum temperature (Deg C)`        0.0615660  0.0092444   6.660 1.17e-10 ***
## log(`Maximum Temperature (Deg C)`)   0.0882774  0.1447815   0.610 0.542467    
## `9am relative humidity (%)`         -0.0160463  0.0067637  -2.372 0.018255 *  
## month2:`9am relative humidity (%)`  -0.0064228  0.0106047  -0.606 0.545162    
## month3:`9am relative humidity (%)`  -0.0133029  0.0082332  -1.616 0.107114    
## month4:`9am relative humidity (%)`  -0.0127516  0.0098064  -1.300 0.194411    
## month5:`9am relative humidity (%)`  -0.0006539  0.0099449  -0.066 0.947616    
## month6:`9am relative humidity (%)`   0.0000504  0.0109956   0.005 0.996345    
## month7:`9am relative humidity (%)`  -0.0045885  0.0106745  -0.430 0.667582    
## month8:`9am relative humidity (%)`   0.0079695  0.0098678   0.808 0.419896    
## month9:`9am relative humidity (%)`  -0.0100449  0.0102626  -0.979 0.328413    
## month10:`9am relative humidity (%)`  0.0153906  0.0098794   1.558 0.120241    
## month11:`9am relative humidity (%)`  0.0007717  0.0086811   0.089 0.929217    
## month12:`9am relative humidity (%)` -0.0062885  0.0086105  -0.730 0.465714    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.4524 on 325 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.6529, Adjusted R-squared:  0.6198 
## F-statistic: 19.72 on 31 and 325 DF,  p-value: < 2.2e-16
```

Step 1: P-values for Catergorical Variables

```
anova(step1_lm)
```

```
## Analysis of Variance Table
##
## Response: sqrt(`Evaporation (mm)`)
##                                      Df Sum Sq Mean Sq F value    Pr(>F)
## day_of_week                           6  2.475  0.4125  2.0151 0.0632359 .
## month                                11 77.520  7.0473 34.4280 < 2.2e-16 ***
## `Minimum temperature (Deg C)`         1 20.110 20.1101 98.2440 < 2.2e-16 ***
## log(`Maximum Temperature (Deg C)`)    1  2.448  2.4477 11.9578 0.0006166 ***
## `9am relative humidity (%)`           1 18.935 18.9348 92.5023 < 2.2e-16 ***
## month:`9am relative humidity (%)`    11  3.640  0.3309  1.6164 0.0926635 .
## Residuals                           325 66.526  0.2047
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

step 2

```
step2_lm <- lm(sqrt(`Evaporation (mm)`)~ day_of_week + month + `Minimum temperature (Deg C)`+
                 log(`Maximum Temperature (Deg C)`)+ `9am relative humidity (%)`,data = df)
```

Step 2: P-values for Quantitative Variables

```
summary(step2_lm)
```

```
##
## Call:
## lm(formula = sqrt(`Evaporation (mm)`) ~ day_of_week + month +
##     `Minimum temperature (Deg C)` + log(`Maximum Temperature (Deg C)`) +
##     `9am relative humidity (%)`, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.51860 -0.27873  0.02511  0.27158  1.87516
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          2.648049   0.527489   5.020 8.39e-07 ***
## day_of_weekMonday                    0.075992   0.090978   0.835  0.40416
## day_of_weekSaturday                  0.237397   0.091609   2.591  0.00997 **
## day_of_weekSunday                    0.095713   0.091261   1.049  0.29503
## day_of_weekThursday                  0.004993   0.091874   0.054  0.95669
## day_of_weekTuesday                   0.118808   0.091483   1.299  0.19494
## day_of_weekWednesday                 0.121912   0.091921   1.326  0.18565
## month2                              -0.107666   0.120238  -0.895  0.37119
## month3                              -0.069042   0.118301  -0.584  0.55987
## month4                              -0.295139   0.127701  -2.311  0.02143 *
## month5                              -0.442327   0.136282  -3.246  0.00129 **
## month6                              -0.608323   0.150214  -4.050 6.37e-05 ***
## month7                              -0.519013   0.161392  -3.216  0.00143 **
## month8                              -0.306639   0.156196  -1.963  0.05045 .
## month9                              -0.288656   0.149631  -1.929  0.05456 .
## month10                             -0.083580   0.130048  -0.643  0.52087
## month11                             -0.042581   0.126302  -0.337  0.73623
## month12                             -0.160415   0.119706  -1.340  0.18113
## `Minimum temperature (Deg C)`        0.064323   0.009100   7.069 9.09e-12 ***
## log(`Maximum Temperature (Deg C)`)  0.082306   0.143930   0.572  0.56781
## `9am relative humidity (%)`         -0.019375   0.002035  -9.522  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.457 on 336 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.6339, Adjusted R-squared:  0.6121
## F-statistic: 29.09 on 20 and 336 DF,  p-value: < 2.2e-16
```

Step 2: P-values for Catergorical Variables

```
anova(step2_lm)
```

```
## Analysis of Variance Table
##
## Response: sqrt(`Evaporation (mm)`)
##                                     Df Sum Sq Mean Sq F value    Pr(>F)
## day_of_week                          6  2.475  0.4125  1.9752 0.0686094 .
## month                               11 77.520  7.0473 33.7471 < 2.2e-16 ***
## `Minimum temperature (Deg C)`        1 20.110 20.1101 96.3008 < 2.2e-16 ***
## log(`Maximum Temperature (Deg C)`)   1  2.448  2.4477 11.7213 0.0006943 ***
## `9am relative humidity (%)`          1 18.935 18.9348 90.6726 < 2.2e-16 ***
## Residuals                          336 70.166  0.2088
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

step 3

```
step3_lm <-lm(sqrt(`Evaporation (mm)`)~ month + `Minimum temperature (Deg C)`+
                log(`Maximum Temperature (Deg C)`)+ `9am relative humidity (%)`,data
= df)
```

Step 3: P-values for Quantitative Variables

```
summary(step3_lm)
```

```
##
## Call:
## lm(formula = sqrt(`Evaporation (mm)`) ~ month + `Minimum temperature (Deg C)` +
##     log(`Maximum Temperature (Deg C)`) + `9am relative humidity (%)`,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.61803 -0.28671  0.01691  0.25310  1.76446
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         2.832974   0.518240   5.467 8.86e-08 ***
## month2                             -0.107193   0.120769  -0.888 0.375390
## month3                             -0.068334   0.118723  -0.576 0.565277
## month4                             -0.294983   0.128022  -2.304 0.021813 *
## month5                             -0.454137   0.136542  -3.326 0.000977 ***
## month6                             -0.613075   0.150417  -4.076 5.71e-05 ***
## month7                             -0.528030   0.161396  -3.272 0.001178 **
## month8                             -0.318643   0.156345  -2.038 0.042311 *
## month9                             -0.286649   0.149955  -1.912 0.056767 .
## month10                            -0.082189   0.130353  -0.631 0.528782
## month11                            -0.050091   0.126695  -0.395 0.692820
## month12                            -0.151449   0.120105  -1.261 0.208178
## `Minimum temperature (Deg C)`       0.065907   0.009037   7.293 2.12e-12 ***
## log(`Maximum Temperature (Deg C)`)  0.043622   0.143194   0.305 0.760830
## `9am relative humidity (%)`        -0.019245   0.002042  -9.425  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4592 on 342 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.6237, Adjusted R-squared:  0.6083
## F-statistic:  40.5 on 14 and 342 DF,  p-value: < 2.2e-16
```

Step 3: P-values for Catergorical Variables

```
anova(step3_lm)
```

```
## Analysis of Variance Table
##
## Response: sqrt(`Evaporation (mm)`)
##                                     Df Sum Sq Mean Sq F value    Pr(>F)
## month                               11 77.713  7.0648 33.5058 < 2.2e-16 ***
## `Minimum temperature (Deg C)`        1 21.007 21.0065 99.6260 < 2.2e-16 ***
## log(`Maximum Temperature (Deg C)`)   1  2.090  2.0897  9.9108  0.001788 **
## `9am relative humidity (%)`          1 18.732 18.7320 88.8390 < 2.2e-16 ***
## Residuals                          342 72.112  0.2109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

step 4

```
step4_lm <-lm(sqrt(`Evaporation (mm)`)~ month + `Minimum temperature (Deg C)`+
                `9am relative humidity (%)`,data = df)
```

Step 4: P-values for Quantitative Variables

```
summary(step4_lm)
```

```
##
## Call:
## lm(formula = sqrt(`Evaporation (mm)`) ~ month + `Minimum temperature (Deg C)` +
##     `9am relative humidity (%)`, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.62696 -0.29486  0.01309  0.25042  1.76245
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    2.974509   0.229292  12.973  < 2e-16 ***
## month2                        -0.109306   0.120410  -0.908 0.364630
## month3                        -0.071780   0.118026  -0.608 0.543476
## month4                        -0.300882   0.126382  -2.381 0.017823 *
## month5                        -0.466600   0.130096  -3.587 0.000384 ***
## month6                        -0.628396   0.141572  -4.439 1.22e-05 ***
## month7                        -0.546195   0.149779  -3.647 0.000307 ***
## month8                        -0.335948   0.145467  -2.309 0.021513 *
## month9                        -0.300400   0.142810  -2.103 0.036150 *
## month10                       -0.088992   0.128256  -0.694 0.488237
## month11                       -0.056596   0.124718  -0.454 0.650264
## month12                       -0.154435   0.119546  -1.292 0.197282
## `Minimum temperature (Deg C)`  0.066757   0.008584   7.777 8.77e-14 ***
## `9am relative humidity (%)`   -0.019424   0.001953  -9.946  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4586 on 343 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.6236, Adjusted R-squared:  0.6094
## F-statistic: 43.72 on 13 and 343 DF,  p-value: < 2.2e-16
```

Step 4: P-values for Catergorical Variables

```
anova(step4_lm)
```

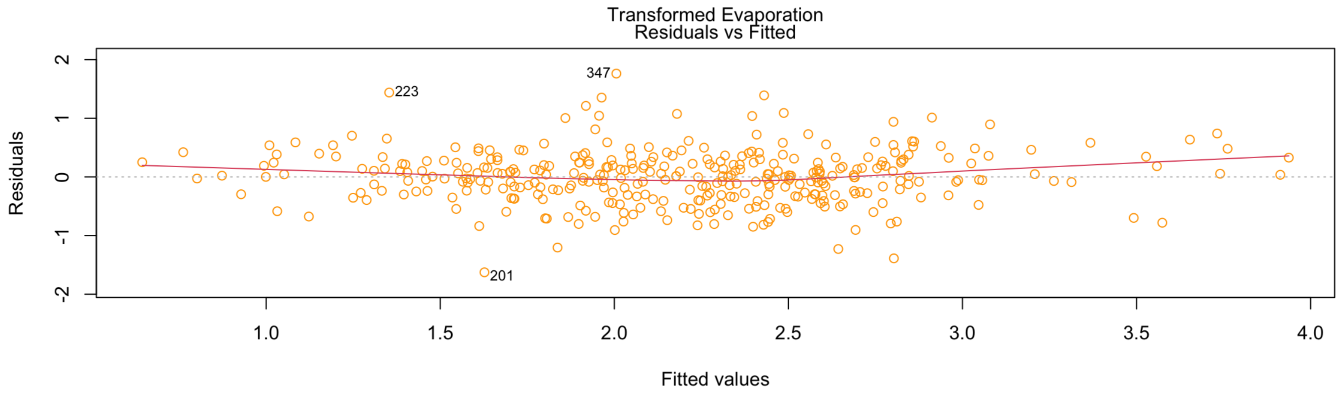
```
## Analysis of Variance Table
##
## Response: sqrt(`Evaporation (mm)`)
##                                 Df Sum Sq Mean Sq F value    Pr(>F)    
## month                           11 77.713  7.0648  33.595 < 2.2e-16 ***
## `Minimum temperature (Deg C)`    1 21.007 21.0065  99.890 < 2.2e-16 ***
## `9am relative humidity (%)`      1 20.802 20.8022  98.919 < 2.2e-16 ***
## Residuals                      343 72.131  0.2103                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

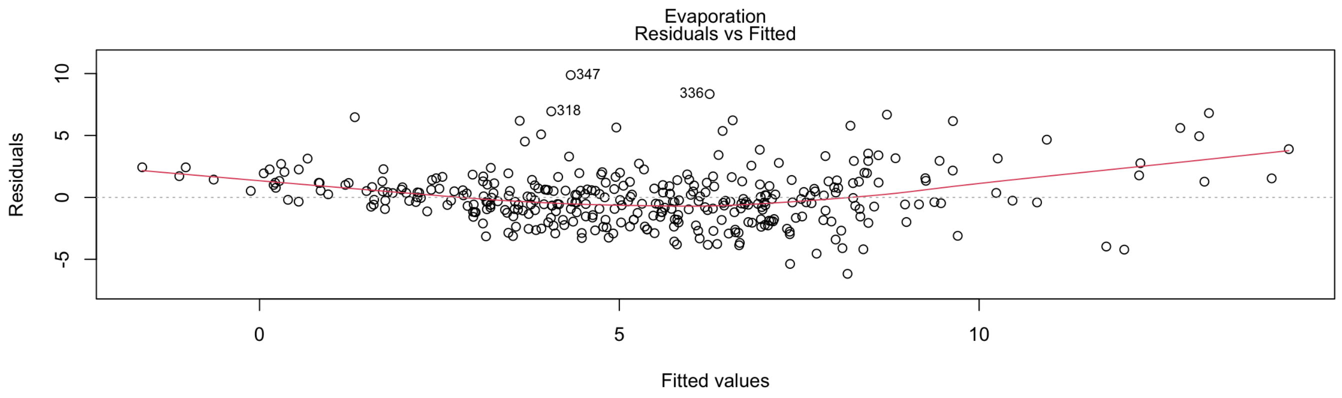
Linear Assumptions

Linear Relationship

```
line = 1
cex = 1
side = 3
adj=0.5

par(mfrow=c(2,1))
plot(step4_lm, which = 1, col = 'orange')
mtext("Transformed Evaporation", side=side, line=line, cex=cex, adj=adj)
plot(lm(`Evaporation (mm)` ~ month + `Minimum temperature (Deg C)`+
          `9am relative humidity (%)`,data = df), which = 1)
mtext("Evaporation", side=side, line=line, cex=cex, adj=adj)
```

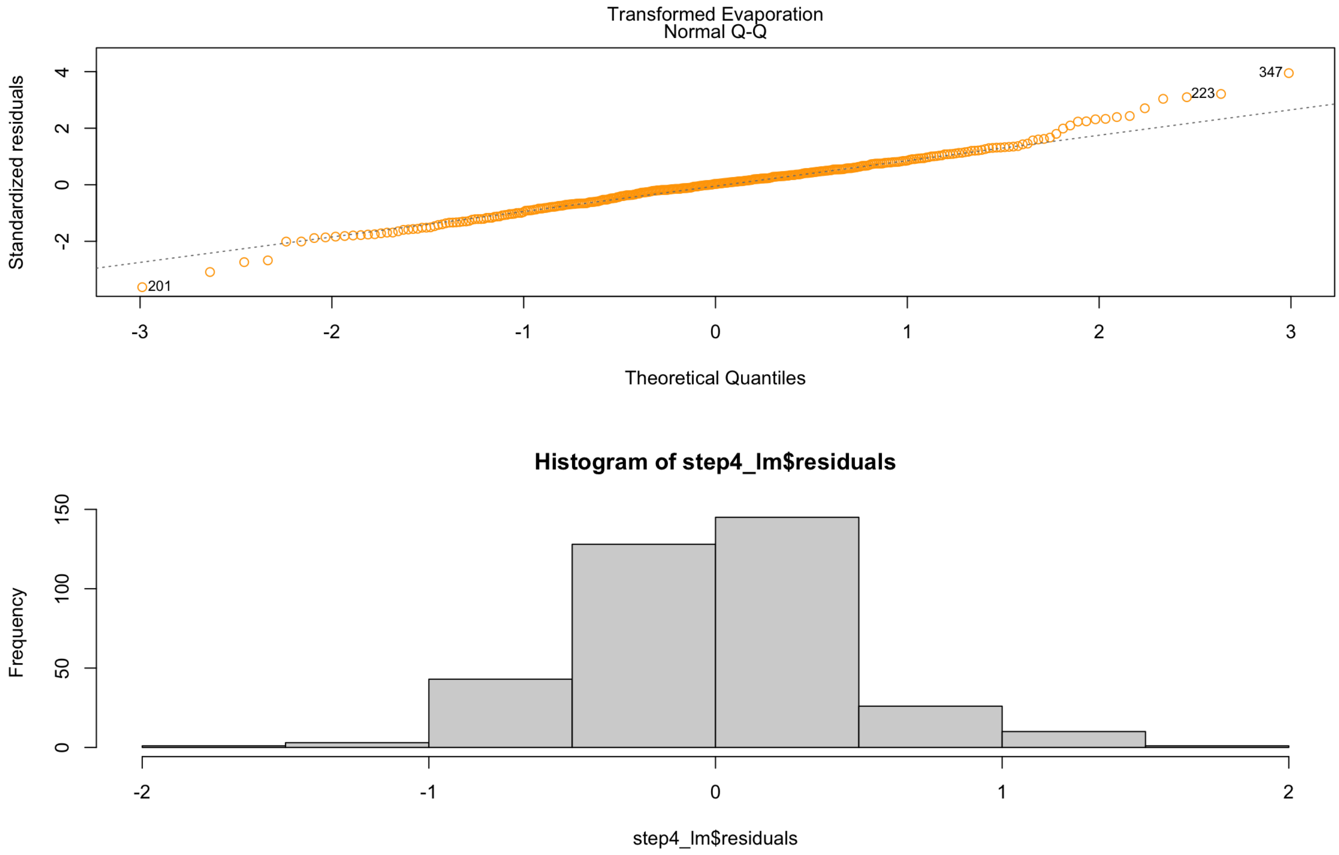




Our residual vs Fitted plot depicts a roughly straight red line about zero on the y-axis indicating the presence of a linear relationship. In comparison to the pre-transformed dependent variable which exhibited a quadratic relationship, modelling the relationship with evaporation on the square root scale has clearly flattened out the red line but still exhibits a slightly convex shape. We do see a slight deviation towards the tails of our plot which could be attributed to some highly influential outliers, but this is within an acceptable range and linearity seems to hold reasonably well. Therefore, the assumption of linearity is justified.

Noramallity in residuals (Normally Distributed)

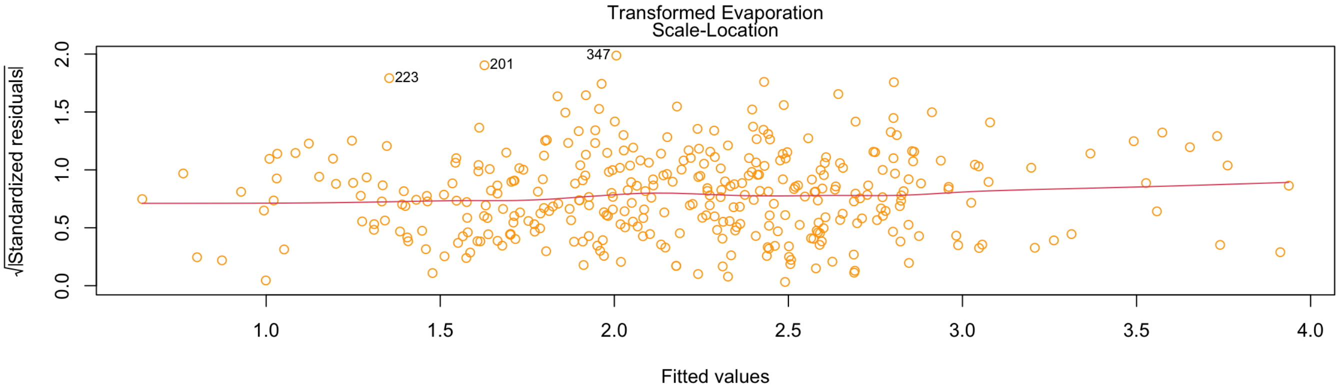
```
par(mfrow=c(2,1))
plot(step4_lm, which = 2, col = 'orange')
mtext("Transformed Evaporation", side=side, line=line, cex=cex, adj=adj)
hist(step4_lm$residuals)
```

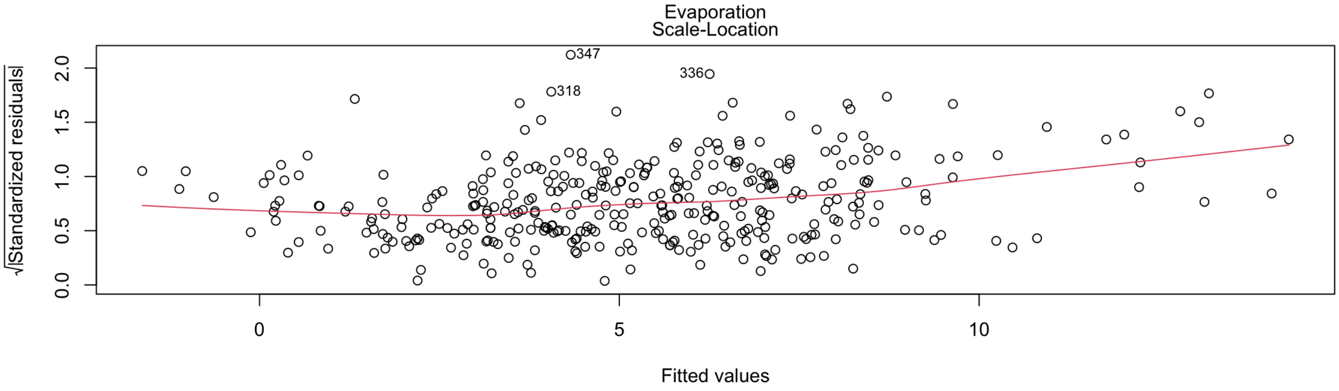


The normal quantile-quantile plot of the residuals and residuals histogram clearly lay along the line depicting the normality of the residuals. We can see slight divergence where the observations lay above 2 and below -2 vaguely following the similar shape to a heavy-skewed distribution (slight tailed depicted). However, the fact that points that lie less than -2 and greater than 2 on the x-axis drift away from the line is somewhat acceptable because it is minimal (barring three labelled points), and the fact that the bulk proportion of our observations lie on the line with minimal deviations allows us to conclude that the residuals are normally distributed.

Homoscedasticity

```
par(mfrow=c(2,1))
plot(step4_lm, which = 3, col = 'orange')
mtext("Transformed Evaporation", side=side, line=line, cex=cex, adj=adj)
plot(lm(`Evaporation (mm)` ~ month + `Minimum temperature (Deg C)`+
          `9am relative humidity (%)`,data = df), which = 3)
mtext("Evaporation", side=side, line=line, cex=cex, adj=adj)
```





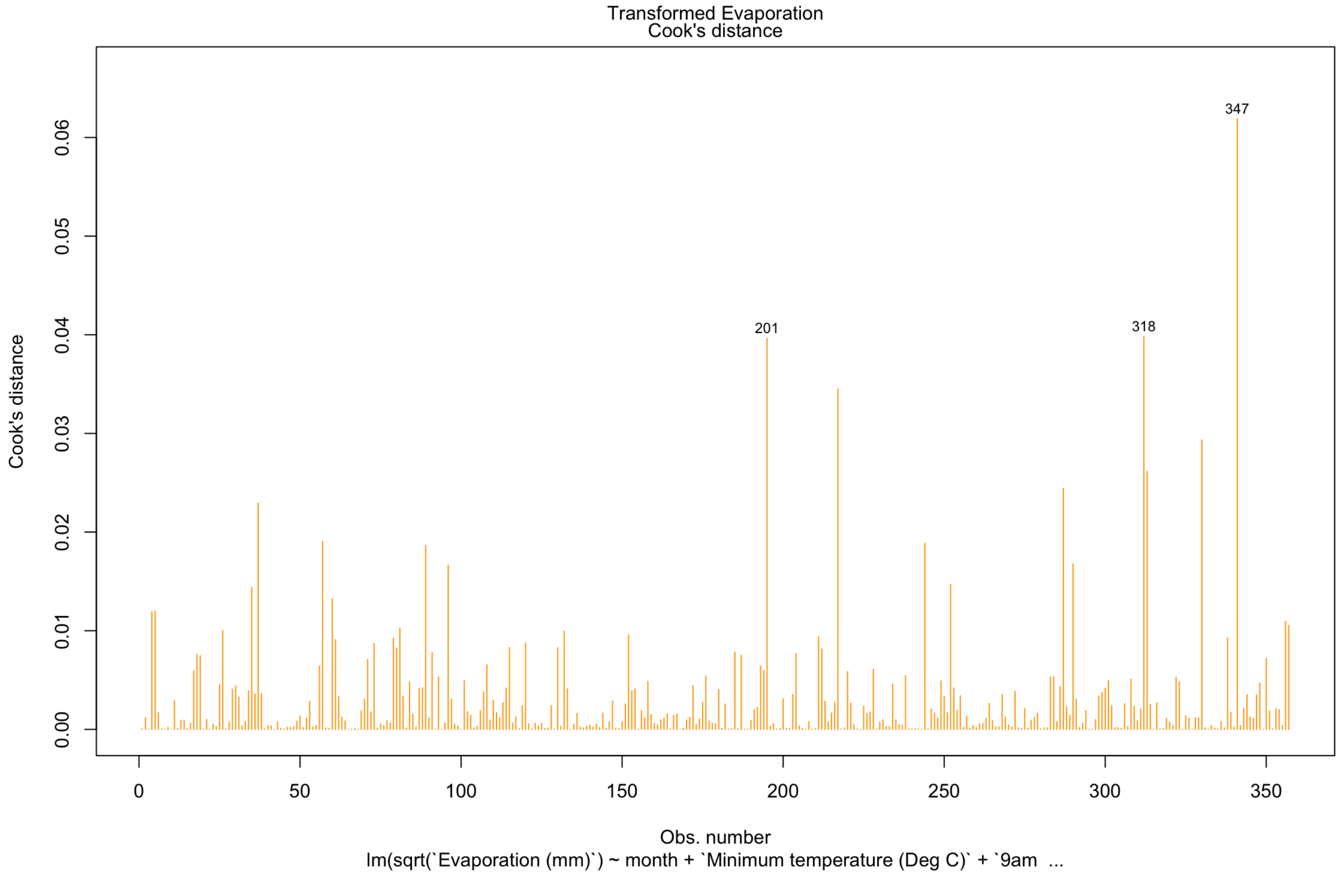
```
ncvTest(step4_lm)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.4206115, Df = 1, p = 0.51663
```

The pre-transformed evaporation scale-location plot clearly shows an upward trend as fitted values increase so too does spread in residuals. Notably, for our final model an argument can be made that there is no clear or strong trend in our data but as we move from left to right there seems to be a small increase in the spread of residuals as fitted values increase according to the red line. However, we don't see that obvious cone shape associated with heteroscedastic distributions and the observations seem to have a fairly symmetric spread. Additionally, using the ncvTest the corresponding p-value of 0.52 > 0.05, this indicates heteroskedasticity is not present. While this is a pretty borderline case, putting more weight towards the ncvTest as opposed to that slight increase in the red line we have sufficient evidence to say that the assumption of constant spread is justified.

Cooks Distance

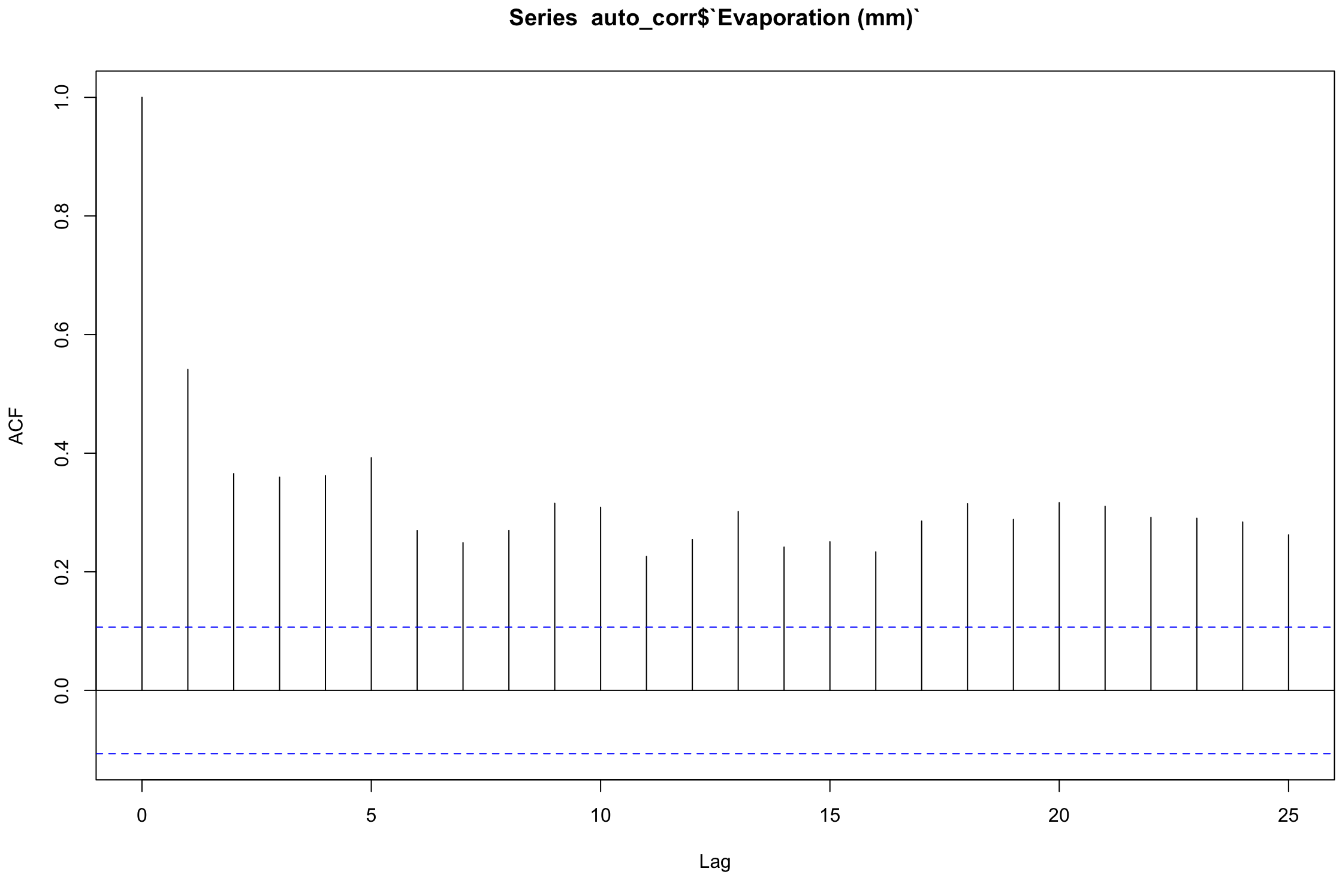
```
plot(step4_lm, which = 4, col = 'orange')
mtext("Transformed Evaporation", side=side, line=line, cex=cex, adj=adj)
```



Notably, data points 201 and 347 appear in every plot and according to Cook’s distance rule of thumb exceeding4/(n-p-1), these observations have high influence on the regression analysis specifically our graphs.

Autocorrelation

```
auto_corr <- df%>%
  na.omit()
acf(auto_corr$`Evaporation (mm)`)
```



```
durbinWatsonTest(step4_lm)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.04010166      1.908389   0.136
##  Alternative hypothesis: rho != 0
```

One key assumption in linear regression that is applicable to our dataset is that there is no correlation between the residuals (independent residuals). Notably, the Durban-Watson test is used to detect the presence of autocorrelation in the residuals of our regression. Since this p-value (0.146) > 0.05, we cannot reject the null hypothesis (no correlation among the residuals). Therefore, there is sufficient evidence that autocorrelation is not present.

Predictions

question 1

```
confidence_1 <- predict(step4_lm, newdata = tibble(month = '2',
                                  `9am relative humidity (%)`=74,
                                  `Minimum temperature (Deg C)` = 13.8),
                                   interval = 'confidence', level = 0.95)^2

prediction_1 <- predict(step4_lm, newdata = tibble(month = '2',
                                  `9am relative humidity (%)`=74,
                                  `Minimum temperature (Deg C)` = 13.8),
                                   interval = 'prediction', level = 0.95)^2
```

question 2

```
confidence_2 <- predict(step4_lm, newdata = tibble(month = '12',
                                   `9am relative humidity (%)`=57,
                                   `Minimum temperature (Deg C)` = 16.4),
                                    interval = 'confidence', level = 0.95)^2

prediction_2 <- predict(step4_lm, newdata = tibble(month = '12',
                                   `9am relative humidity (%)`=57,
                                   `Minimum temperature (Deg C)` = 16.4),
                                    interval = 'prediction', level = 0.95)^2
```

```
# question 3
confidence_3 <- predict(step4_lm, newdata = tibble(month = '1',
                                   `9am relative humidity (%)`=35,
                                   `Minimum temperature (Deg C)` = 26.5),
                                    interval = 'confidence', level = 0.95)^2

prediction_3 <- predict(step4_lm, newdata = tibble(month = '1',
                                   `9am relative humidity (%)`=35,
                                   `Minimum temperature (Deg C)` = 26.5),
                                    interval = 'prediction', level = 0.95)^2
```

```
# question 4
confidence_4 <- predict(step4_lm, newdata = tibble(month = '7',
                                   `9am relative humidity (%)`=76,
                                   `Minimum temperature (Deg C)` = 6.8),
                                    interval = 'confidence', level = 0.95)^2

prediction_4 <- predict(step4_lm, newdata = tibble(month = '7',
                                   `9am relative humidity (%)`=76,
                                   `Minimum temperature (Deg C)` = 6.8),
                                    interval = 'prediction', level = 0.95)^2
```

Confidence interval Table

```
confidence_interval <- tribble(
  ~Month,~`Min Temperature (Deg C)`, ~`9am relative Humidity (%)`, ~lower,~fit,~uppe
r,
  'February', 13.8, 74, confidence_1[2],confidence_1[1],confidence_1[3],
  'December', 16.4, 57, confidence_2[2],confidence_2[1],confidence_2[3],
  'January', 26.5, 35, confidence_3[2],confidence_3[1],confidence_3[3],
  'July', 6.8, 76, confidence_4[2],confidence_4[1],confidence_4[3]
)
confidence_interval
```

```
## # A tibble: 4 x 6
##   Month    `Min Temperature (Deg C… `9am relative Humidity (%… lower   fit upper
##   <chr>                       <dbl>                      <dbl> <dbl> <dbl> <dbl>
## 1 February                     13.8                         74  4.72  5.52  6.38
## 2 December                     16.4                         57  6.94  7.88  8.89
## 3 January                      26.5                         35 14.7  16.5  18.5 
## 4 July                          6.8                         76  1.50  1.98  2.52
```

Prediction interval Table

```
prediction_interval <- tribble(
  ~Month,~`Min Temperature (Deg C)`, ~`9am relative Humidity (%)`, ~lower,~fit,~uppe
r,
  'February', 13.8, 74, prediction_1[2],prediction_1[1],prediction_1[3],
  'December', 16.4, 57, prediction_2[2],prediction_2[1],prediction_2[3],
  'January', 26.5, 35, prediction_3[2],prediction_3[1],prediction_3[3],
  'July', 6.8, 76, prediction_4[2],prediction_4[1],prediction_4[3]
)
prediction_interval
```

```
## # A tibble: 4 x 6
##   Month    `Min Temperature (Deg C… `9am relative Humidity (%… lower   fit upper
##   <chr>                       <dbl>                      <dbl> <dbl> <dbl> <dbl>
## 1 February                     13.8                         74 2.05   5.52 10.7 
## 2 December                     16.4                         57 3.57   7.88 13.9 
## 3 January                      26.5                         35 9.81  16.5  25.0 
## 4 July                          6.8                         76 0.236  1.98  5.41
```