Final Project

PSTAT126: Regression Analysis, Winter 2025

true

2025-02-01

Due Date

The deadline for this step is: Friday, January 31 at 11:59 PM.

Formation of groups:

Sick Puppies

- Julianne Hussain (jhussain)
- Lily Yousefian (lilyyousefian)
- Andres Avelarr (andresavelar)
- Brian Ngan (Brianngan)

Tentative Role Distribution Table

Name	Responsibilities
Julianne Hussain	generating plots
Lily Yousefian	interpreting plots
Andres Avelar	majority of coding
Brian Ngan	organizing

Data Information

Data Name/Title: Algerian Forest Fires

Author/Owner: Abid, Faroudja

Date of Publication: 10/21/2019

Publication Venue: UCI Machine Learning Repository

Retrieval Date: 01/29/25

Link: https://archive.ics.uci.edu/dataset/547/algerian+forest+fires+dataset

Variable discription:

The dataset includes 244 instances that regroup a data of two regions of Algeria,namely the Bejaia region located in the northeast of Algeria and the Sidi Bel-abbes region located in the northwest of Algeria.

Name	Label	Type	Categories	No. of valid obs.
Variable 1	Day	numeric	Days of month 1-31	244
Variable 2	Month	numeric	Months, 6-9	244
Variable 3	Year	numeric	2012	244
Variable 4	Temperature	numeric	temperature noon (temperature max) in Celsius degrees: 22 to 42	244
Variable 5	Relative Humidity	numeric	Relative Humidity in %: 21 to 90	244
Variable 6	Wind Speed	numeric	Wind speed in km/h: 6 to 29	244
Variable 7	Rain	numeric	Total Day in mm: 0 to 16.8	244
Variable 8	FFMC	numeric	Fine Fuel Moisture Code (FFMC) index from the FWI system: 28.6 to 92.5	235
Variable 9	DMC	numeric	Duff Moisture Code (DMC) index from the FWI system: 1.1 to 65.9	241
Variable 10	DC	numeric	Drought Code (DC) index from the FWI system: 7 to 220.4	242
Variable 11	ISI	numeric	Initial Spread Index (ISI) index from the FWI system: 0 to 18.5	243
Variable 12	BUI	numeric	Buildup Index (BUI) index from the FWI system: 1.1 to 68	244
Variable 13	FWI	numeric	Fire Weather Index (FWI) Index: 0 to 31.1	243
Variable 14	Classes	factor	Two classes, Fire & Not Fire	243

```r
#optimized
data <- read.csv("Algerian_forest_fires_dataset_UPDATE.csv", header = FALSE, skip = 2)
data_clean <- data[!grepl("Sidi-Bel Abbes Region Dataset", data$V1), ]
header_rows <- grep("day", data_clean$V1, ignore.case = TRUE)
if (length(header_rows) > 0) data_clean <- data_clean[-header_rows, ]
colnames(data_clean) <- c("day", "month", "year", "Temperature", "RH", "Ws", "Rain",
                          "FFMC", "DMC", "DC", "ISI", "BUI", "FWI", "Classes")
data_clean <- transform(data_clean,
                        day = as.numeric(day),
                        month = as.numeric(month),
                        year = as.numeric(year),
                        Temperature = as.numeric(Temperature),
                        RH = as.numeric(RH),
                        Ws = as.numeric(Ws),
                        Rain = as.numeric(Rain),
                        FFMC = as.numeric(FFMC),
                        DMC = as.numeric(DMC),
                        DC = as.numeric(DC),
                        ISI = as.numeric(ISI),
                        BUI = as.numeric(BUI),
                        FWI = as.numeric(FWI),
                        Classes = trimws(tolower(Classes)))

# Validation criteria
criteria <- list(
  day = c(1, 31),
  month = c(6, 9),
  year = c(2012, 2012),
  Temperature = c(22, 42),
  RH = c(21, 90),
  Ws = c(6, 29),
  Rain = c(0, 16.8),
  FFMC = c(28.6, 92.5),
  DMC = c(1.1, 65.9),
  DC = c(7, 220.4),
  ISI = c(0, 18.5),
  BUI = c(1.1, 68),
  FWI = c(0, 31.1)
)
valid_classes <- c("fire", "not fire")

# Function to validate variables
validate_variable <- function(variable, range = NULL, categories = NULL) {
  if (!is.null(range)) {
    return(sum(!is.na(variable) & variable >= range[1] & variable <= range[2], na.rm = TRUE))
  } else if (!is.null(categories)) {
    return(sum(!is.na(variable) & variable %in% categories, na.rm = TRUE))
  }
}

# Validate all variables
valid_obs <- sapply(names(criteria), function(var) validate_variable(data_clean[[var]], range = criteri
valid_obs["Classes"] <- validate_variable(data_clean$Classes, categories = valid_classes)

# Create the results table
num_valid_obs <- data.frame(Variable = c(names(criteria), "Classes"), Valid_Observations = valid_obs)

# Print the results
print(num_valid_obs)
```

```
##                Variable Valid_Observations
## day                 day                244
## month             month                244
## year               year                244
## Temperature Temperature                244
## RH                   RH                244
## Ws                          Ws                  244
## Rain                      Rain                  244
## FFMC                      FFMC                  235
## DMC                        DMC                  241
## DC                          DC                  242
## ISI                        ISI                  243
## BUI                        BUI                  244
## FWI                        FWI                  243
## Classes                Classes                  243
```

```
data_clean <- subset(data_clean, Classes == "fire")

# Check the number of observations in the filtered dataset
cat("Number of fire observations:", nrow(data_clean), "\n")
```

```
## Number of fire observations: 137
```

Initial Insights:

We can see that the data was messy and required some cleaning, there are also some variables with less than 244 valid observations. We will most likely have to filter out the classes column to have more accurate results.

Research Questions:

How does Temperature affect the Fire Weather Index (FWI) during fire events?

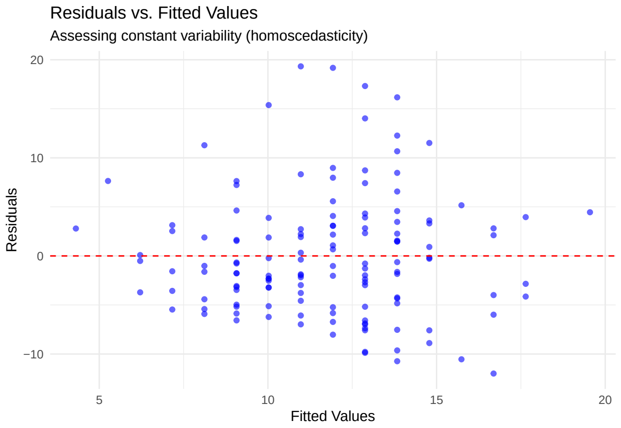
How do additional weather factors such as Relative Humidity (RH), Rain, and Wind Speed (Ws) affect FWI during fire events?

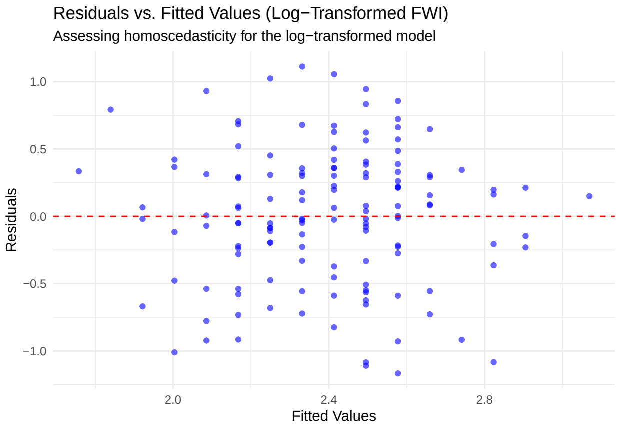
Do the Codes and Indices (FFMC, DMS, DC, ISI, BUI) significantly predict FWI during fire events?

Hypothesis:

Null Hypothesis: Temperature does not have an effect on FWI during fire events.

Alternative Hypothesis: Higher Temperature is positively associated with FWI during fire events.



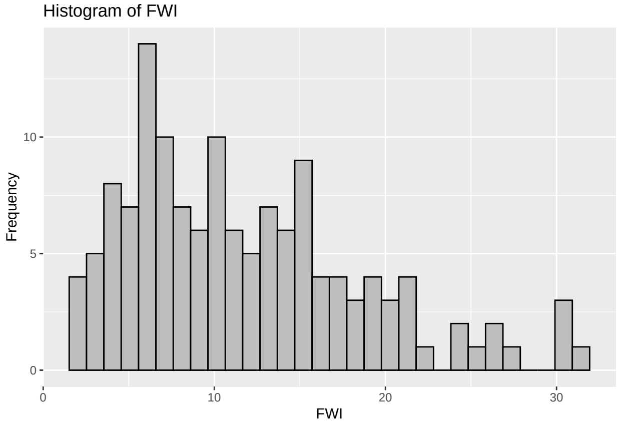


We log-transformed our simple linear regression model because we needed to stabilize the variance.

Categorical Variable:

Our response variable is Fire Weather Index (FWI) which we chose to split into thirds to preserve the distribution.

```
##        0% 33.33333% 66.66667%      100%
##   1.70000   7.30000  13.83333  31.10000
```



```
##       day            month            year       Temperature         RH
##  Min.   : 2.00   Min.   :6.000   Min.   :2012   Min.   :26.0   Min.   :21.00
##  1st Qu.:10.00   1st Qu.:7.000   1st Qu.:2012   1st Qu.:32.0   1st Qu.:45.00
##  Median :18.00   Median :8.000   Median :2012   Median :34.0   Median :56.00
##  Mean   :17.34   Mean   :7.526   Mean   :2012   Mean   :33.8   Mean   :56.42
##  3rd Qu.:24.00   3rd Qu.:8.000   3rd Qu.:2012   3rd Qu.:36.0   3rd Qu.:66.00
##  Max.   :31.00   Max.   :9.000   Max.   :2012   Max.   :42.0   Max.   :88.00
##        Ws             Rain              FFMC            DMC
##  Min.   : 8.00   Min.   :0.00000   Min.   :80.20   Min.   : 3.40
##  1st Qu.:14.00   1st Qu.:0.00000   1st Qu.:85.30   1st Qu.:12.10
##  Median :15.00   Median :0.00000   Median :87.80   Median :18.00
##  Mean   :15.32   Mean   :0.09635   Mean   :87.54   Mean   :21.05
##  3rd Qu.:17.00   3rd Qu.:0.00000   3rd Qu.:89.40   3rd Qu.:25.80
##  Max.   :21.00   Max.   :6.00000   Max.   :96.00   Max.   :65.90
##        DC              ISI              BUI             FWI
##  Min.   :  9.70   Min.   : 2.600   Min.   : 5.10   Min.   : 1.70
##  1st Qu.: 34.10   1st Qu.: 4.700   1st Qu.:13.70   1st Qu.: 6.30
##  Median : 54.20   Median : 6.800   Median :19.20   Median :10.50
##  Mean   : 70.82   Mean   : 7.423   Mean   :24.02   Mean   :11.73
##  3rd Qu.: 96.80   3rd Qu.: 9.200   3rd Qu.:30.40   3rd Qu.:15.70
##  Max.   :220.40   Max.   :19.000   Max.   :68.00   Max.   :31.10
##    Classes          FWI_category
##  Length:137         Low   :47
##  Class :character   Medium:44
##  Mode  :character   High  :46
##
##
##
```

Table 3: Skewness and Kurtosis of Weather Variables

Variable	Skewness	Kurtosis
Temperature	-0.0278237	2.945840
RH	-0.1019675	2.485445
Ws	-0.2187552	3.237929
Rain	9.7025284	103.439977
FFMC	0.0003518	2.623560
DMC	1.3014849	4.487733
DC	1.0848362	3.378900
ISI	1.0367894	3.626574
BUI	1.1877341	3.901933
FWI	0.8323184	3.160659

Because Temperature is skewed left with thin tails, we know that there is generally not significant outliers in the data, meaning there is not a high occurrence of extremes in temperature.The left skew tells us that most of the temperatures collected were higher, with low temperatures happening very rarely.

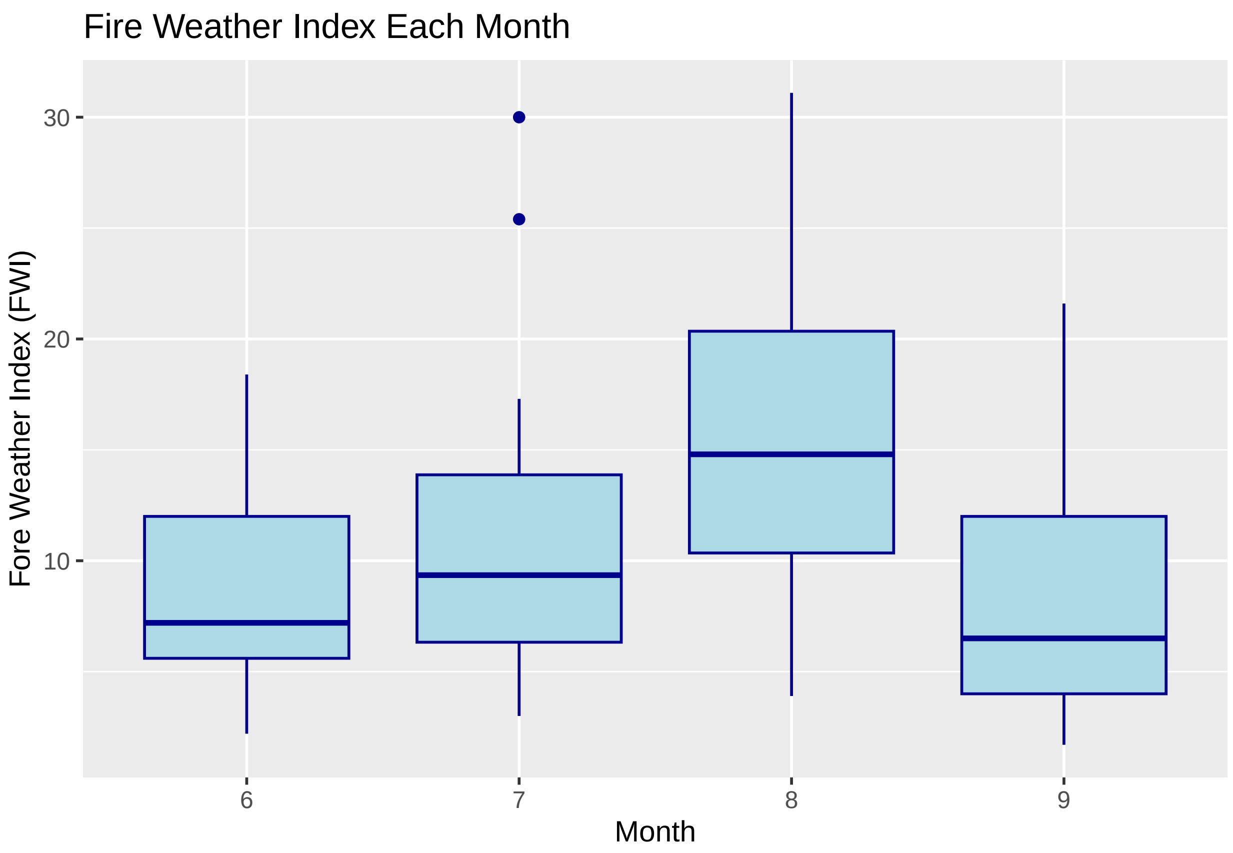
Because relative humidity is skewed left with thin tails, we assume the same about this variable. It generally does not have a lot of extremes, and most of the data is centered around higher humidity, with a very small portion of the data being lower humidity.

Wind speed is left skewed with thin tails, which tells us that there are mostly low wind speeds. Occasionally there are higher wind speeds, but they don't occur very frequently.

Rain is right skewed and has significantly thick tails. This tells us that even though this region receives little to no rain, when rain occurs it is in small amounts with very few, if any, outliers.

We can use the same line of reasoning to interpret the remaining variables.

```
## [1] 6 7 8 9
```

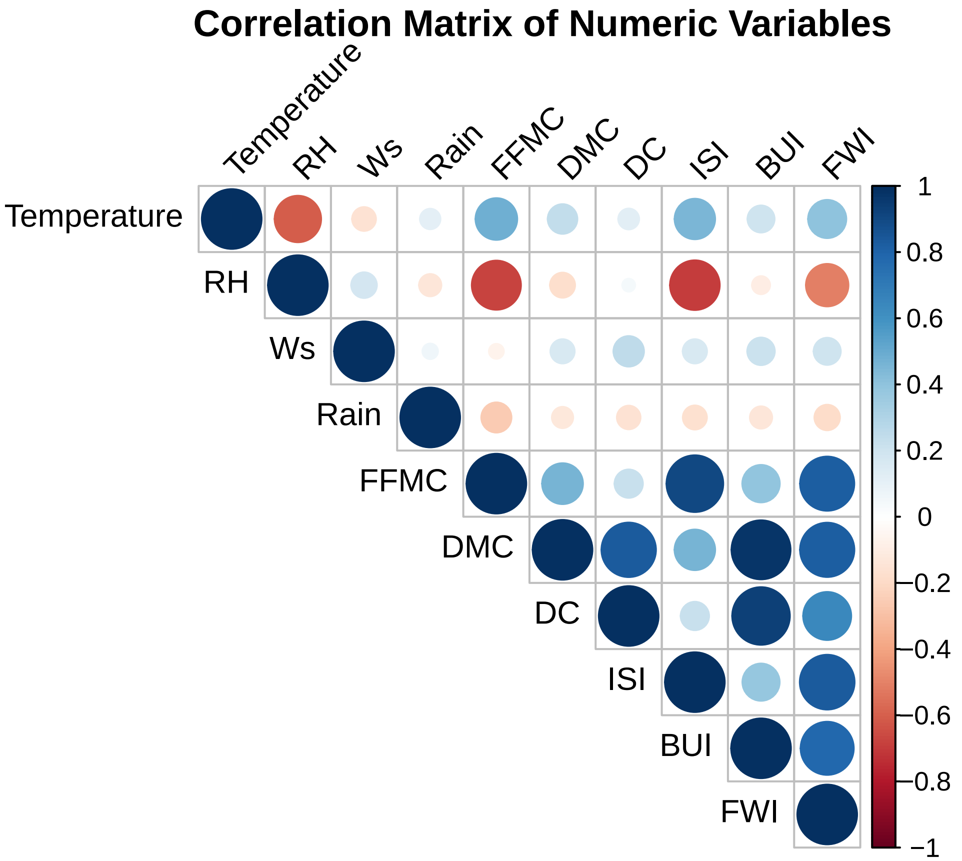


Month 6 (June): Shows the lowest FWI distribution overall, with a relatively low median and minimal spread

Month 7 (July): Has a moderate median FWI but shows two high outliers, indicating that while most FWI values remain in a moderate range, some instances reach much higher levels

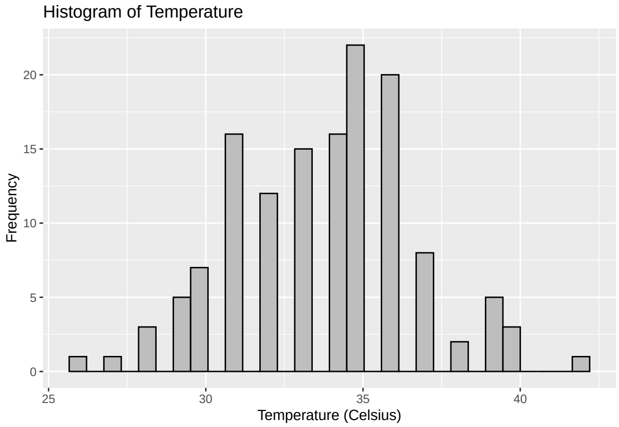
Month 8 (August): Shows the highest median FWI and a wider spread, suggesting that fire risk peaks and becomes more variable during this month

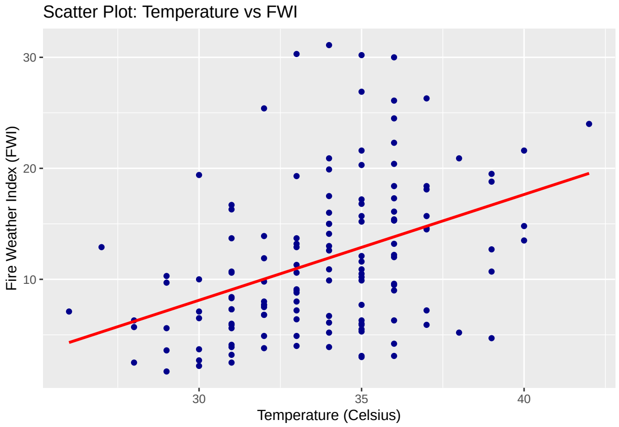
Month 9 (September): Has a median FWI lower than August's but still above June's, with a broad range indicating variability in fire risk late in the season



We chose to partition the variables into two categories: Weather factors: Temperature, RH, Ws, and Rain, additionally month as fitting. Codes and Indices: FFMC, DMC, DC, ISI, BUI Response variable: FWI

Histogram and Scatter Plot:

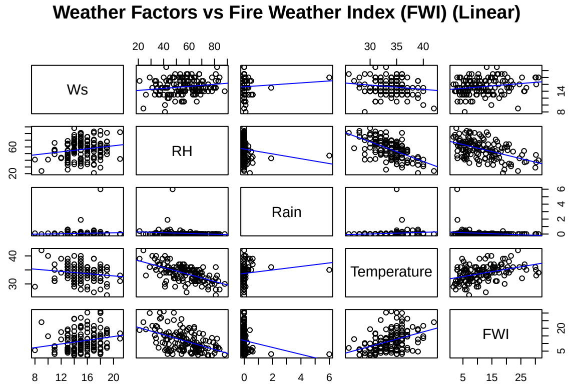


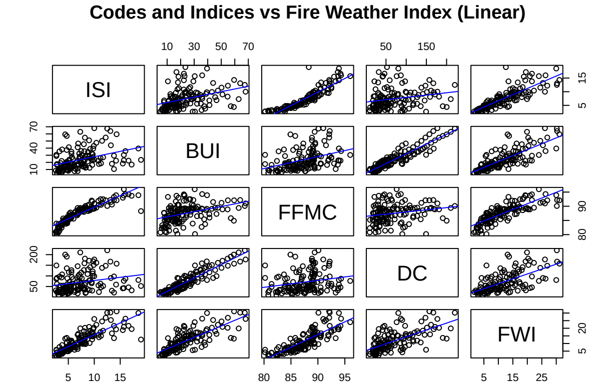


Histogram: We can see that the distribution is slightly right-skewed, with most observations being around 303 - 35s. Although a few temperatures reach nearly highs of 40s and lows 25s, over all indicating generally warm conditions.

Scatter plot: A clear positive trend shows that higher temperatures are associated with an increased Fire Weather Index, suggesting that heat elevates fire risk. However, the spread of points indicates that even at similar temperatures, FWI can vary, implying that other factors also influence fire conditions.

Pair Plots:





WS vs FWI: The linear trend shows a positive relationship, meaning that when WS increases, FWI tends to increase. Therefore the more wind speed we have the more our fire weather index is.

RH vs FWI: The plot shows a negative slope meaning the moisture helps reduce the risk of fire.

Rain vs FWI: The linear line shows a negative relationship, meaning the more rainfall we have a lower potential for fire.

Temperature vs FWI: The linear relationship is clearly positive meaning that higher temperatures tend to increase the risk of fire spread.

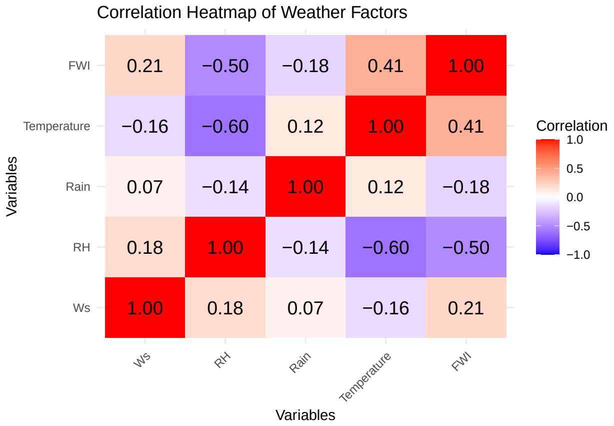
ISI vs FWI: It shows a strong positive linear trend, ISI increases FWI also steadily rises.

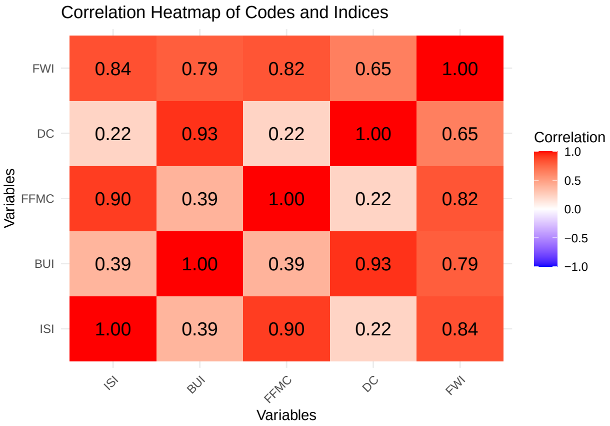
BUI vs FWI: It shows a positive slope, meaning a higher BUI value relates to a higher FWI level.

FFMC vs FWI: It shows an upward trend, meaning that when FFMC increases, FWI increases as well.

DC vs FWI: It's showing a strong and positive relation, meaning higher DC values correspond with higher FWI levels.

Correlation Heatmaps





For the correlation of weather factors: FWI is positively correlated with Temperature (0.57) but negatively correlated with both Rain (-0.57) and Relative Humidity (RH) (-0.58) Wind Speed (Ws) shows weak correlations with the other variables (e.g., only 0.03 with FWI) Temperature is strongly negatively correlated with RH (-0.65) and moderately negatively correlated with Rain (-0.32) RH and Rain each show moderate negative correlations with FWI, reflecting how moisture factors help lower fire potential

For the correlation of codes and Indices: All codes (ISI, BUI, FFMC, DC) show positive correlations with FWI, confirming that they each contribute to the overall fire risk measure ISI has the strongest correlation with FWI (0.92), followed by BUI (0.86), DC (0.74), and FFMC (0.69) Among the codes themselves, BUI and DC exhibit a very strong correlation (0.91), while ISI is moderately correlated with both BUI (0.64) and DC (0.70)

Hypotheses for Multiple Linear Regression

Null Hypothesis: None of the weather factors have a significant effect on FWI during fire events. Alternative Hypothesis: At least one of the weather factors has a significant effect on FWI during fire events.

Regression model application

```
##
## Call:
## lm(formula = FWI ~ Temperature + Ws + RH + Rain + month, data = data_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8190  -3.0652  -0.1958   2.7043  11.9525
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.320373  10.263763   0.908    0.366
## Temperature  0.007681   0.232980   0.033    0.974
## Ws           0.922319   0.199518   4.623 9.07e-06 ***
## RH          -0.258819   0.038460  -6.730 5.02e-10 ***
## Rain        -3.353467   0.785575  -4.269 3.78e-05 ***
## month7       2.095813   1.303898   1.607    0.110
## month8       5.614336   1.342133   4.183 5.28e-05 ***
## month9       1.639258   1.477401   1.110    0.269
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.829 on 129 degrees of freedom
## Multiple R-squared:  0.5251, Adjusted R-squared:  0.4993
## F-statistic: 20.37 on 7 and 129 DF,  p-value: < 2.2e-16
```

The adjusted r-squared value of 0.4993 means that approximately 49.93% of the variance in the model is explained by these predictors, which is not significant enough to deem this a reliable model. The p-value is less than 2.2e-16 which is significantly less than the 0.05 level of significance. Therefore we can reject our null hypothesis and say that Based on our results from the correlation matrix, our box plot, and scatter plot, we chose to include ISI, DMC, BUI, DC, FFMC, Temperature, Ws, and month in our multiple linear regression model.

Applying model selection technique

Null Hypothesis: The Codes and Indices do not have a significant relationship with FWI during fire events. Alternative Hypothesis: At least one of the fire indices has a significant relationship with FWI during fire events.

First we define a full model to run stepwise optimization. We used forward and backward stepwise selection to identify the most important predictors. Based on our results from the correlation matrix, our box plot, and scatter plot, we chose to include ISI, DMC, BUI, DC, FFMC, Temperature, Ws, and month in our multiple linear regression model.

```
##
## Call:
## lm(formula = FWI ~ ISI + DMC + BUI + DC + FFMC + RH + Temperature +
##     Rain + Ws, data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3756 -0.3798 -0.0190  0.4934  2.7670
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -52.85891    7.70129  -6.864 2.65e-10 ***
## ISI           0.62569    0.08261   7.574 6.56e-12 ***
## DMC           0.21439    0.13041   1.644  0.10266
## BUI          -0.02582    0.17252  -0.150  0.88126
## DC            0.03030    0.02003   1.513  0.13286
## FFMC          0.60993    0.08723   6.993 1.37e-10 ***
## RH           -0.01607    0.01338  -1.201  0.23200
## Temperature  -0.04064    0.04545  -0.894  0.37289
## Rain          0.22761    0.22787   0.999  0.31976
## Ws            0.18114    0.05859   3.092  0.00244 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.19 on 127 degrees of freedom
## Multiple R-squared:  0.9716, Adjusted R-squared:  0.9696
## F-statistic: 483.2 on 9 and 127 DF,  p-value: < 2.2e-16
```

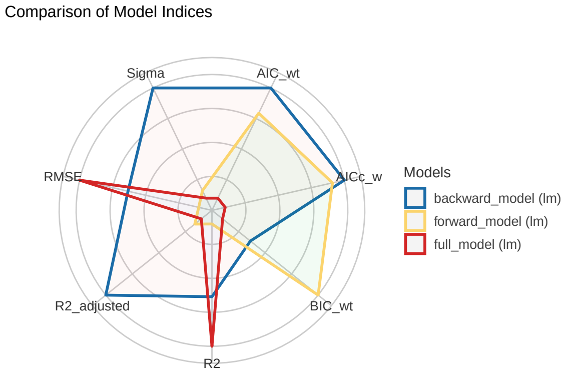
```
## Start:  AIC=527.22
## FWI ~ 1
##
##               Df Sum of Sq    RSS    AIC
## + ISI          1    4467.6 1866.3 361.81
## + DMC          1    4316.9 2017.0 372.45
## + FFMC         1    4275.7 2058.2 375.22
## + BUI          1    3948.1 2385.8 395.45
## + DC           1    2639.5 3694.4 455.36
## + RH           1    1611.9 4722.0 488.98
## + Temperature  1    1054.6 5279.3 504.27
## + Ws           1     266.5 6067.4 523.33
## + Rain         1     207.5 6126.4 524.65
## <none>                     6333.9 527.22
##
```

```
## Step:  AIC=361.81
## FWI ~ ISI
##
##               Df Sum of Sq     RSS    AIC
## + BUI          1   1608.93  257.41  92.41
## + DMC          1   1548.70  317.64 121.21
## + DC           1   1393.28  473.06 175.78
## + FFMC         1    141.64 1724.70 353.00
## + RH           1     68.70 1797.64 358.67
## + Ws           1     29.81 1836.53 361.60
## <none>                     1866.34 361.81
## + Rain         1     13.58 1852.76 362.81
## + Temperature  1      4.12 1862.22 363.51
##
## Step:  AIC=92.41
## FWI ~ ISI + BUI
##
##               Df Sum of Sq    RSS    AIC
## + FFMC         1    57.008 200.41 60.110
## + RH           1     5.811 251.60 91.279
## <none>                     257.41 92.407
## + DMC          1     1.501 255.91 93.605
## + Ws           1     1.168 256.25 93.783
## + DC           1     0.925 256.49 93.913
## + Temperature  1     0.538 256.88 94.120
## + Rain         1     0.131 257.28 94.337
##
## Step:  AIC=60.11
## FWI ~ ISI + BUI + FFMC
##
##               Df Sum of Sq    RSS    AIC
## + Ws           1   10.7912 189.62 54.527
## <none>                     200.41 60.110
## + Rain         1    2.3544 198.05 60.491
## + RH           1    0.7352 199.67 61.606
## + Temperature  1    0.4727 199.93 61.786
## + DMC          1    0.0783 200.33 62.056
## + DC           1    0.0370 200.37 62.085
##
## Step:  AIC=54.53
## FWI ~ ISI + BUI + FFMC + Ws
##
##               Df Sum of Sq    RSS    AIC
## + RH           1    4.3826 185.23 53.323
## <none>                     189.62 54.527
## + Rain         1    2.5486 187.07 54.673
## + DMC          1    1.1376 188.48 55.702
## + DC           1    0.1706 189.44 56.403
## + Temperature  1    0.0001 189.61 56.527
##
## Step:  AIC=53.32
## FWI ~ ISI + BUI + FFMC + Ws + RH
##
##               Df Sum of Sq    RSS    AIC
## <none>                     185.23 53.323
## + Temperature  1   0.87743 184.35 54.673
## + DMC          1   0.87501 184.36 54.674
## + Rain         1   0.58912 184.64 54.887
## + DC           1   0.13823 185.09 55.221
```

```
##
## Call:
## lm(formula = FWI ~ ISI + BUI + FFMC + Ws + RH, data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4356 -0.3242 -0.0480  0.5725  2.6663
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -50.079974   7.141442  -7.013 1.12e-10 ***
## ISI           0.640611   0.081212   7.888 1.06e-12 ***
## BUI           0.251554   0.008179  30.754  < 2e-16 ***
## FFMC          0.563957   0.082328   6.850 2.60e-10 ***
## Ws            0.180008   0.056331   3.196  0.00175 **
## RH           -0.019644   0.011158  -1.761  0.08065 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.189 on 131 degrees of freedom
## Multiple R-squared:  0.9708, Adjusted R-squared:  0.9696
## F-statistic: 869.7 on 5 and 131 DF,  p-value: < 2.2e-16
```

```
## Start:  AIC=57.18
## FWI ~ ISI + DMC + BUI + DC + FFMC + RH + Temperature + Rain +
##     Ws
##
##               Df Sum of Sq    RSS     AIC
## - BUI          1     0.032 179.74  55.201
## - Temperature  1     1.132 180.84  56.037
## - Rain         1     1.412 181.12  56.249
## - RH           1     2.041 181.75  56.724
## <none>                     179.71  57.177
## - DC           1     3.238 182.95  57.623
## - DMC          1     3.824 183.53  58.062
## - Ws           1    13.528 193.24  65.120
## - FFMC         1    69.189 248.90  99.798
## - ISI          1    81.172 260.88 106.240
##
## Step:  AIC=55.2
## FWI ~ ISI + DMC + DC + FFMC + RH + Temperature + Rain + Ws
##
##               Df Sum of Sq    RSS     AIC
## - Temperature  1     1.124 180.87  54.055
## - Rain         1     1.385 181.13  54.253
## - RH           1     2.232 181.97  54.892
## <none>                     179.74  55.201
## - Ws           1    13.508 193.25  63.129
## - DC           1    59.716 239.46  92.501
## - FFMC         1    71.665 251.41  99.172
## - ISI          1    81.986 261.73 104.683
## - DMC          1   174.098 353.84 145.994
##
## Step:  AIC=54.06
## FWI ~ ISI + DMC + DC + FFMC + RH + Rain + Ws
##
##        Df Sum of Sq    RSS     AIC
## - Rain  1     1.081 181.95  52.871
## - RH    1     1.462 182.33  53.159
## <none>              180.87  54.055
## - Ws    1    14.467 195.33  62.597
## - DC    1    58.634 239.50  90.525
## - FFMC  1    70.621 251.49  97.215
## - ISI   1    82.289 263.15 103.428
## - DMC   1   174.527 355.39 144.594
##
## Step:  AIC=52.87
## FWI ~ ISI + DMC + DC + FFMC + RH + Ws
##
##        Df Sum of Sq    RSS     AIC
## <none>              181.95  52.871
## - RH    1     3.370 185.32  53.386
## - Ws    1    16.331 198.28  62.647
## - DC    1    57.568 239.51  88.533
## - FFMC  1    70.761 252.71  95.878
## - ISI   1    81.383 263.33 101.519
## - DMC   1   184.227 366.17 146.688
```

```
##
## Call:
## lm(formula = FWI ~ ISI + DMC + DC + FFMC + RH + Ws, data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3833 -0.3825 -0.0279  0.5512  2.7458
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -51.842998   7.091618  -7.310 2.42e-11 ***
## ISI           0.622847   0.081680   7.625 4.53e-12 ***
## DMC           0.197826   0.017243  11.473  < 2e-16 ***
## DC            0.026404   0.004117   6.413 2.42e-09 ***
## FFMC          0.581590   0.081794   7.110 6.90e-11 ***
## RH           -0.017213   0.011093  -1.552 0.123156
## Ws            0.194465   0.056929   3.416 0.000849 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.183 on 130 degrees of freedom
## Multiple R-squared:  0.9713, Adjusted R-squared:  0.9699
## F-statistic: 732.6 on 6 and 130 DF,  p-value: < 2.2e-16
## # Comparison of Model Performance Indices
##
## Name           | Model | AIC (weights) | AICc (weights) | BIC (weights) |    R2
## ---------------------------------------------------------------------------------
## forward_model  |    lm | 444.1 (0.417) |  445.0 (0.459) | 464.6 (0.774) | 0.971
## backward_model |    lm | 443.7 (0.522) |  444.8 (0.506) | 467.0 (0.225) | 0.971
## full_model     |    lm | 448.0 (0.061) |  450.1 (0.036) | 480.1 (<.001) | 0.972
##
## Name           | R2 (adj.) |  RMSE | Sigma
## ------------------------------------------
## forward_model  |     0.970 | 1.163 | 1.189
## backward_model |     0.970 | 1.152 | 1.183
## full_model     |     0.970 | 1.145 | 1.190
```

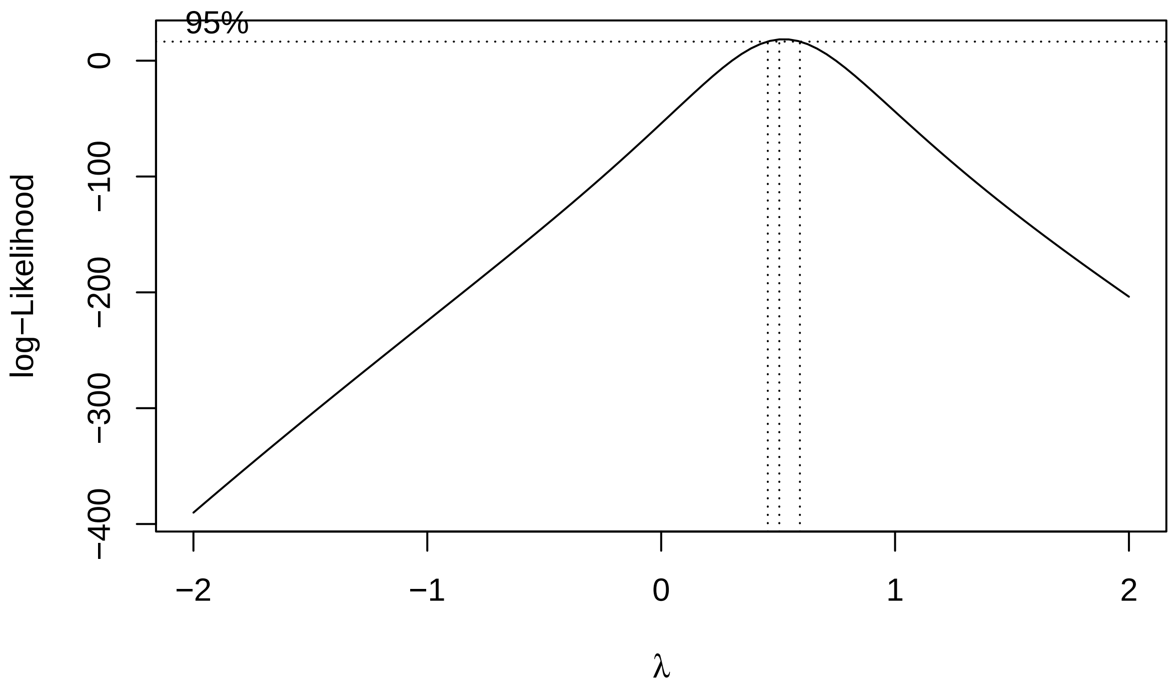


According to the above performance model, the most reliable multiple linear regression model can be derived from the backward model. This gives us $Y = -51.843 + 0.623ISI + 0.198DMC + 0.026DC + 0.582FFMC - 0.017RH + 0.194Ws + \epsilon$. This means that the most correlated variables to a high FWI during fire events are ISI, DMC, DC, FFMC, RH and Ws.

```
##
## Call:
## lm(formula = FWI ~ ISI + DMC + DC + FFMC + RH + Ws, data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3833 -0.3825 -0.0279  0.5512  2.7458
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -51.842998   7.091618  -7.310 2.42e-11 ***
## ISI           0.622847   0.081680   7.625 4.53e-12 ***
## DMC           0.197826   0.017243  11.473  < 2e-16 ***
## DC            0.026404   0.004117   6.413 2.42e-09 ***
## FFMC          0.581590   0.081794   7.110 6.90e-11 ***
## RH           -0.017213   0.011093  -1.552 0.123156
## Ws            0.194465   0.056929   3.416 0.000849 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

```
## Residual standard error: 1.183 on 130 degrees of freedom
## Multiple R-squared:  0.9713, Adjusted R-squared:  0.9699
## F-statistic: 732.6 on 6 and 130 DF,  p-value: < 2.2e-16
```

```
# Backward model with log transformation on FWI (adding 1 to avoid log(0))
model_log_backward <- lm(log(FWI + 1) ~ ISI + DMC + DC + FFMC + RH + Ws, data = data_clean)
model_diag_log_backward <- data.frame(
  fitted = fitted(model_log_backward),
  residuals = resid(model_log_backward)
)

# Residual plot for the log-transformed model
p_log <- ggplot(model_diag_log_backward, aes(x = fitted, y = residuals)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(
    title = "Residuals vs. Fitted (Log-Transformed Model)",
    x = "Fitted Values",
    y = "Residuals"
  ) +
  theme_minimal()

# Fit the backward model with the original (non-transformed) FWI
model_backward <- lm(FWI ~ ISI + DMC + DC + FFMC + RH + Ws, data = data_clean)
model_diag_backward <- data.frame(
  fitted = fitted(model_backward),
  residuals = resid(model_backward)
)

# 2. Use Box-Cox to determine the optimal lambda for the backward model
# Note: FWI must be strictly positive. If FWI contains 0, consider adding a small constant.
bc <- boxcox(model_backward, plotit = TRUE)
```

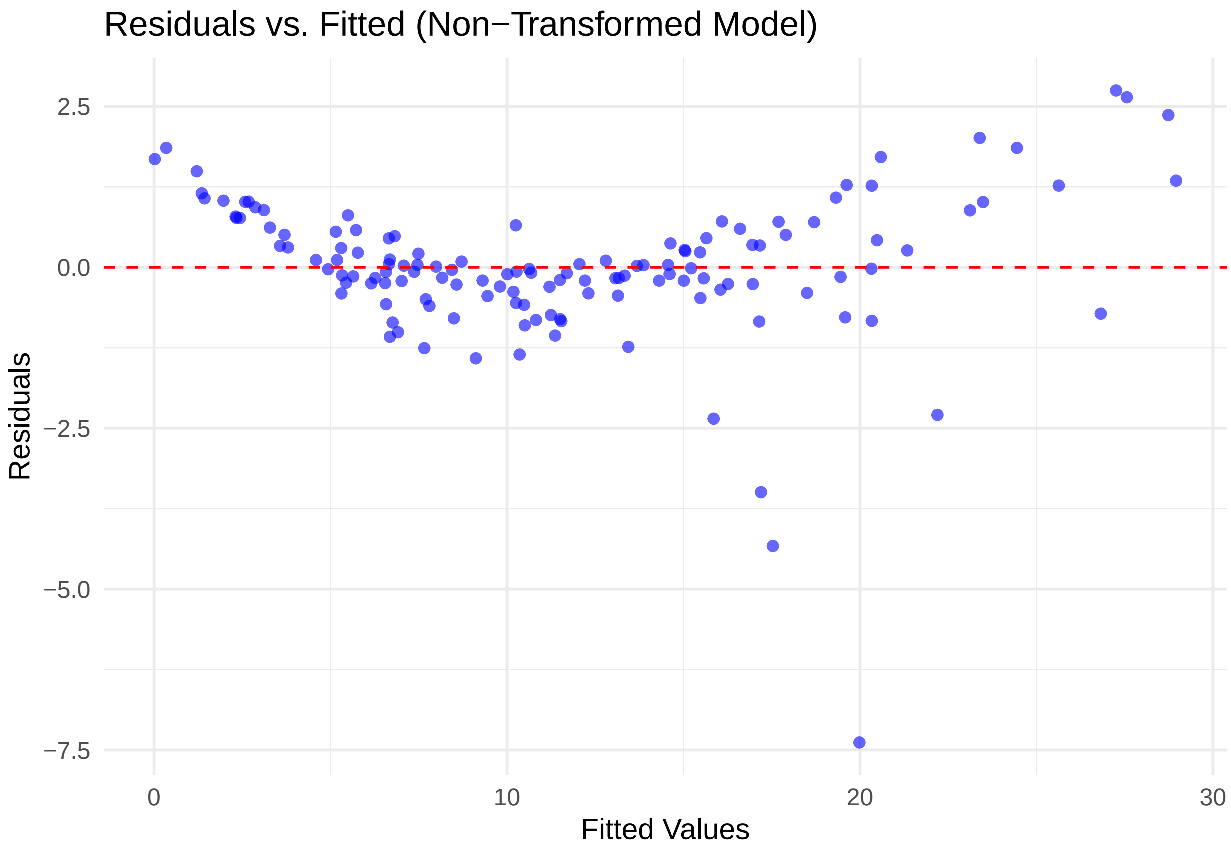


```
lambda_opt <- bc$x[which.max(bc$y)]
cat("Optimal lambda from Box-Cox:", lambda_opt, "\n")
```

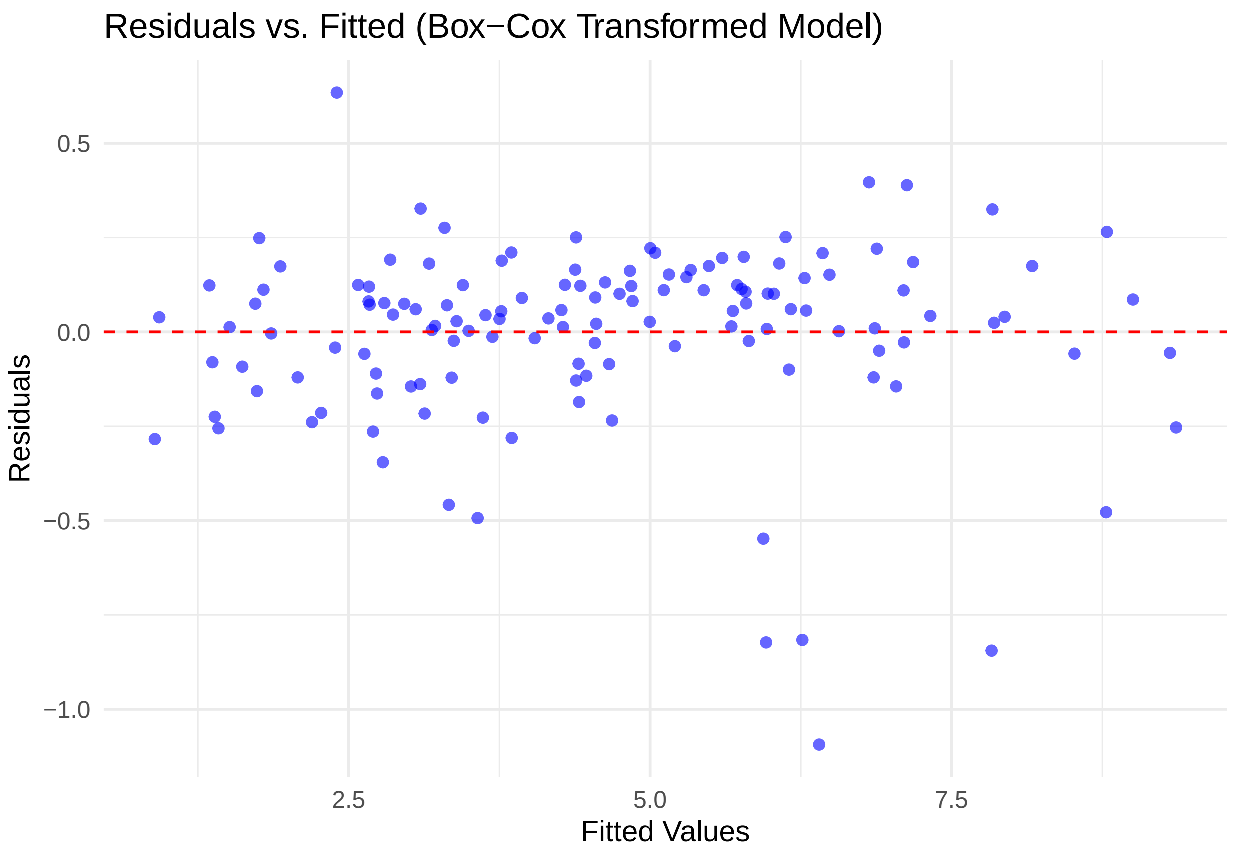
```
## Optimal lambda from Box-Cox: 0.5050505
```

```
# 3. Transform FWI using the optimal lambda:
#    If lambda != 0, transformation is (FWI^lambda - 1)/lambda; if lambda ~ 0, it is equivalent to log
data_clean$FWI_BC <- if (abs(lambda_opt) < 1e-6) log(data_clean$FWI) else ((data_clean$FWI^lambda_opt -

# 4. Fit the backward model using the Box-Cox transformed response
model_bc <- lm(FWI_BC ~ ISI + DMC + DC + FFMC + RH + Ws, data = data_clean)
model_diag_bc <- data.frame(
  fitted = fitted(model_bc),
  residuals = resid(model_bc)
)

# 5. Create residual plots for both models
# Non-transformed model residual plot
p_nontransformed <- ggplot(model_diag_backward, aes(x = fitted, y = residuals)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Residuals vs. Fitted (Non-Transformed Model)",
       x = "Fitted Values", y = "Residuals") +
  theme_minimal()
```

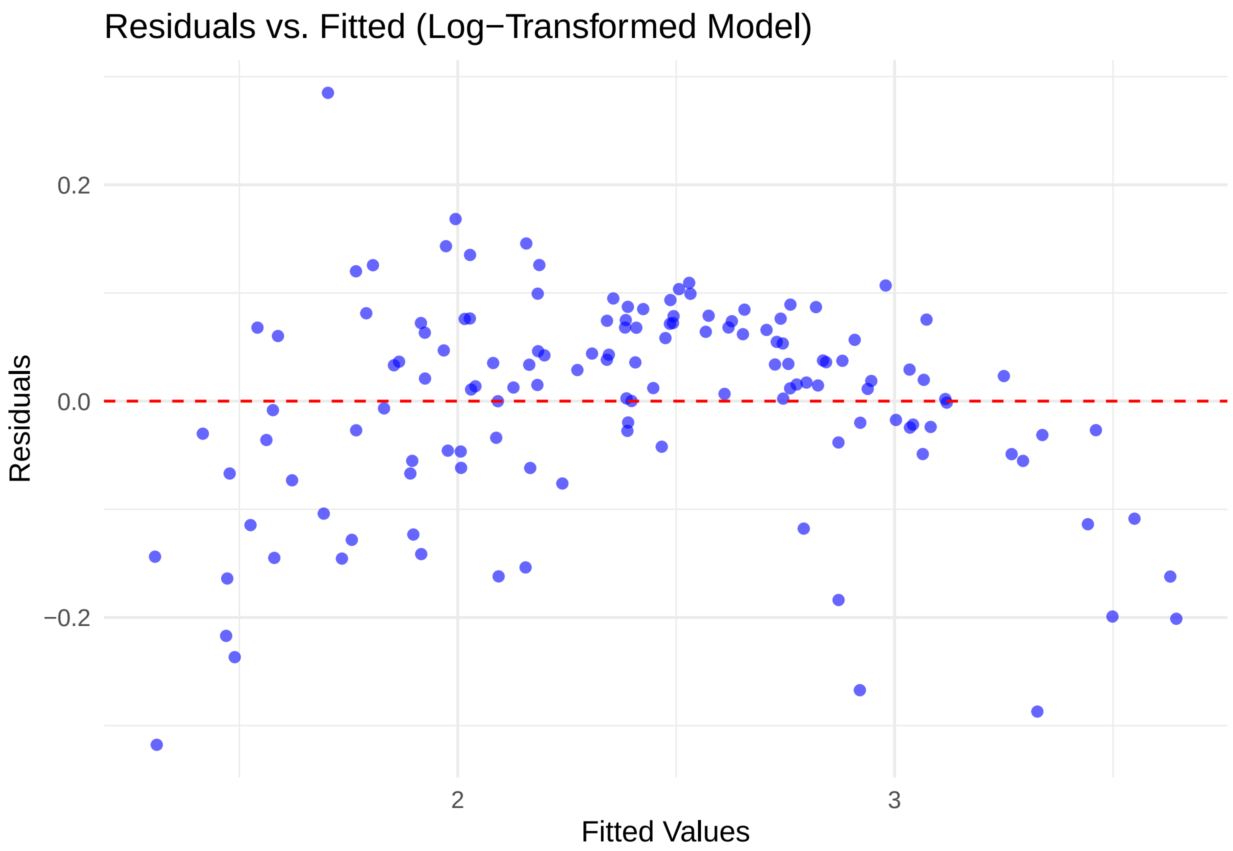
```r
# Box-Cox transformed model residual plot
p_boxcox <- ggplot(model_diag_bc, aes(x = fitted, y = residuals)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Residuals vs. Fitted (Box-Cox Transformed Model)",
       x = "Fitted Values", y = "Residuals") +
  theme_minimal()
print(p_nontransformed)
```



```r
print(p_boxcox)
```


Residuals vs. Fitted (Box−Cox Transformed Model)
0.5
0.0
−0.5
−1.0
Residuals
2.5
5.0
7.5
Fitted Values

```
print(p_log)
```



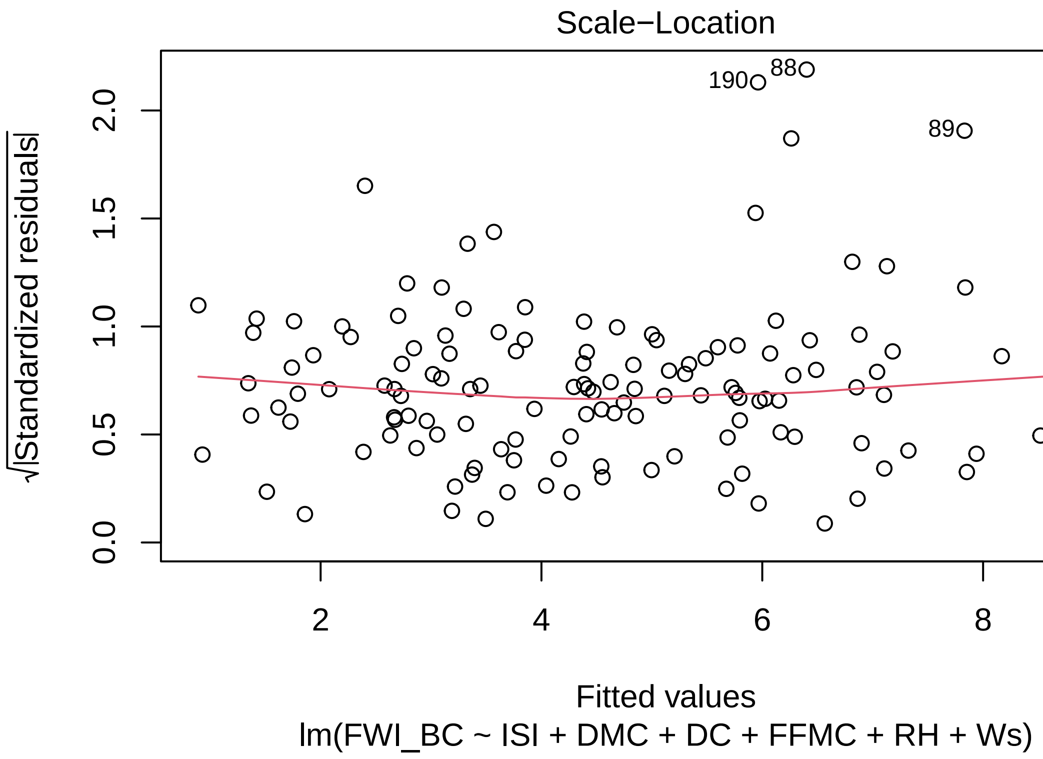
```
bptest(model_backward)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_backward
## BP = 61.857, df = 6, p-value = 1.886e-11
```

```
bptest(model_bc)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_bc
## BP = 20.54, df = 6, p-value = 0.002218
```

```
plot(model_bc, which = 3)
```



```
library(car)
ncvTest(model_bc)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 12.7909, Df = 1, p = 0.00034831
```

```
model_poly <- lm(FWI ~ poly(Temperature, 2) + RH + Ws + ISI + DMC + DC + FFMC, data = data_clean)
summary(model_poly)
```

```
##
## Call:
## lm(formula = FWI ~ poly(Temperature, 2) + RH + Ws + ISI + DMC +
##     DC + FFMC, data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2876 -0.3669 -0.0228  0.5174  2.7577
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -52.158991   7.141417  -7.304 2.65e-11 ***
## poly(Temperature, 2)1  -1.113989   1.536172  -0.725  0.46967
## poly(Temperature, 2)2   0.653047   1.282631   0.509  0.61153
## RH                     -0.021628   0.012289  -1.760  0.08081 .
## Ws                      0.197084   0.058586   3.364  0.00101 **
## ISI                     0.612794   0.083765   7.316 2.49e-11 ***
## DMC                     0.198655   0.017376  11.432  < 2e-16 ***
## DC                      0.026833   0.004162   6.447 2.12e-09 ***
## FFMC                    0.587893   0.082465   7.129 6.59e-11 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.188 on 128 degrees of freedom
## Multiple R-squared:  0.9715, Adjusted R-squared:  0.9697
## F-statistic: 544.6 on 8 and 128 DF,  p-value: < 2.2e-16
```

```
model_int1 <- lm(FWI ~ Temperature * RH + Ws + ISI + DMC + DC + FFMC, data = data_clean)
summary(model_int1)
```

```
##
## Call:
## lm(formula = FWI ~ Temperature * RH + Ws + ISI + DMC + DC + FFMC,
##     data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8887 -0.3751  0.0344  0.5079  2.7037
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -59.982284   8.861326  -6.769 4.20e-10 ***
## Temperature      0.180496   0.128501   1.405 0.162554
## RH               0.109203   0.074138   1.473 0.143214
## Ws               0.212682   0.058011   3.666 0.000359 ***
## ISI              0.572519   0.085657   6.684 6.47e-10 ***
## DMC              0.200363   0.017181  11.662  < 2e-16 ***
## DC               0.027331   0.004122   6.631 8.45e-10 ***
## FFMC             0.609927   0.082553   7.388 1.70e-11 ***
## Temperature:RH  -0.004038   0.002267  -1.781 0.077261 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.175 on 128 degrees of freedom
## Multiple R-squared:  0.9721, Adjusted R-squared:  0.9704
## F-statistic: 557.4 on 8 and 128 DF,  p-value: < 2.2e-16
```

```
# Model 2: Interaction between Wind Speed (Ws) and RH.
model_int2 <- lm(FWI ~ Temperature + RH + Ws * RH + ISI + DMC + DC + FFMC, data = data_clean)
summary(model_int2)
```

```
##
## Call:
## lm(formula = FWI ~ Temperature + RH + Ws * RH + ISI + DMC + DC +
##     FFMC, data = data_clean)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2366 -0.4683 -0.0905  0.7283  2.4157
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -59.086419   7.956865  -7.426 1.39e-11 ***
## Temperature  -0.027162   0.044160  -0.615  0.53960
## RH            0.088195   0.048003   1.837  0.06849 .
## Ws            0.577342   0.173678   3.324  0.00116 **
## ISI           0.589445   0.081571   7.226 3.97e-11 ***
## DMC           0.197743   0.016976  11.648  < 2e-16 ***
## DC            0.027510   0.004085   6.735 4.99e-10 ***
## FFMC          0.612987   0.081520   7.519 8.48e-12 ***
## RH:Ws        -0.007156   0.003044  -2.351  0.02025 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.165 on 128 degrees of freedom
## Multiple R-squared:  0.9726, Adjusted R-squared:  0.9709
## F-statistic: 567.7 on 8 and 128 DF,  p-value: < 2.2e-16
```

```r
# Model 3: Interaction between Temperature and Wind Speed.
model_int3 <- lm(FWI ~ Temperature * Ws + RH + ISI + DMC + DC + FFMC, data = data_clean)
summary(model_int3)
```

```
##
## Call:
## lm(formula = FWI ~ Temperature * Ws + RH + ISI + DMC + DC + FFMC,
##     data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4055 -0.3704 -0.0340  0.5254  2.7049
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -50.358611   9.063590  -5.556 1.53e-07 ***
## Temperature     -0.047653   0.191826  -0.248   0.8042
## Ws               0.160426   0.429595   0.373   0.7094
## RH              -0.020972   0.012337  -1.700   0.0916 .
## ISI              0.621101   0.082214   7.555 7.03e-12 ***
## DMC              0.198037   0.017381  11.394  < 2e-16 ***
## DC               0.026649   0.004193   6.355 3.35e-09 ***
## FFMC             0.585938   0.082929   7.065 9.16e-11 ***
## Temperature:Ws   0.000910   0.012663   0.072   0.9428
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.19 on 128 degrees of freedom
## Multiple R-squared:  0.9714, Adjusted R-squared:  0.9696
## F-statistic: 543.5 on 8 and 128 DF,  p-value: < 2.2e-16
```

```
# Model 4: Interaction between ISI and FFMC (fire weather indices might interact),
# with other variables added additively.
model_int4 <- lm(FWI ~ Temperature + RH + Ws + ISI * FFMC + DMC + DC, data = data_clean)
summary(model_int4)
```

```
##
## Call:
## lm(formula = FWI ~ Temperature + RH + Ws + ISI * FFMC + DMC +
##     DC, data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1175 -0.2749  0.1017  0.4627  3.0902
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -45.713234   6.628810  -6.896 2.19e-10 ***
## Temperature  -0.057807   0.040831  -1.416    0.159
## RH           -0.016252   0.011091  -1.465    0.145
## Ws            0.297209   0.055401   5.365 3.67e-07 ***
## ISI          -3.930962   0.844930  -4.652 8.07e-06 ***
## FFMC          0.536692   0.074869   7.168 5.37e-11 ***
## DMC           0.191428   0.015689  12.201  < 2e-16 ***
## DC            0.028610   0.003767   7.595 5.67e-12 ***
## ISI:FFMC      0.048281   0.008927   5.409 3.01e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.073 on 128 degrees of freedom
## Multiple R-squared:  0.9767, Adjusted R-squared:  0.9753
## F-statistic: 671.4 on 8 and 128 DF,  p-value: < 2.2e-16
```

After testing with both interaction terms and polynomial regressors, we found that the RH:Ws and ISI:FFMC interactions are statistically significant (p-values of 0.02025 and 3.01e-07 respectively). These should be retained in our final model since they add meaningful information about how combined effects influence FWI.

```
final_model <- lm(FWI ~ ISI * FFMC + RH * Ws + DMC + DC, data = data_clean)
summary(final_model)
```

```
##
## Call:
## lm(formula = FWI ~ ISI * FFMC + RH * Ws + DMC + DC, data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7498 -0.3032  0.0298  0.5327  2.8963
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -59.874898   6.777345  -8.835 6.61e-15 ***
## ISI          -4.661306   0.819618  -5.687 8.33e-08 ***
## FFMC          0.568348   0.071197   7.983 7.09e-13 ***
## RH            0.164582   0.042659   3.858  0.00018 ***
## Ws            0.934979   0.160024   5.843 4.01e-08 ***
## DMC           0.189943   0.014828  12.810  < 2e-16 ***
## DC            0.029791   0.003565   8.357 9.22e-14 ***
## ISI:FFMC      0.055511   0.008631   6.431 2.30e-09 ***
## RH:Ws        -0.011427   0.002721  -4.199 4.99e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.014 on 128 degrees of freedom
## Multiple R-squared:  0.9792, Adjusted R-squared:  0.9779
## F-statistic:   754 on 8 and 128 DF,  p-value: < 2.2e-16
```

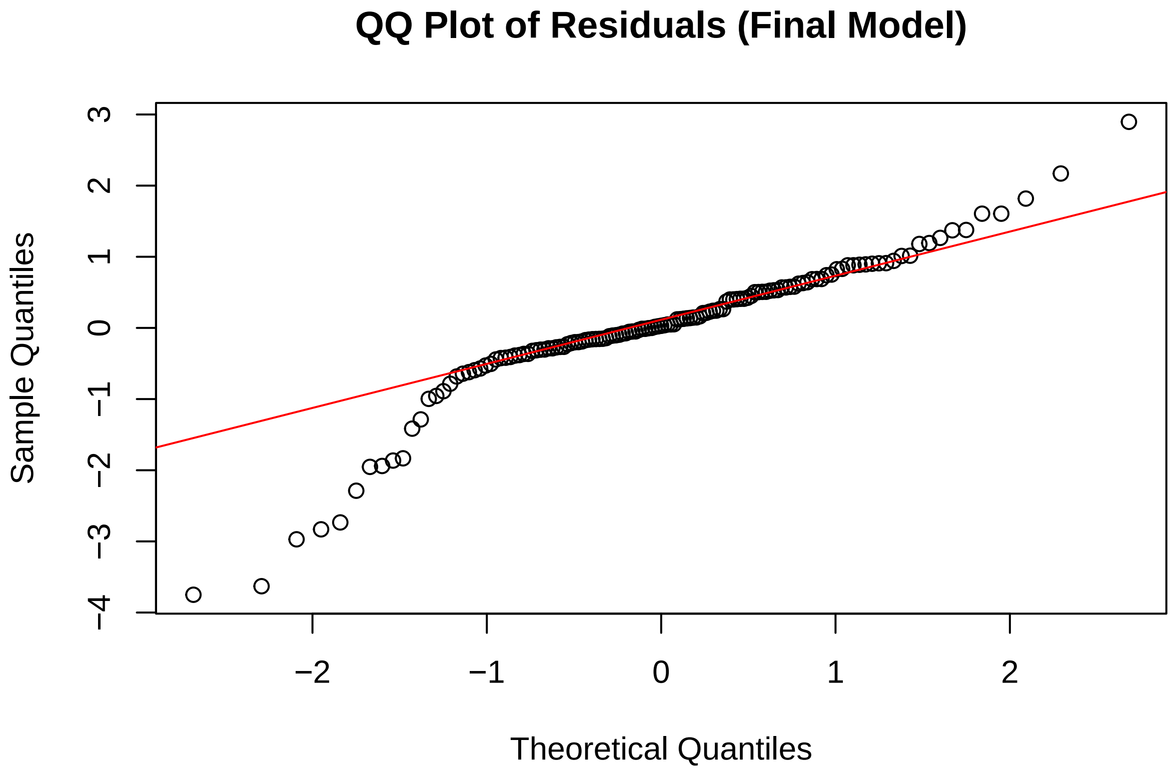
Diagnostic Checking

We center the values to standardize and lower the VIF of the interaction terms without changing the relationship.

```r
data_clean$cISI <- scale(data_clean$ISI, center = TRUE, scale = FALSE)
data_clean$cFFMC <- scale(data_clean$FFMC, center = TRUE, scale = FALSE)
# Then create interaction using centered variables
final_model_centered <- lm(FWI ~ cISI * cFFMC + RH * Ws + DMC + DC, data = data_clean)
vif(final_model_centered)
```

```
##       cISI      cFFMC         RH         Ws        DMC         DC cISI:cFFMC
##  15.826506  12.399694  48.757215  18.027564   4.560170   4.003508   2.206748
##      RH:Ws
##  72.177120
```

```r
qqnorm(resid(final_model), main = "QQ Plot of Residuals (Final Model)")
qqline(resid(final_model), col = "red")
```

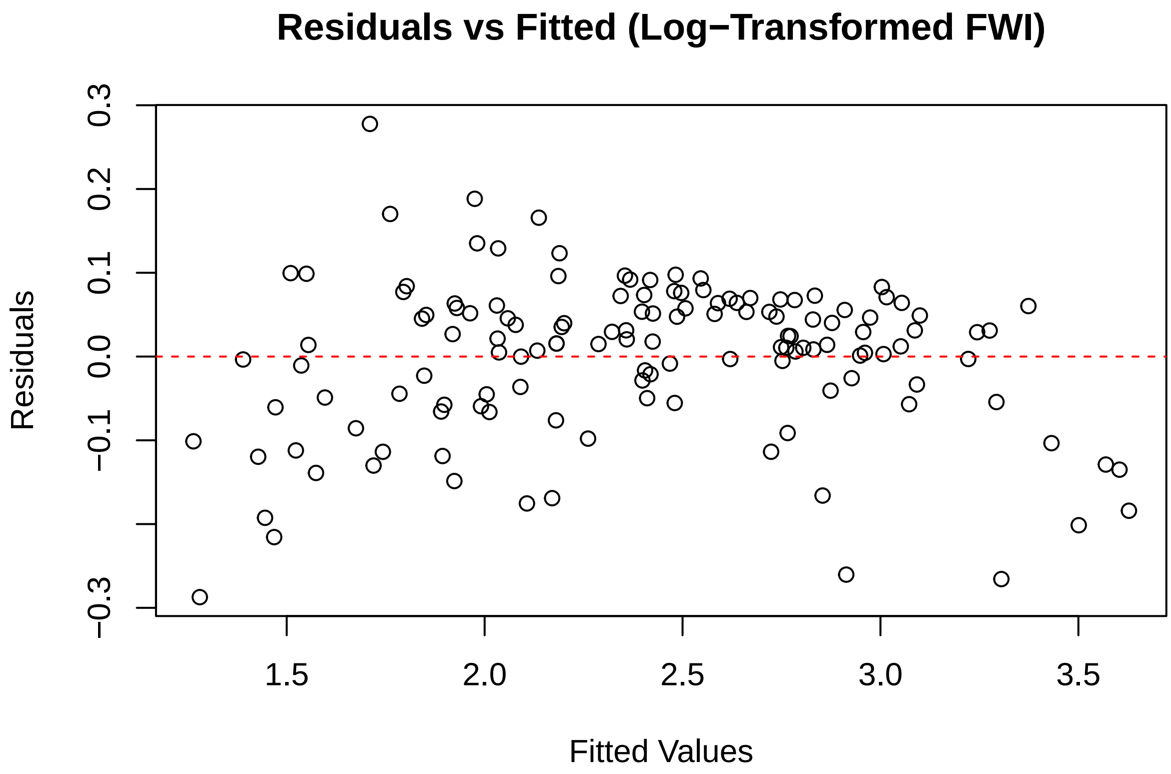


```
final_model_log <- lm(log(FWI + 1) ~ cISI * cFFMC + RH * Ws + DMC + DC, data = data_clean)
summary(final_model_log)
```

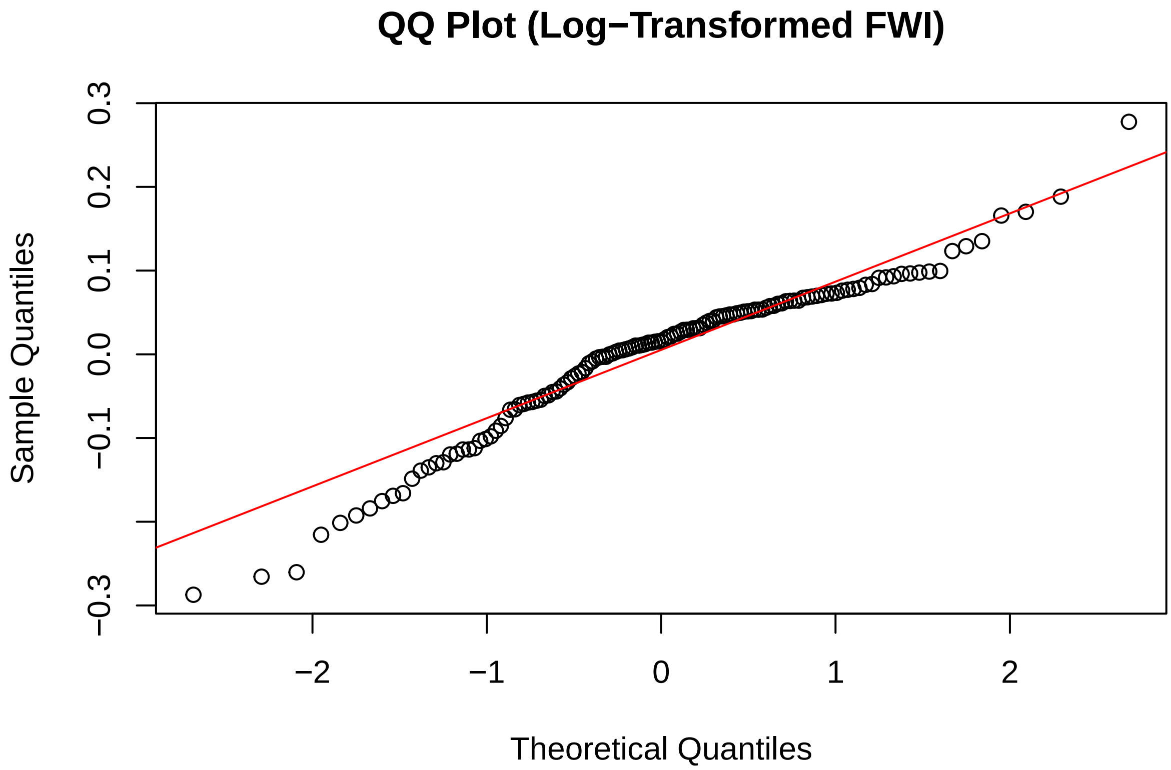
```
##
## Call:
## lm(formula = log(FWI + 1) ~ cISI * cFFMC + RH * Ws + DMC + DC,
##     data = data_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28720 -0.04972  0.01558  0.06026  0.27768
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.8178078  0.2359269   7.705 3.16e-12 ***
## cISI         0.0182654  0.0088485   2.064 0.041015 *
## cFFMC        0.1014731  0.0086287  11.760  < 2e-16 ***
## RH          -0.0054664  0.0040324  -1.356 0.177602
## Ws           0.0154602  0.0151263   1.022 0.308674
## DMC          0.0101877  0.0014016   7.269 3.18e-11 ***
## DC           0.0026878  0.0003370   7.977 7.33e-13 ***
## cISI:cFFMC  -0.0028089  0.0008159  -3.443 0.000779 ***
## RH:Ws        0.0003179  0.0002572   1.236 0.218831
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.09585 on 128 degrees of freedom
## Multiple R-squared:  0.9725, Adjusted R-squared:  0.9708
## F-statistic: 566.5 on 8 and 128 DF,  p-value: < 2.2e-16
```

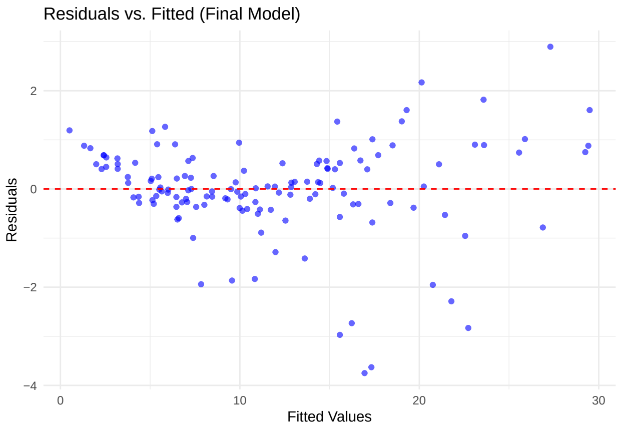
```r
plot(fitted(final_model_log), resid(final_model_log),
     xlab = "Fitted Values", ylab = "Residuals",
     main = "Residuals vs Fitted (Log-Transformed FWI)")
abline(h = 0, col = "red", lty = 2)
```

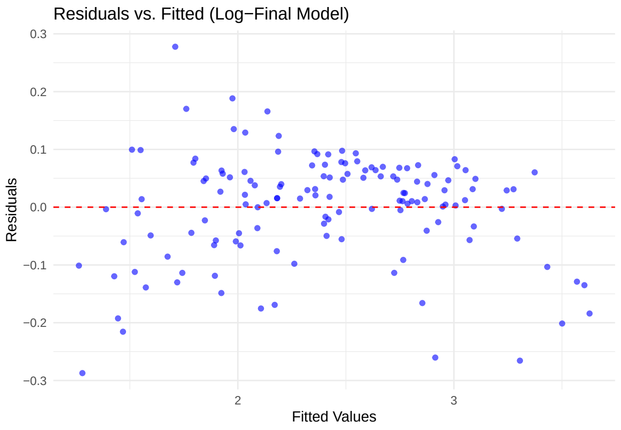


```r
# QQ Plot
qqnorm(resid(final_model_log), main = "QQ Plot (Log-Transformed FWI)")
qqline(resid(final_model_log), col = "red")
```



```
resid_df_original <- data.frame(
  fitted = fitted(final_model),
  residuals = resid(final_model),
  model = "Original (Untransformed)"
)

resid_df_log <- data.frame(
  fitted = fitted(final_model_log),
  residuals = resid(final_model_log),
  model = "Log-Transformed"
)

# Plot for the original final model
p_original <- ggplot(resid_df_original, aes(x = fitted, y = residuals)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residuals vs. Fitted (Final Model)",
       x = "Fitted Values",
       y = "Residuals") +
  theme_minimal()

# Plot for the log-transformed final model
p_log <- ggplot(resid_df_log, aes(x = fitted, y = residuals)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residuals vs. Fitted (Log-Final Model)",
       x = "Fitted Values",
       y = "Residuals") +
  theme_minimal()


print(p_original)
print(p_log)
```





Abstract:

This study analyzes the Algerian Forest Fires dataset to identify the atmospheric factors and fire indices that best predict the Fire Weather Index (FWI) during fire events. After rigorous data cleaning and exploratory analysis, multiple regression models were developed, including the use of Box-Cox transformations and interaction terms, to optimize predictive performance. The final model, which explains over 97% of the variability in FWI, highlights the significant contributions of key fire indices such as ISI, DMC, and FFMC. Diagnostic tests confirm that, after appropriate transformations, the regression assumptions are reasonably met. These findings offer valuable insights for forest management and suggest directions for future research.

Introduction:

The dataset, sourced from the UCI Machine Learning Repository, comprises daily weather observations from two Algerian regions (Bejaia and Sidi Bel-abbes). It includes measurements such as temperature, relative humidity (RH), wind speed (Ws), and rainfall, alongside several fire indices (FFMC, DMC, DC, ISI, BUI) that are used to compute the Fire Weather Index (FWI). This analysis exclusively focuses on observations labeled as "fire" to accurately assess the risk factors present during actual fire events. The primary objective is to identify the atmospheric conditions and fire indices that are most predictive of the FWI, thereby enhancing our understanding of fire risk and informing effective forest management strategies.

Research Questions:

1. How does Temperature affect the Fire Weather Index (FWI) during fire events?
2. How do additional weather factors such as Relative Humidity (RH), Rain, and Wind Speed (Ws) affect FWI during fire events?
3. Do the Codes and Indices (FFMC, DMS, DC, ISI, BUI) significantly predict FWI during fire events?

Hypotheses:

- Null Hypothesis:
  None of the examined weather factors or fire indices significantly influence FWI during fire events.
- Alternative Hypothesis:
  At least one of the fire indices has a significant relationship with FWI during fire events.

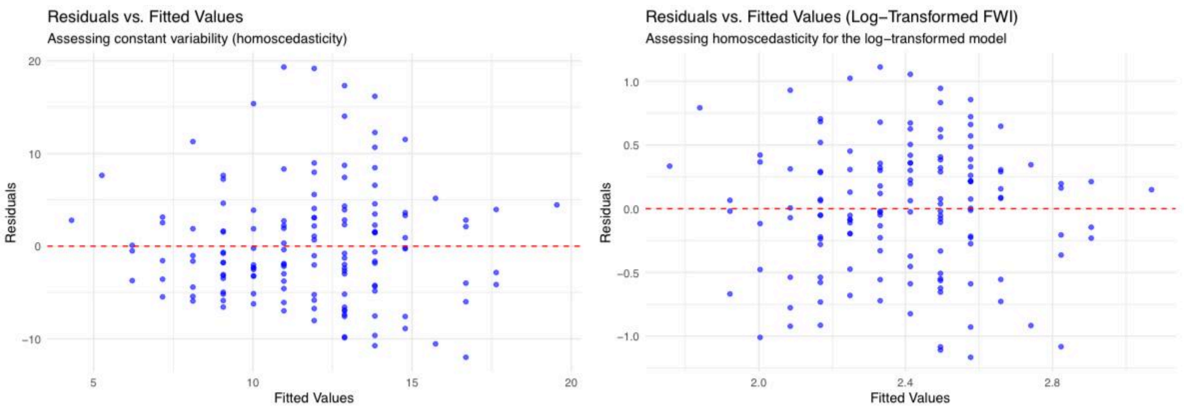
Data Processing:

Prior to analysis, extensive data cleaning and preprocessing were performed. Raw CSV data was imported, and irrelevant rows along with repeated headers were removed. Variables were then converted to appropriate numeric types. A set of validation criteria ensured that each variable's values fell within expected ranges (e.g., temperature between 22°C and 42°C, RH between 21% and 90%). We then filtered the data so that only observations with the class, “fire”, were kept so we could focus the analysis on actual fire events. Missing values and outliers were carefully identified and handled during the cleaning process. In later stages, the response variable (FWI) was transformed using both log and the Box-Cox approaches to stabilize variance and improve the model fit. Continuous predictors were considered for standardization to ensure comparability across variables.

Exploratory Data Analysis (EDA):

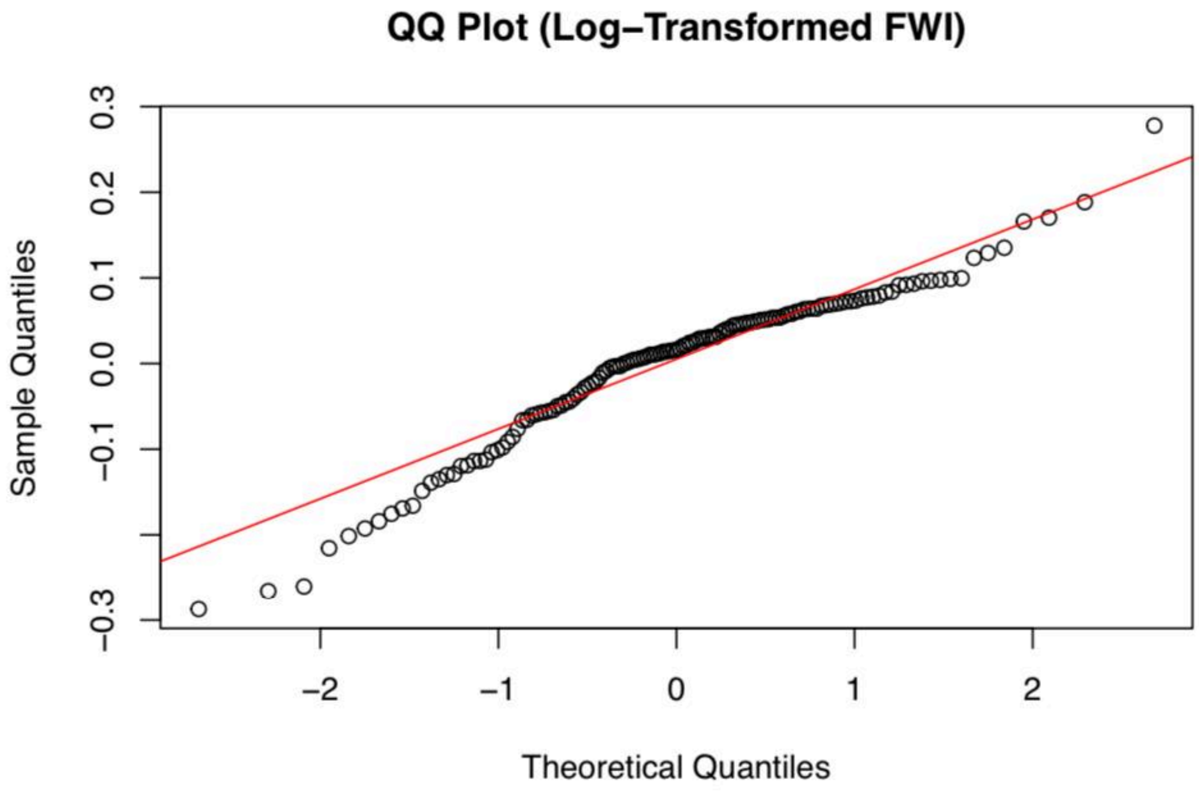
Summary statistics were computed for the weather variables and fire indices, providing insights into the central tendency and dispersion of the data. For example, the average temperature and its standard deviation helped establish a baseline understanding of the atmospheric conditions. The histograms and density plots revealed that while variables such as temperature and relative humidity (RH) exhibited relatively symmetric distributions, the rainfall variable was highly right-skewed. Skewness and kurtosis metrics confirmed these observations, indicating that certain variables might require transformation prior to modeling. Correlation matrices and heatmaps showcased strong associations among key predictors. Notably, temperature showed a positive relationship with the FWI (r ≈ 0.57), whereas RH and rain were negatively correlated with FWI. Among the fire indices, ISI, FFMC, DMC, and DC were strongly positively correlated with FWI. Pair plots further supported these findings, visually confirming the linear relationships that informed subsequent regression model development.

Visualizations:



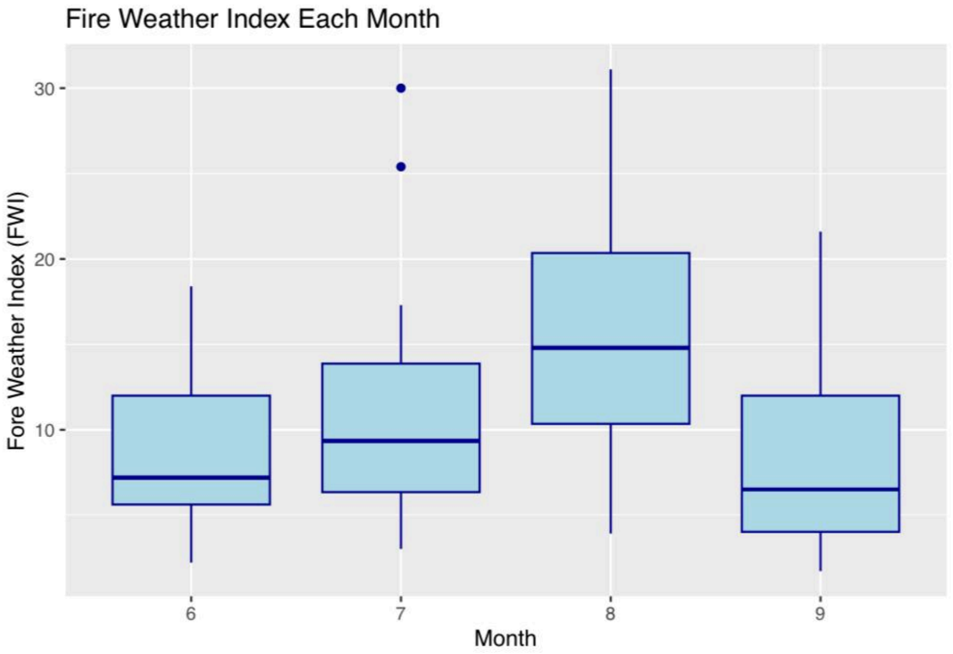
Caption:

Residuals vs. Fitted Plot: This plot checks the linearity assumption and homoscedasticity of the model errors. A random scatter of residuals around zero suggests that the model fits the data well. Comparing the original and log-transformed plots demonstrates whether transformation improves the residual distribution.

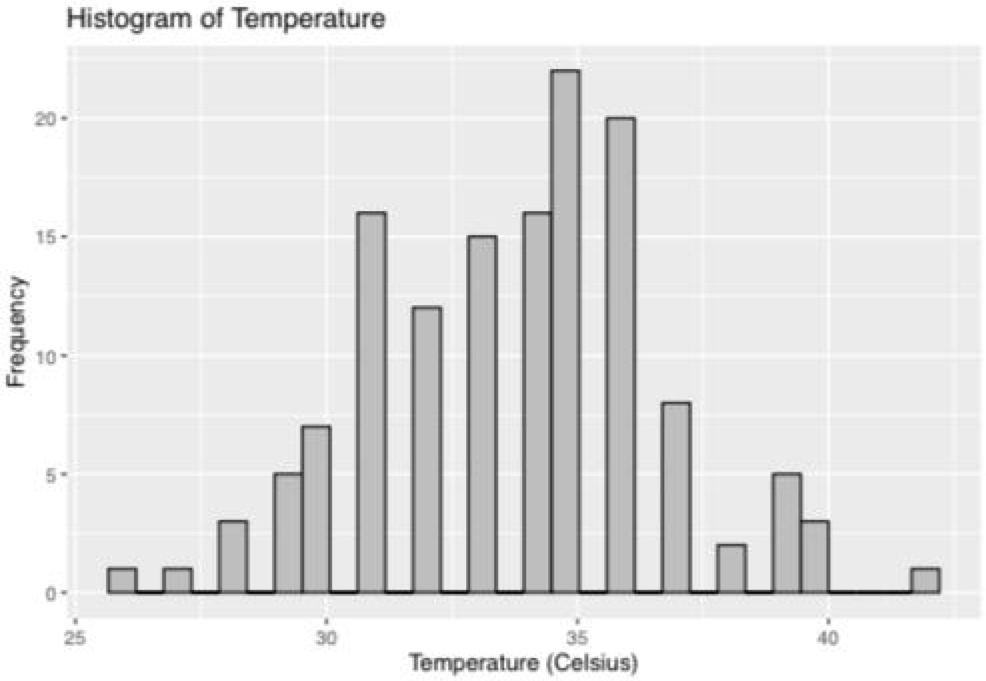


Caption:

Q-Q Plot: This plot compares the quantiles of the model residuals to those of a normal distribution. Points closely following the reference line indicate that the residuals are normally distributed.

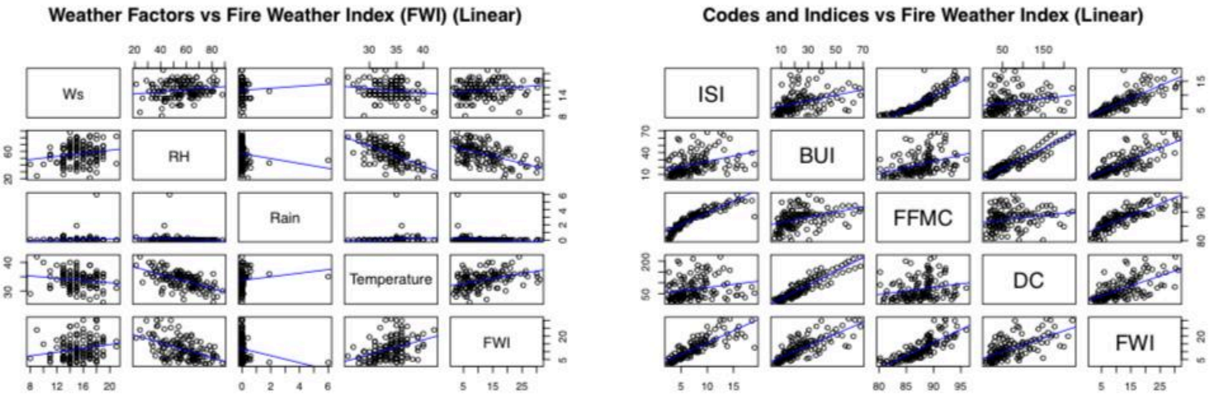


Caption:
Boxplot of FWI by Month: This plot displays the distribution of FWI for each month. It highlights the seasonal variations and potential differences in fire risk across the fire season.

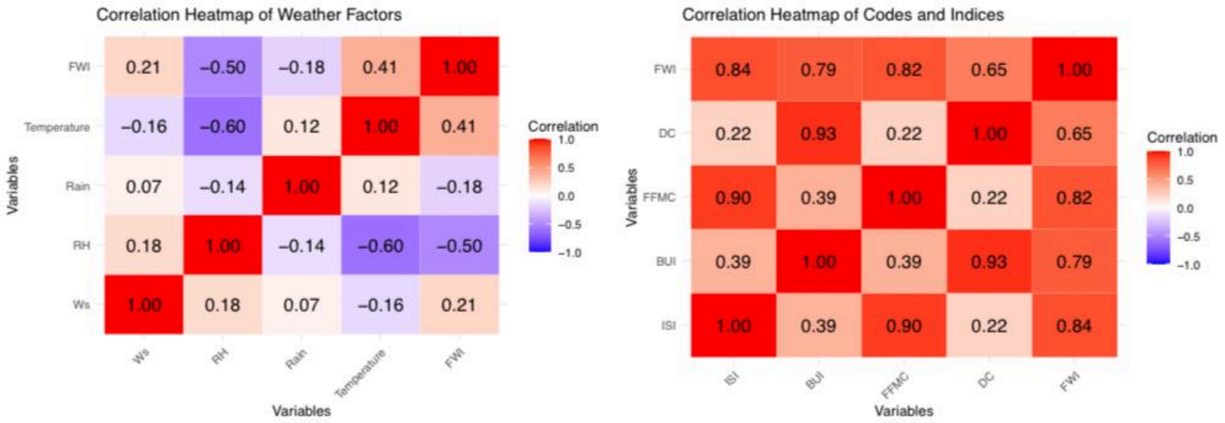


Caption:

Histograms: These plots illustrate the distributions of key variables. The Temperature histogram shows the central tendency and spread of daily temperatures. The FWI histogram reveals the distribution of the Fire Weather Index and highlights any skewness or outliers.



Caption:
Pair Plots: These plots depict the bivariate relationships between sets of predictors and FWI. The inclusion of fitted regression lines helps visualize the linear trends and supports the choice of predictors for the regression analysis.



Caption:
Correlation Heatmap: This heatmap visualizes the strength and direction of pairwise correlations among variables. These help identify which predictors are most strongly associated with the Fire Weather Index, assisting in the variable selection process.

Explain Your Regression Model:

Two modeling strategies were used in our analysis: Meteorological Factors Model and Fire Indices and Combined Model. Initially, we developed a simple linear regression model using only the meteorological variables, Temperature, Relative Humidity (RH), Wind Speed (Ws), and Rainfall, as well as month as a categorical predictor. While some predictors in this model were

statistically significant, the adjusted R^2 was only around 0.50. This indicated that these variables alone explained roughly 50% of the variability in the Fire Weather Index (FWI), suggesting that other factors might be influencing fire risk. After that, given the strong correlations observed between FWI and the fire indices (ISI, DMC, DC, FFMC), we expanded our approach by incorporating these indices. We employed sequential variable selection techniques (both forward and backward stepwise selection) to identify the most relevant predictors. This process resulted in a combined model that not only includes key fire indices but also important interaction terms that capture the joint effects of certain variables.
Our final model is:
FWI = −59.87 − 4.66 × ISI + 0.57 × FFMC + 0.16 × RH + 0.93 × Ws + 0.19 × DMC + 0.03 × DC + 0.056 × (ISI × FFMC) − 0.011 × (RH × Ws) + ϵ

This final model had an adjusted R^2 of approximately 0.978, meaning that nearly 97.8% of the variability in FWI is explained by the selected predictors and interaction terms. The diagnostic tests confirmed that the final model meets key regression assumptions. The inclusion of interactions (between ISI and FFMC, and between RH and Ws) suggests that the combined effects of these variables are crucial in understanding fire behavior. For example, the interaction between ISI and FFMC implies that the impact of one fire index on FWI depends on the level of the other, which highlights the complex dynamics of fire risk factors. These modeling decisions directly address our research questions and provide valuable insights for forest management and future research.

Interpret Your Results:

The final regression model explains approximately 97.8% of the variability in the Fire Weather Index (FWI) (adjusted R^2 is approximately 0.978), which indicates an excellent fit. Since the overall p-value, $2.2x10^{-16}$ is less than the significance level $\alpha = .05$, the overall model is statistically significant.

β_0 (Intercept): the intercept represents that the baseline FWI when all predictors are zero is -59.87.
β_1 (ISI): a one-unit increase in the Initial Spread Index (ISI) is associated with a 4.66 unit decrease in FWI when FFMC is zero.
β_2 (FFMC): a one-unit increase in the Fine Fuel Moisture Code (FFMC) is linked to a 0.57 unit increase in FWI.
β_2 (RH) : each one-unit increase in RH increases FWI by 0.16 units.
β_3 (Ws) : each one-unit increase in Ws raising FWI by 0.93 units.
β_4 (DMC) : each one-unit increase in DMC is associated with an increase of 0.19 units in the Fire Weather Index (FWI).
β_5 (DC) : each one-unit increase in DC is associated with a 0.03 unit increase in FWI
β_6 (ISI:FFMC) : for every one-unit increase in FFMC, the effect of ISI on FWI increases by 0.056 units.

β_7 (RH:Ws) : for every one-unit increase in RH, the effect of Ws on FWI decreases by 0.011 units.

All predictors in the final model have p-values less than 0.05, which indicates that they are all statistically significant. The coefficients for ISI, FFMC, RH, Ws, DMC, and DC are all significant, suggesting that both meteorological factors and fire indices independently contribute to predicting FWI. The interaction term between ISI and FFMC and between RH and Ws are both significant. These interactions reveal that the effect of one predictor on FWI is dependent on the level of another. For example, FFMC changes the effect of ISI, showing that these indices interact rather than just adding their separate effects.

Check Assumptions:

We check linearity and homoscedasticity by plotting residuals vs. fitted values for both the original and log-transformed models. The plots showed a random scatter around zero for both the original and log-transformed models, which supports the linearity assumption. The plot also shows that the residuals have a constant spread across the fitted values, which supports the homoscedasticity assumption. In addition to the plot, we further tested for homoscedasticity by using a Breush-Pagan test. The p-value we got from the Breusch-Pagan test was less than the $\alpha = 0.05$ significance level, so we can conclude that the data is homoscedastic. We also made a Q-Q plot. The Q-Q plot showed that the residuals closely follow the reference line, which suggests that the errors are normally distributed. We checked for independence by inspecting the residual plots to see if there were any patterns or trends that might suggest autocorrelation. Since we did not see any patterns or trends, we can assume independence. Since the log and Box-Cox transformations helped stabilize the variance and improve the model's diagnostic plots, it is recommended to use the transformed model when making predictions.

Recommendations:

Our final model has an excellent fit with an adjusted R^2 of approximately 0.978, but this unusually high value suggests that the model may be overfitting the data. To address this, we recommend implementing cross-validation to assess the model's performance on independent datasets and reducing complexity by removing predictors that don't contribute much to the overall fit. Regularization methods such as ridge or lasso regression could also help by shrinking overly complex coefficients. Additionally, some of our predictors have a high VIF value which indicates multicollinearity. To combat this, we recommend combining highly correlated predictors into composite variables or removing redundant ones. Finally, the significant interactions, specifically between ISI and FFMC and between RH and wind speed, showcase the importance of considering joint effects rather than treating variables in isolation. Future fire risk models should include these interaction effects and be regularly updated with new data to ensure they remain accurate.

Conclusions:

This study analyzed the Algerian Forest Fires dataset to predict the Fire Weather Index (FWI) using multiple regression analysis. Our final model, which includes both key fire indices (ISI, DMC, DC, FFMC) and meteorological factors (RH, Ws) along with significant interaction terms (ISI × FFMC and RH × Ws), explained nearly 97.8% of the variability in FWI. These results suggest that the combined effects of these variables are crucial for understanding fire risk during fire events. Diagnostic checks indicated that after appropriate transformations (log and Box-Cox), the model meets the assumptions of linearity, normality of errors, homoscedasticity, and independence. However, the unusually high adjusted R^2 raises concerns about potential overfitting, and some predictors have high VIF values which suggests the presence of multicollinearity. Overall, while the model shows strong explanatory power, limitations related to overfitting and multicollinearity need to be acknowledged. Future work should focus on validating the model with new data and exploring alternative modeling techniques to ensure accurate and generalizable predictions.