# Non-Parametric Analysis of the Forest Fires Dataset

Parvez Khan (241080082)
Piyush Kumar (241080083)
Prottush Das (241080085)
Riya Sharma (241080088)
Sahil Kumar Kalsaria (241080091)

## Contents

## 1. Introduction

Forest fires remain a significant threat, causing environmental, economic, and social devastation globally. Understanding and accurately predicting the **extent of the burned area** is crucial for effective resource allocation in fire management and for developing preventative strategies.

This study utilizes the **Forest Fires dataset (Cortez and Morais, 2007)** from the **UCI Machine Learning Repository**, which contains **517** entries (fire instances) and **13** variables, including spatial coordinates, temporal indicators (**month**, **day**), four components of the **Fire Weather Index (FWI)**, and four direct meteorological variables (**temperature**, **relative humidity**, **wind**, and **rain**). The target variable is the **total burned area (area)**.

Given the **highly skewed distribution** of the burned area, which often violates **normality assumptions**, **nonparametric statistical methods** and **specialized regression techniques** are often employed to ensure robust and interpretable results. Analyses often combine exploratory and inferential techniques. **Nonparametric tests** and **rank-based correlation measures** can assess group differences and quantify associations among variables. Preliminary results often reveal significant relationships between the **FWI components (FFMC, DMC, DC, ISI)**, **temperature (temp)**, and **wind (wind)** with the magnitude of the **burned area**.

Overall, the findings demonstrate the effectiveness of using a combination of **spatial**, **temporal**, and **meteorological** data in **data-driven fire management** and **predictive modeling** of fire extent.

## 2. Dataset Description

This study utilizes the **Forest Fires dataset**, originally contributed by **P. Cortez and A. Morais (2007)**, often sourced from the **UCI Machine Learning Repository**. The data comprises **517** instances, representing recorded forest fires, primarily in the **Montesinho Natural Park** in the **northeast of Portugal**.

The dataset captures **multivariate spatial, temporal, and meteorological information** relevant to the extent of a forest fire. It includes **13** attributes covering spatial coordinates (**X**, **Y**), temporal variables (**month**, **day**), four Fire Weather Index components (**FFMC**, **DMC**, **DC**, **ISI**) from the Canadian FWI system, and four direct meteorological measures (**temp**, **RH**, **wind**, **rain**).

The target variable (**area**), representing the **total burned area (in hectares)**, is a **continuous numeric variable**. Statistical inspection reveals that the **area variable is heavily skewed toward zero**, with **247 of the 517 instances** recording **0.0 ha** of burned area. The dataset contains **no missing values**, and the features are a mix of **categorical**, **integer**, and **continuous** data types.

## 3. Loading the Dataset and Libraries

The analysis utilizes several **R packages**, including **tseries** and **evd** for statistical testing, **dplyr** and **tidyr** for data manipulation, **ggplot2** for visualization, **fitdistrplus** for distribution fitting, **corrplot** for correlation analysis, and **stats** for core statistical functions. The **evd** package is particularly useful for **fitting extreme value distributions**, which is relevant given the presence of **extreme fire events** in the dataset.

The dataset is loaded from a **CSV file**, and the `attach()` function is used to make variable names directly accessible in the R environment, simplifying subsequent analyses. Initial inspection of column names confirms the **presence of all expected variables**.

```
##     X Y month day FFMC   DMC    DC  ISI temp RH wind rain  area
## 144 1 2   jul sat 90.0  51.3 296.3  8.7 16.6 53  5.4    0  0.71
## 361 6 5   sep fri 92.5 122.0 789.7 10.2 18.4 42  2.2    0  1.09
## 458 1 4   aug wed 91.7 191.4 635.9  7.8 19.9 50  4.0    0 82.75
## 290 7 4   jul sat 91.6 104.2 474.9  9.0 24.3 30  1.8    0  0.00
## 89  1 2   sep sun 93.5 149.3 728.6  8.1 25.3 36  3.6    0  0.00
## 121 3 4   aug mon 91.5 145.4 608.2 10.7 10.3 74  2.2    0  0.00
## 249 8 6   aug wed 93.1 157.3 666.7 13.5 28.7 28  2.7    0  0.00
```

## 4. Data Overview & Summary

### Columns

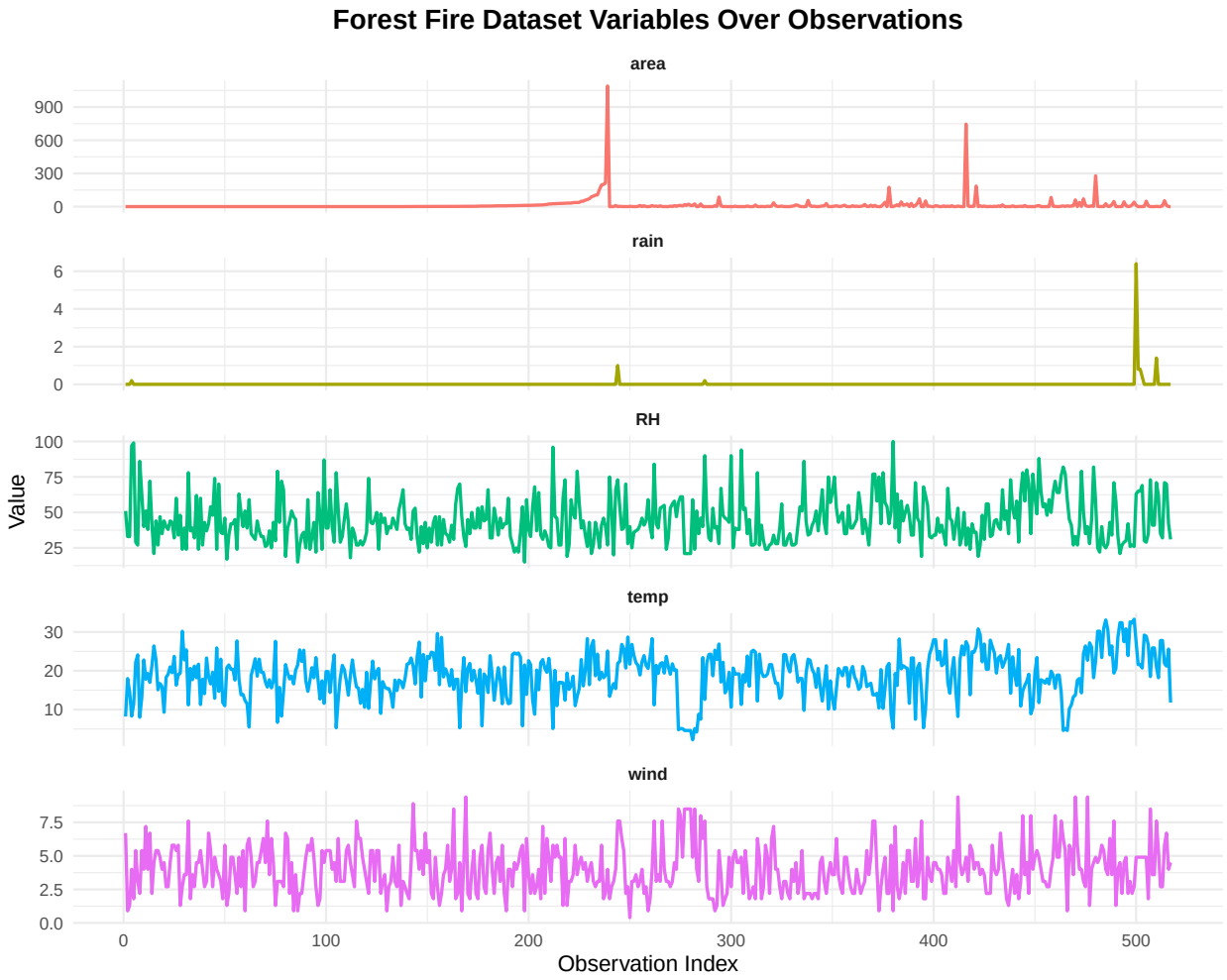
The dataset comprises categorical variables (month, day) that capture temporal patterns, and continuous variables (FFMC, DMC, DC, ISI, temp, RH, wind, rain, area) that quantify fire weather indices, meteorological conditions, and fire severity. Understanding the distribution and behavior of these variables across observations is essential for identifying patterns and potential relationships.

### Summary Statistics

Exploratory visualization reveals temporal patterns and variable distributions across the 517 fire instances. The following time series plot illustrates how key meteorological variables and the burned area evolve across observations, helping identify trends, seasonality, and potential outliers:

```
##        X               Y          month               day
##  Min.   :1.000   Min.   :2.0   Length:517         Length:517
##  1st Qu.:3.000   1st Qu.:4.0   Class :character   Class :character
##  Median :4.000   Median :4.0   Mode  :character   Mode  :character
##  Mean   :4.669   Mean   :4.3
##  3rd Qu.:7.000   3rd Qu.:5.0
```

```
##  Max.   :9.000   Max.   :9.0
##       FFMC             DMC              DC             ISI
##  Min.   :18.70   Min.   :  1.1   Min.   :  7.9   Min.   : 0.000
##  1st Qu.:90.20   1st Qu.: 68.6   1st Qu.:437.7   1st Qu.: 6.500
##  Median :91.60   Median :108.3   Median :664.2   Median : 8.400
##  Mean   :90.64   Mean   :110.9   Mean   :547.9   Mean   : 9.022
##  3rd Qu.:92.90   3rd Qu.:142.4   3rd Qu.:713.9   3rd Qu.:10.800
##  Max.   :96.20   Max.   :291.3   Max.   :860.6   Max.   :56.100
##       temp             RH              wind            rain
##  Min.   : 2.20   Min.   : 15.00   Min.   :0.400   Min.   :0.00000
##  1st Qu.:15.50   1st Qu.: 33.00   1st Qu.:2.700   1st Qu.:0.00000
##  Median :19.30   Median : 42.00   Median :4.000   Median :0.00000
##  Mean   :18.89   Mean   : 44.29   Mean   :4.018   Mean   :0.02166
##  3rd Qu.:22.80   3rd Qu.: 53.00   3rd Qu.:4.900   3rd Qu.:0.00000
##  Max.   :33.30   Max.   :100.00   Max.   :9.400   Max.   :6.40000
##       area
##  Min.   :   0.00
##  1st Qu.:   0.00
##  Median :   0.52
##  Mean   :  12.85
##  3rd Qu.:   6.57
##  Max.   :1090.84
```




The visualization highlights the variability in meteorological conditions and the sporadic nature of burned area, with most observations showing zero or minimal area burned, punctuated by occasional extreme fire events.

### Target Variable Summary

To investigate the temporal patterns and behavior of different variables across specific months, two subsets of the original dataset were created:

- **august_data**: Contains all fire instances that occurred during the month of August
- **september_data**: Contains all fire instances that occurred during the month of September

These monthly subsets allow for focused analysis of how meteorological conditions, FWI components, and other predictors behave differently during these critical fire months. By examining these subsets separately, we can identify month-specific patterns and relationships that may be masked in the full dataset analysis.

A binary fire occurrence variable is created to distinguish between days with and without burned area:

This categorization enables comparative analysis between fire and no-fire conditions, facilitating two-sample non-parametric tests.

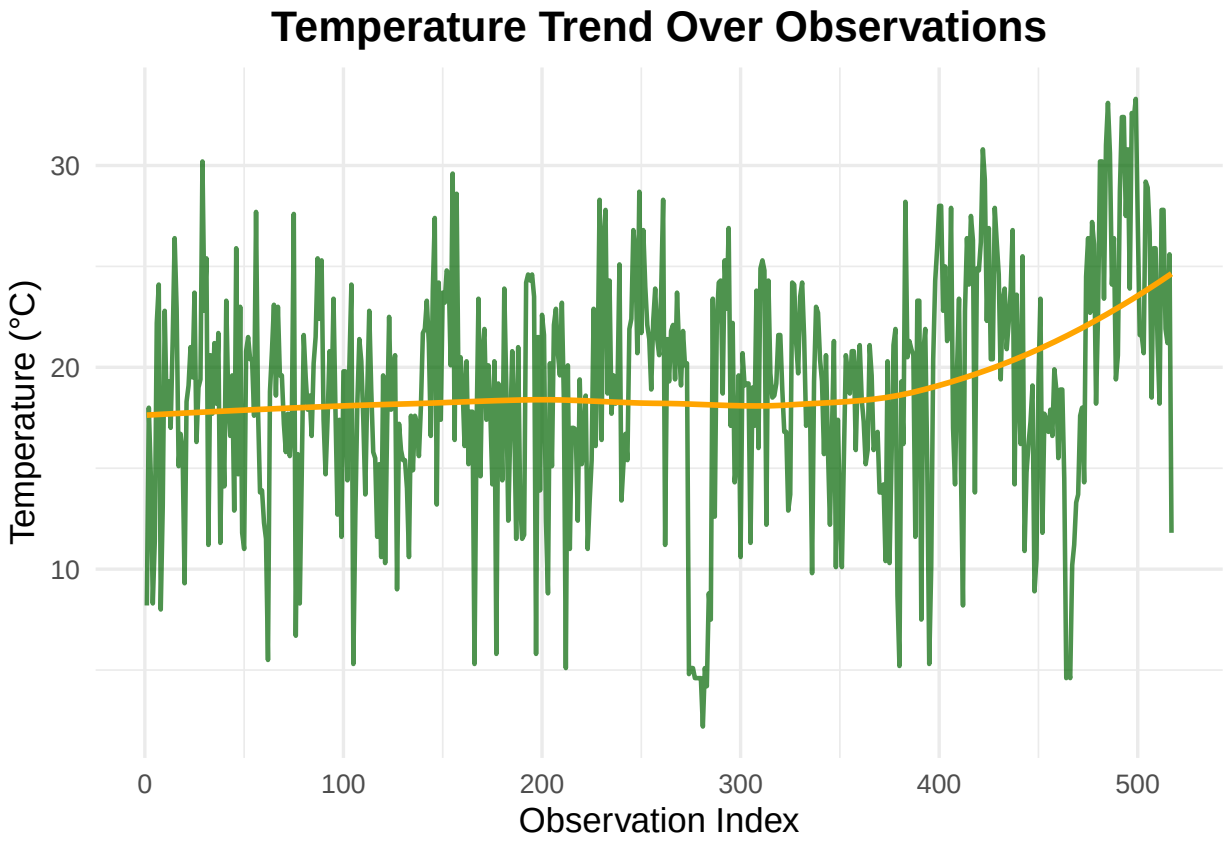
## 5. Non-Parametric Tests

### 5.1 One-Sample Tests

**5.1.1 Runs Test for Randomness** The runs test checks whether the sequence of observations (e.g., temperature, humidity, wind, area, rain) occurs in a random order over time. For our dataset, this helps determine if high and low values of environmental variables (like temperature, humidity, etc.) are randomly distributed, or if there's an underlying pattern such as clustering or trends.

**Temperature**

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



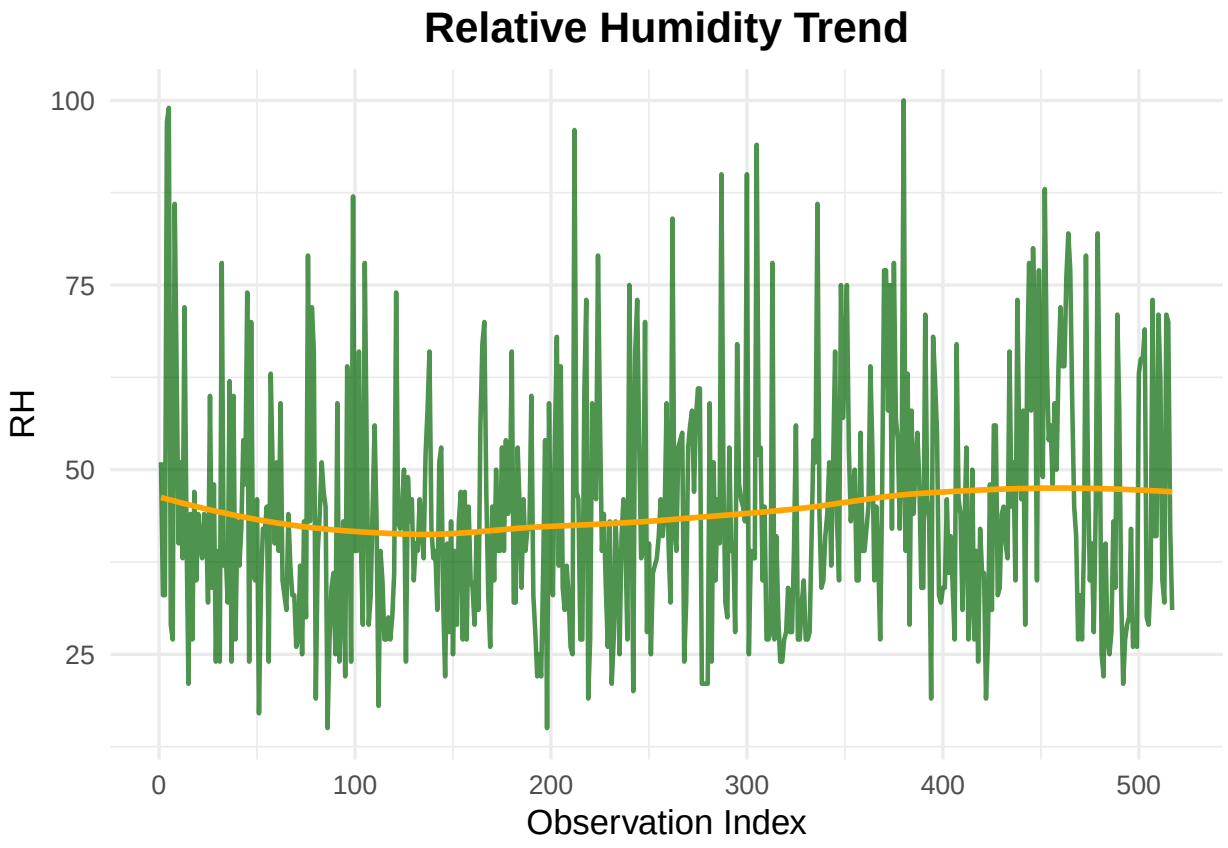
- H0: High and low temperature days occur randomly.
- H1: High and low temperature days do not occur randomly.

```
##
##  Runs Test
##
## data:  as.factor(temp_binary)
## Standard Normal = -6.028, p-value = 1.66e-09
## alternative hypothesis: two.sided
```

- p-value = 1.66e-09 : Reject H0
- **Conclusion:** Temperature sequences exhibit significant non-randomness ($p < 0.001$), indicating systematic temporal patterns.

### Relative Humidity

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



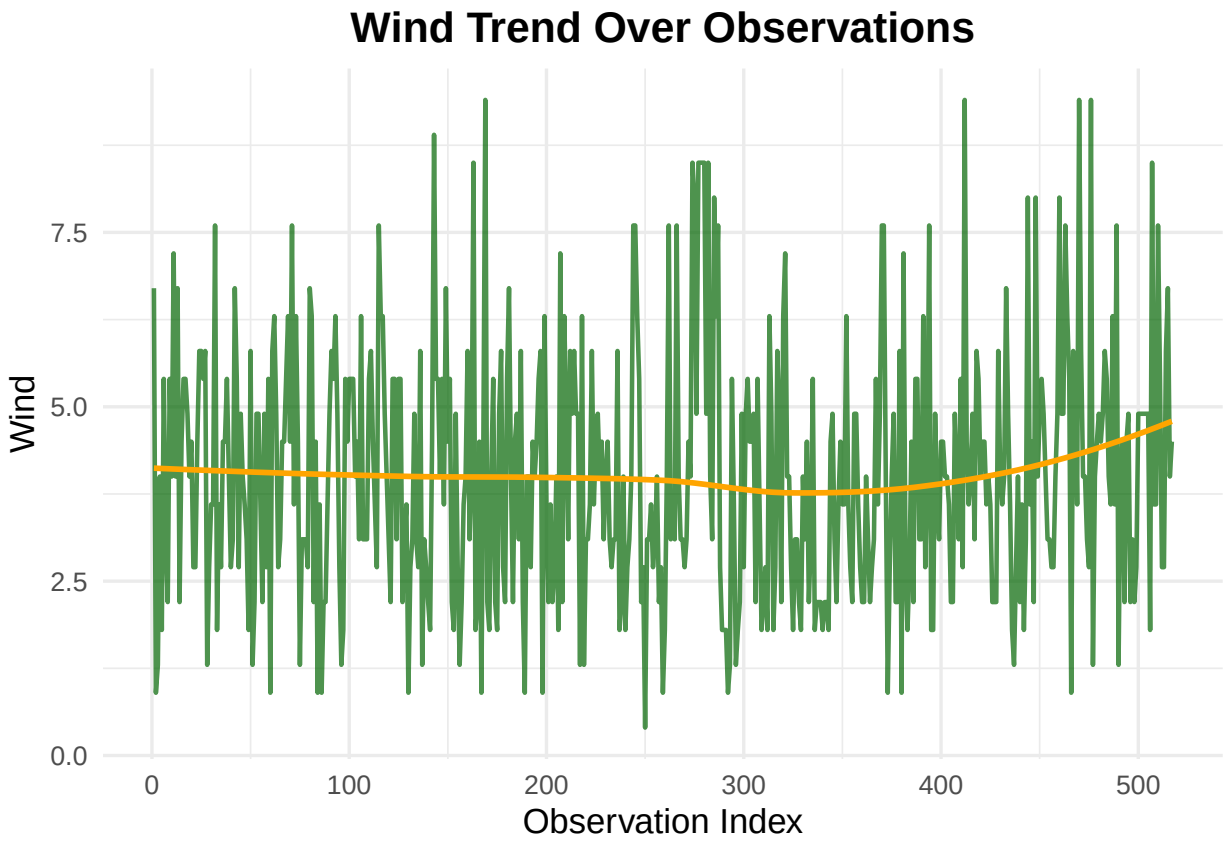
- H0: High and low RH days occur randomly
- H1: High and low RH days do not occur randomly

```
##
##  Runs Test
##
## data:  as.factor(RH_binary)
## Standard Normal = -2.8685, p-value = 0.004124
## alternative hypothesis: two.sided
```

- p-value = 0.004124 : Reject H0
- **Conclusion:** Relative humidity shows clustering or trend (p = 0.004).

### Wind Speed

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




- H0: High and low wind days occur randomly.
- H1: High and low wind days do not occur randomly.

```
##
##  Runs Test
##
## data:  as.factor(wind_binary)
## Standard Normal = -2.6967, p-value = 0.007003
## alternative hypothesis: two.sided
```

- p-value = 0.007003 : Reject H0
- **Conclusion:** The sequence of high and low wind speed days is not random (p = 0.007), indicating a systematic or periodic pattern in wind behavior over time.

**Burned Area**

- H0: High and low burned area days occur randomly.
- H1: High and low burned area days do not occur randomly.

```
##
##  Runs Test
##
## data:  as.factor(area_binary)
## Standard Normal = -12.194, p-value < 2.2e-16
## alternative hypothesis: two.sided
```

- p-value < 2.2e-16 : Reject H0
- **Conclusion:** The sequence of high and low burned area days is not random (p < 0.05), indicating a systematic pattern or clustering of fire-affected days over time.

**Rain Occurrence**

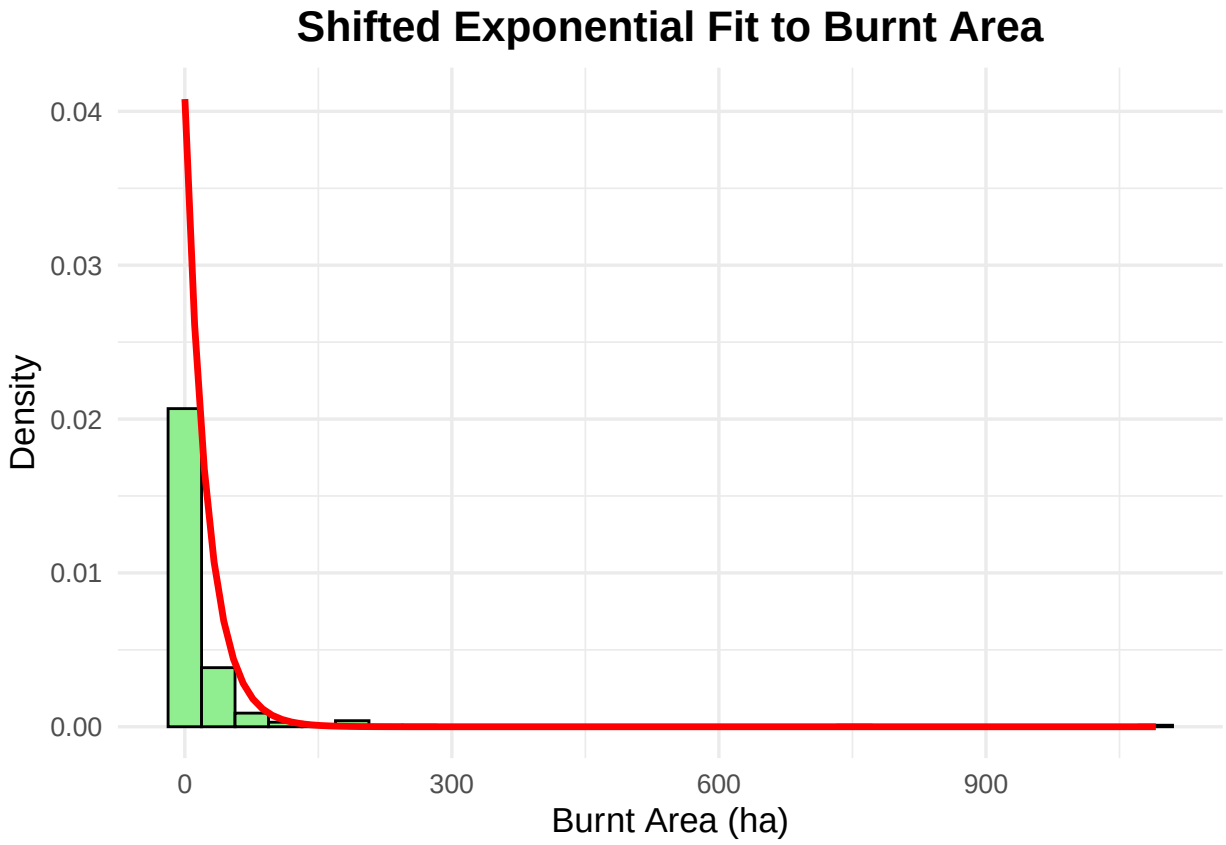
- H0: Rain/no-rain days occur randomly.
- H1: Rain/no-rain days do not occur randomly.

```
##
##  Runs Test
##
## data:  as.factor(rain_binary)
## Standard Normal = -8.5718, p-value < 2.2e-16
## alternative hypothesis: two.sided
```

- p-value < 2.2e-16 : Reject H0
- **Conclusion:** All variables reject randomness (p < 0.05), confirming systematic temporal patterns, trends, or clustering in meteorological conditions and fire occurrences.

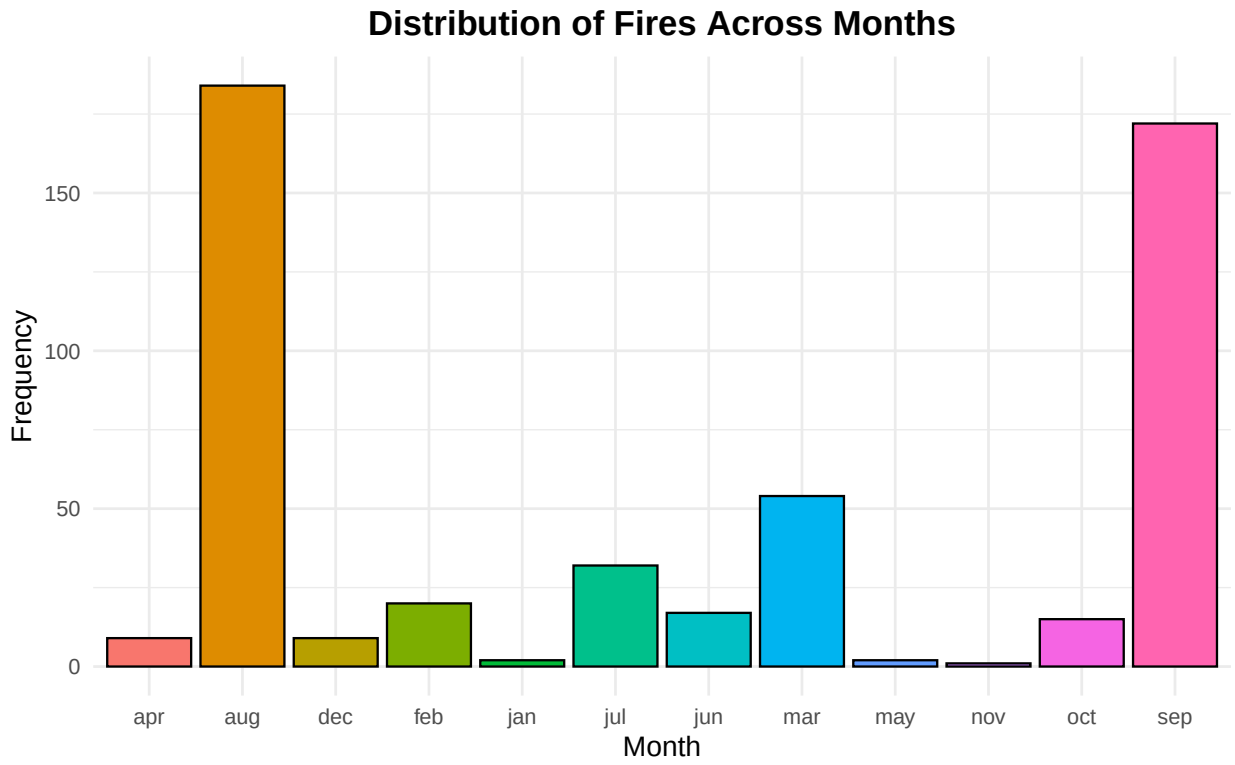
**5.1.2 Chi-Square Goodness-of-Fit Test** The chi-square goodness-of-fit test assesses whether categorical data follow a specified theoretical distribution. Here, it examines whether forest fire occurrences are uniformly distributed across the months. A significant result (small p-value) means fire counts vary significantly by month, implying a seasonal pattern in fire incidence (e.g., more fires in certain months).

- H0: Forest fires are uniformly distributed across months.
- H1: Forest fires are not uniformly distributed across months.




```
##
##  Chi-squared test for given probabilities
##
## data:  month_table
## X-squared = 1072.1, df = 11, p-value < 2.2e-16
```

- p-value < 2.2e-16 : Reject H0




- **Conclusion:** Fire frequency varies significantly by month ($p < 0.001$), revealing strong seasonal patterns with peak fire activity concentrated in specific months.

**5.1.3 One-Sample Kolmogorov-Smirnov Test** The one-sample K–S test compares the empirical distribution of a variable to a theoretical distribution (e.g., normal, exponential, Gumbel). It evaluates whether the observed data could plausibly come from that distribution.

- For burned area, the test checks if it follows a shifted exponential distribution.
- For temperature, relative humidity, and wind, the test examines normality.
- For extreme value tests (Gumbel) on monthly minimum RH and maximum area, the K–S test checks if the extremes follow an Extreme Value (Type I) distribution. This is appropriate for analyzing environmental extremes like droughts or maximum fire sizes.

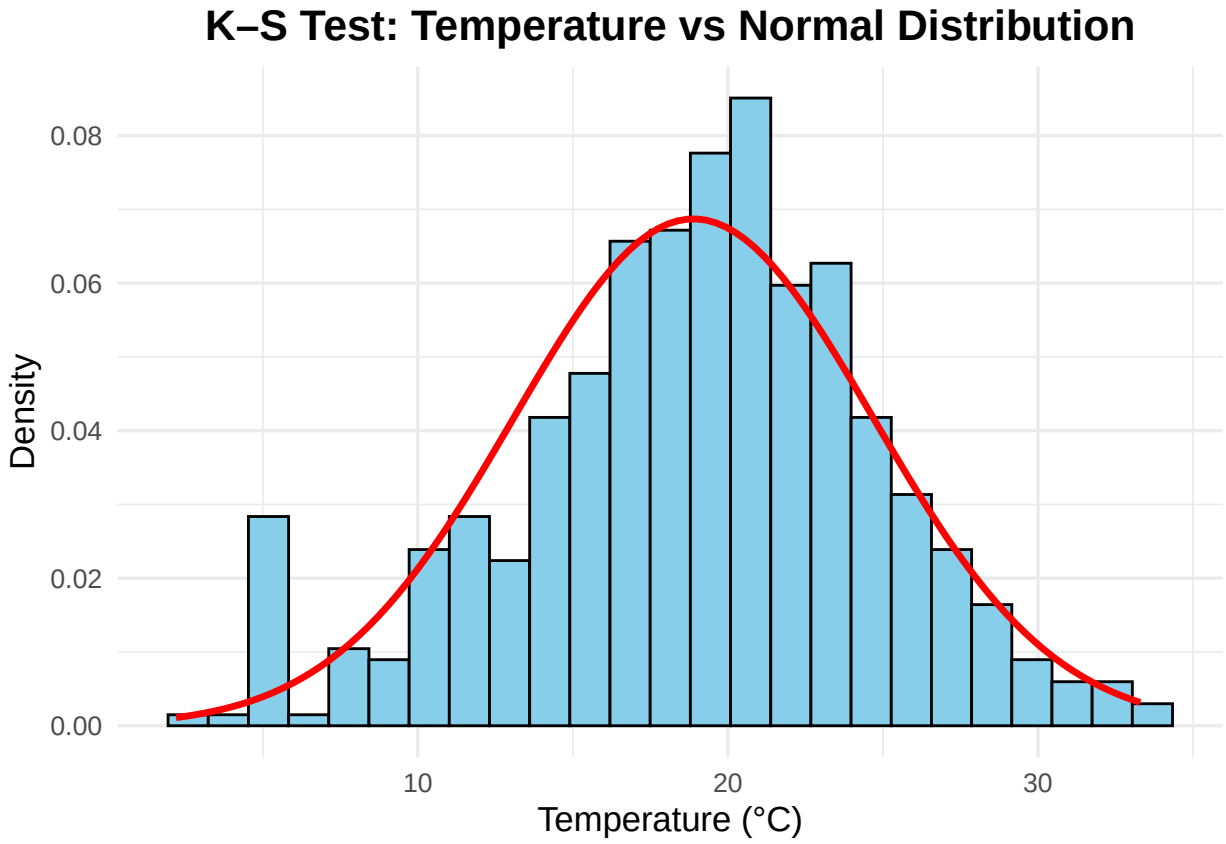
### Burned Area (Shifted Exponential)

- H0: The burnt area follows a shifted exponential distribution
- H1: The burnt area does not follow a shifted exponential distribution

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  area_pos
## D = 0.33614, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

- p-value < 2.2e-16 : Reject H0
- **Conclusion:** Burned area does not follow a shifted exponential distribution.

### Temperature (Normal Distribution)




- H0: Temperature follows a normal distribution.
- H1: Temperature does not follow a normal distribution.

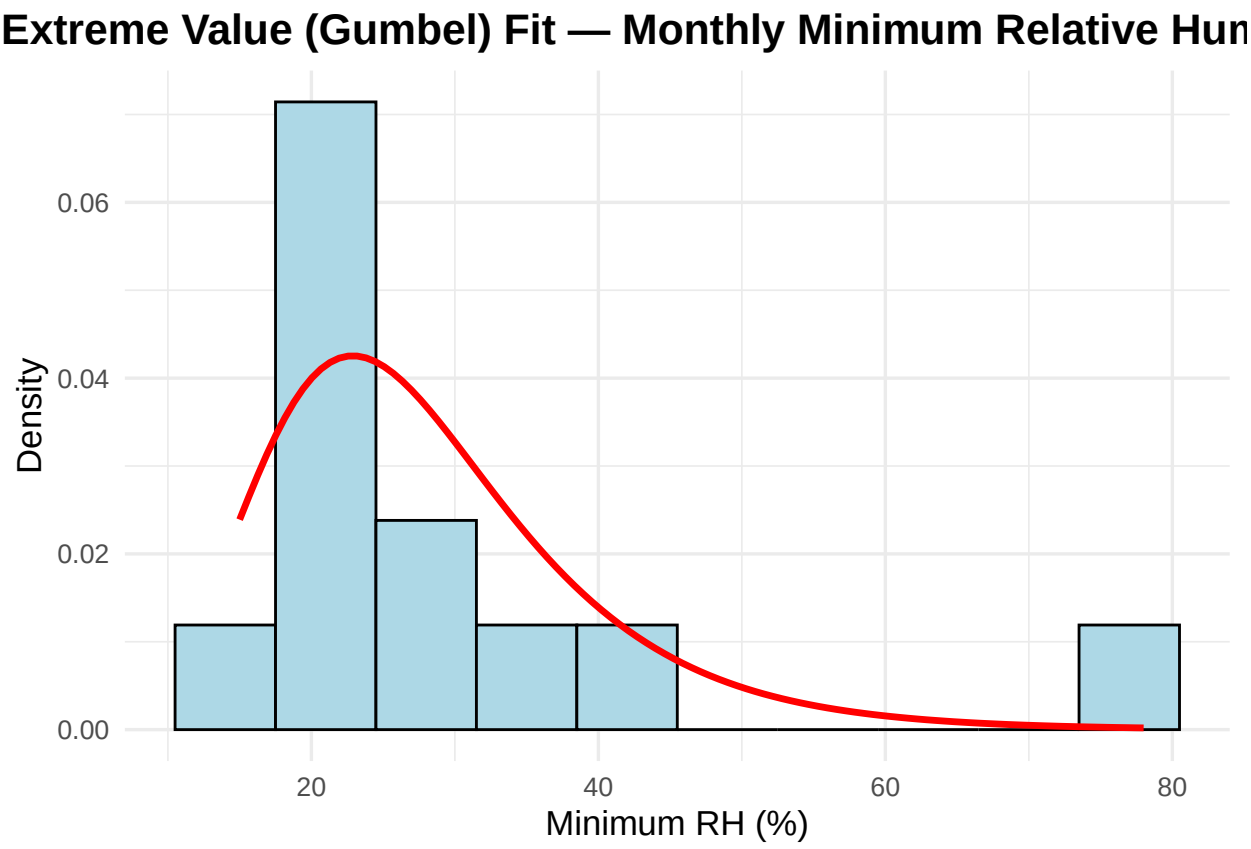
```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  temp
## D = 0.050766, p-value = 0.1392
## alternative hypothesis: two-sided
```

- p-value = 0.1392 : Accept H0
- **Conclusion:** Temperature is consistent with normal distribution (p = 0.139).

### Monthly Minimum RH (Gumbel Distribution)

- H0: Minimum RH values follow a Gumbel distribution.
- H1: Minimum RH values do not follow a Gumbel distribution.

```
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  RH_monthly_min$min_RH
## D = 0.20959, p-value = 0.5962
## alternative hypothesis: two-sided
```



- p-value = 0.5962 : Accept H0
- **Conclusion:** Monthly minimum RH follows Gumbel distribution (p = 0.596), appropriate for extreme value analysis.

**Monthly Maximum Burned Area (Gumbel Distribution)**

- H0: Maximum burnt areas follow a Gumbel distribution
- H1: Maximum burnt areas do not follow a Gumbel distribution

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  area_monthly_max$max_area
## D = 0.39941, p-value = 0.04348
## alternative hypothesis: two-sided
```

- p-value = 0.04348 : Reject H0
- **Conclusion:** Monthly maximum burned area marginally rejects Gumbel fit (p = 0.043).

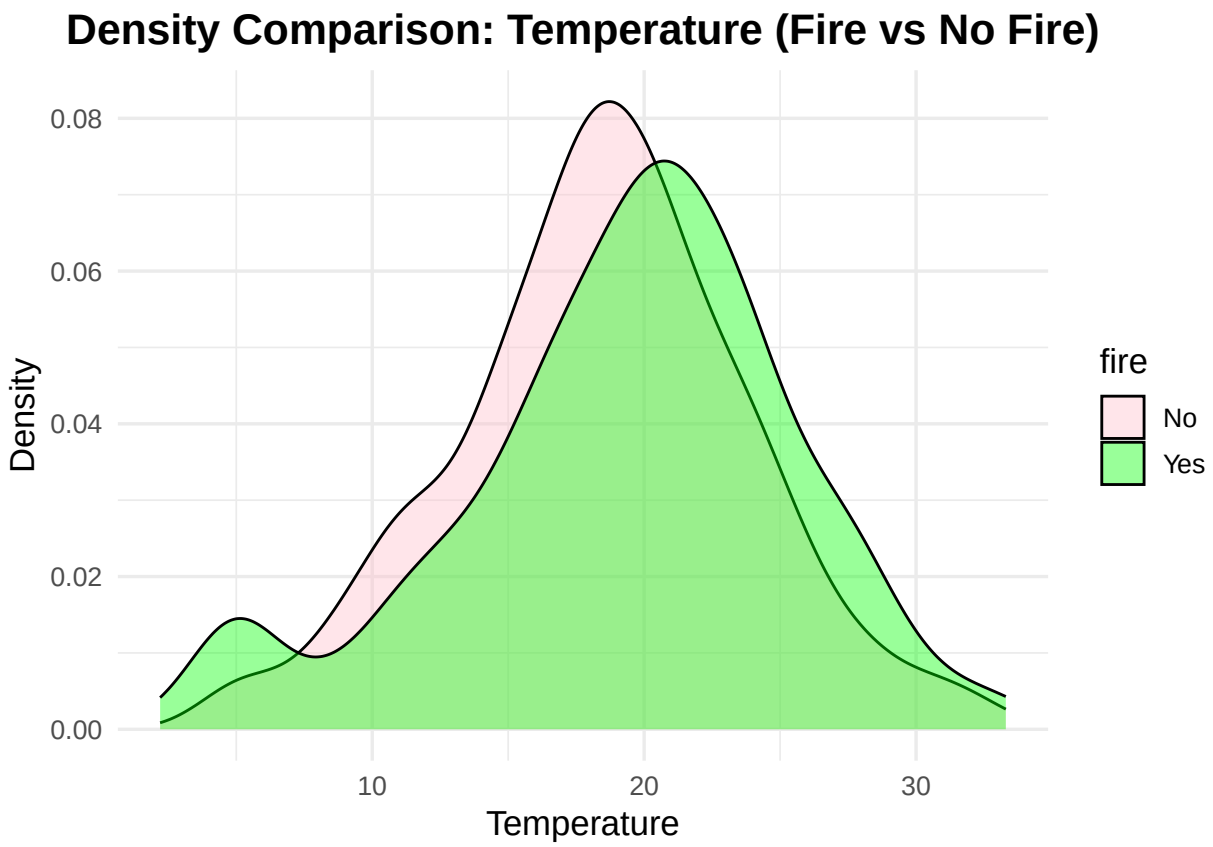
### 5.2 Two-Sample Tests

**5.2.1 Location Tests (Wilcoxon Rank-Sum Test)** The Wilcoxon rank-sum test is a nonparametric alternative to the two-sample t-test. It tests whether the median values of two groups differ, without assuming normality. In the forest fire context, it compares medians of key weather and fire indices across:

- Fire vs No-Fire days (e.g., temperature, DMC)
- August vs September (e.g., DC, wind, temperature)

A significant result indicates differences in median weather conditions, revealing which variables drive fire likelihood and intensity.

**Temperature vs Fire Occurrence**

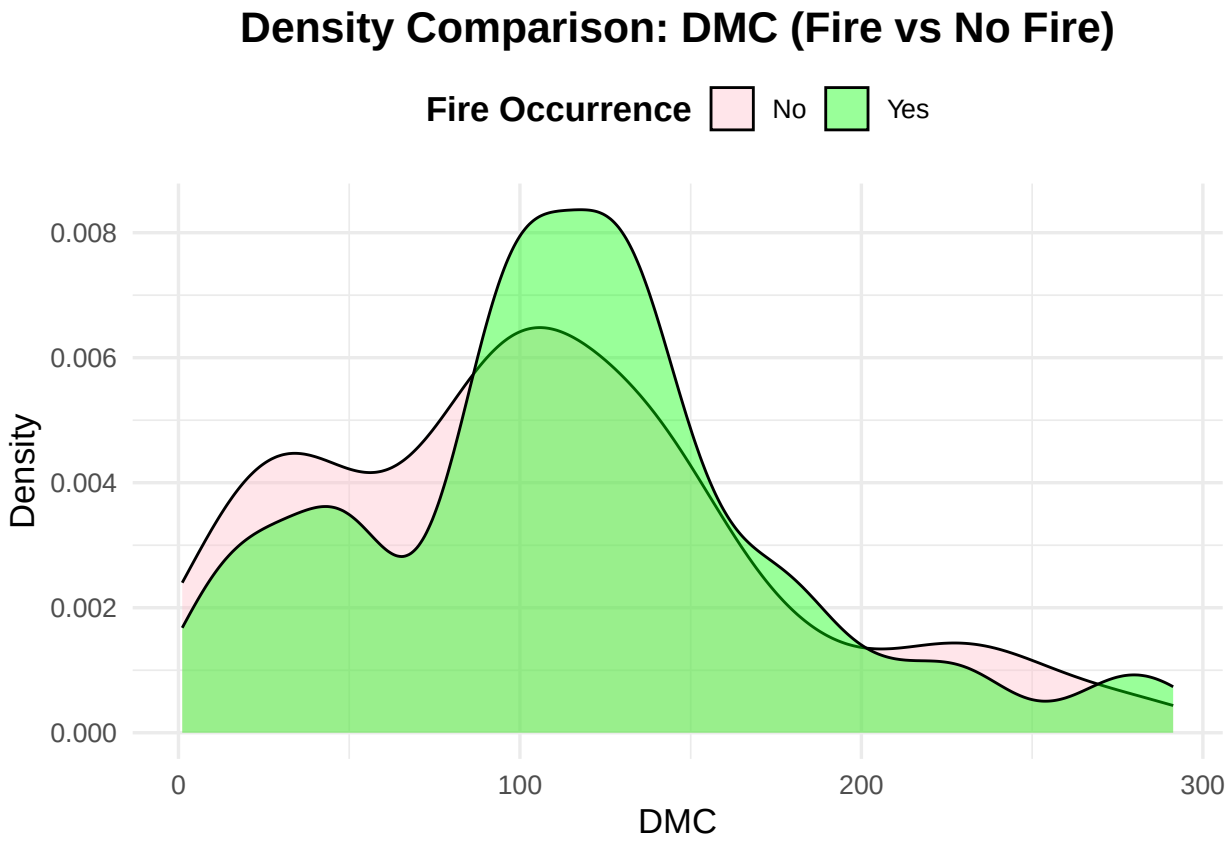



- H0: Median temperature is equal for days with and without burned area
- H1: Median temperature is higher on days with burned area

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  df[df$fire == "Yes", "temp"] and df[df$fire == "No", "temp"]
## W = 37485, p-value = 0.007353
## alternative hypothesis: true location shift is greater than 0
```

- p-value = 0.007353 : Reject H0
- **Conclusion:** Days with burned area have significantly higher median temperatures (p = 0.007), confirming temperature's role in fire risk.

**DMC vs Fire Occurrence**




- H0: Median DMC is equal for days with and without burned area
- H1: Median DMC is higher on days with burned area

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  df[df$fire == "Yes", "DMC"] and df[df$fire == "No", "DMC"]
## W = 36382, p-value = 0.03675
## alternative hypothesis: true location shift is greater than 0
```

- p-value = 0.03675 : Reject H0
- **Conclusion:** Higher median DMC on burned days (p = 0.037) indicates drier duff layer conditions increase fire vulnerability.

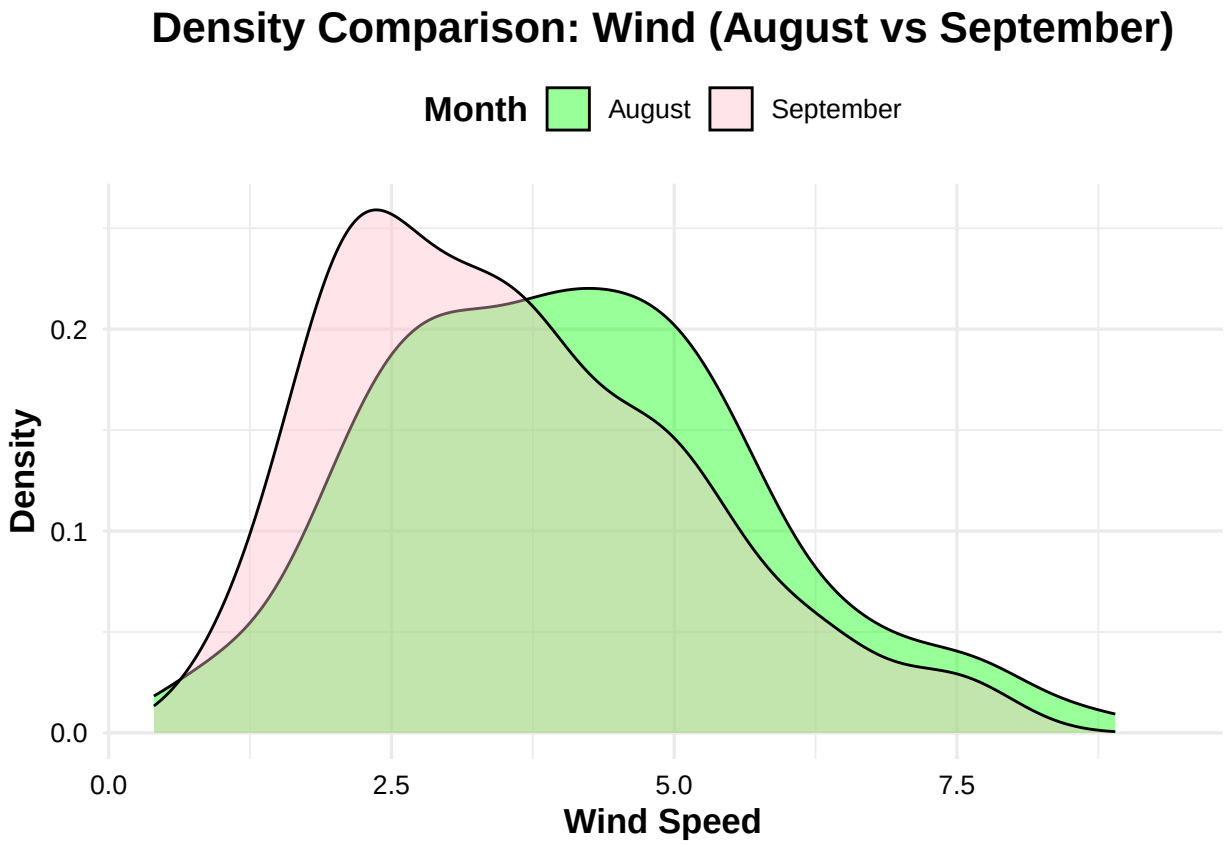
**DC: August vs September**

- H0: Median DC in September <= August
- H1: Median DC in September > August

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  aug_data$DC and sept_data$DC
## W = 3920, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
```

- p-value < 2.2e-16 : Reject H0
- **Conclusion:** September exhibits significantly higher DC ($p < 0.001$), reflecting progressive seasonal drought.

**Wind: August vs September**




- H0: Median wind in August <= September
- H1: Median wind in August > September

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  aug_data$wind and sept_data$wind
## W = 19007, p-value = 0.0004959
## alternative hypothesis: true location shift is greater than 0
```

- p-value = 0.0004959 : Reject H0
- **Conclusion:** August has stronger winds ($p < 0.001$), increasing fire spread potential despite lower drought.

**Temperature: August vs September**

- H0: Median temperature in August <= September
- H1: Median temperature in August > September

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  aug_data$temp and sept_data$temp
## W = 19904, p-value = 1.308e-05
## alternative hypothesis: true location shift is greater than 0
```

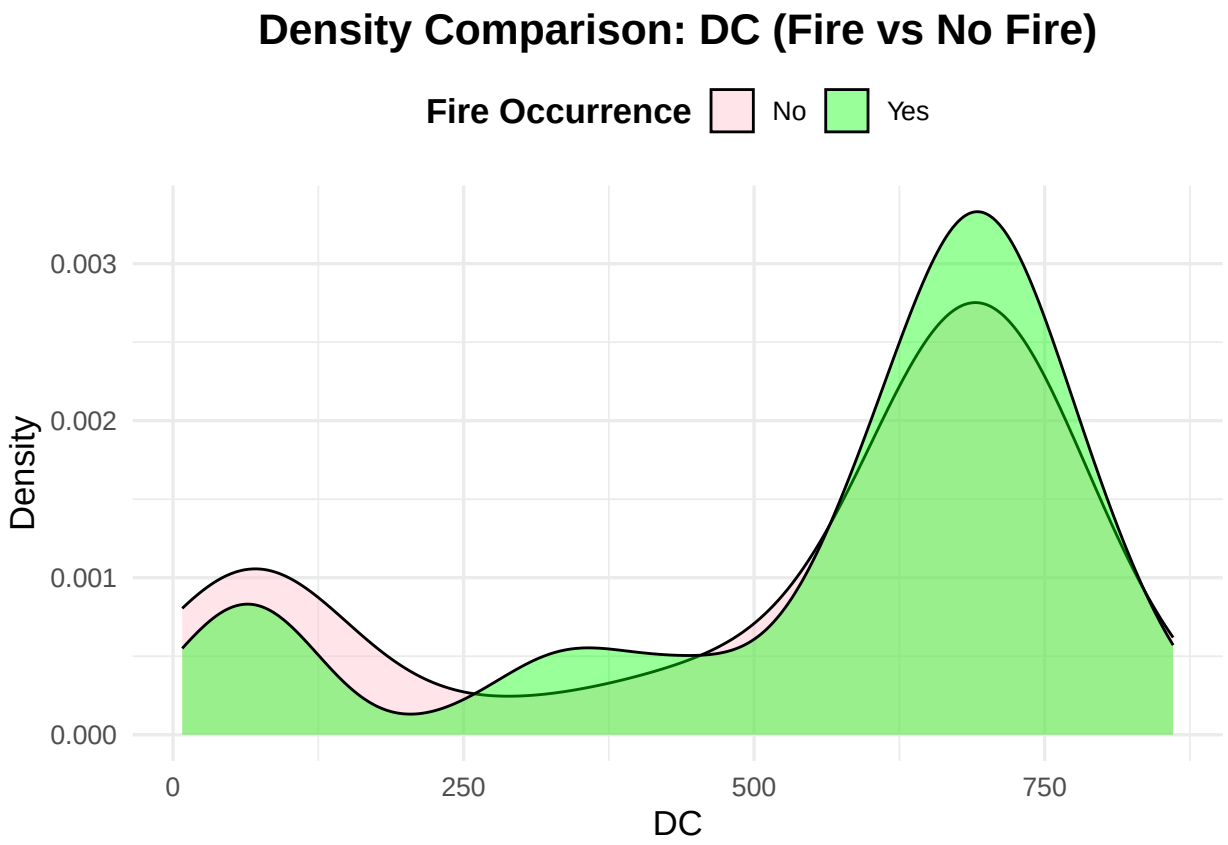
- p-value = 1.308e-05 : Reject H0
- **Conclusion:** August temperatures exceed September ($p < 0.001$), creating distinct monthly fire risk profiles.

**5.2.2 Scale Tests (Mood's Test)** Mood's test examines whether two samples differ in variance (dispersion) rather than location. It tests if the variability of one variable is significantly greater or smaller across two groups. In the forest fire dataset, it's used to check whether the variability of indices (like DC, FFMC, ISI) changes between:

- Fire vs No-Fire days, or
- August vs September.

Significant differences in variability imply that environmental stability or fluctuation may affect fire behavior

**DC Variability: Fire vs No Fire**

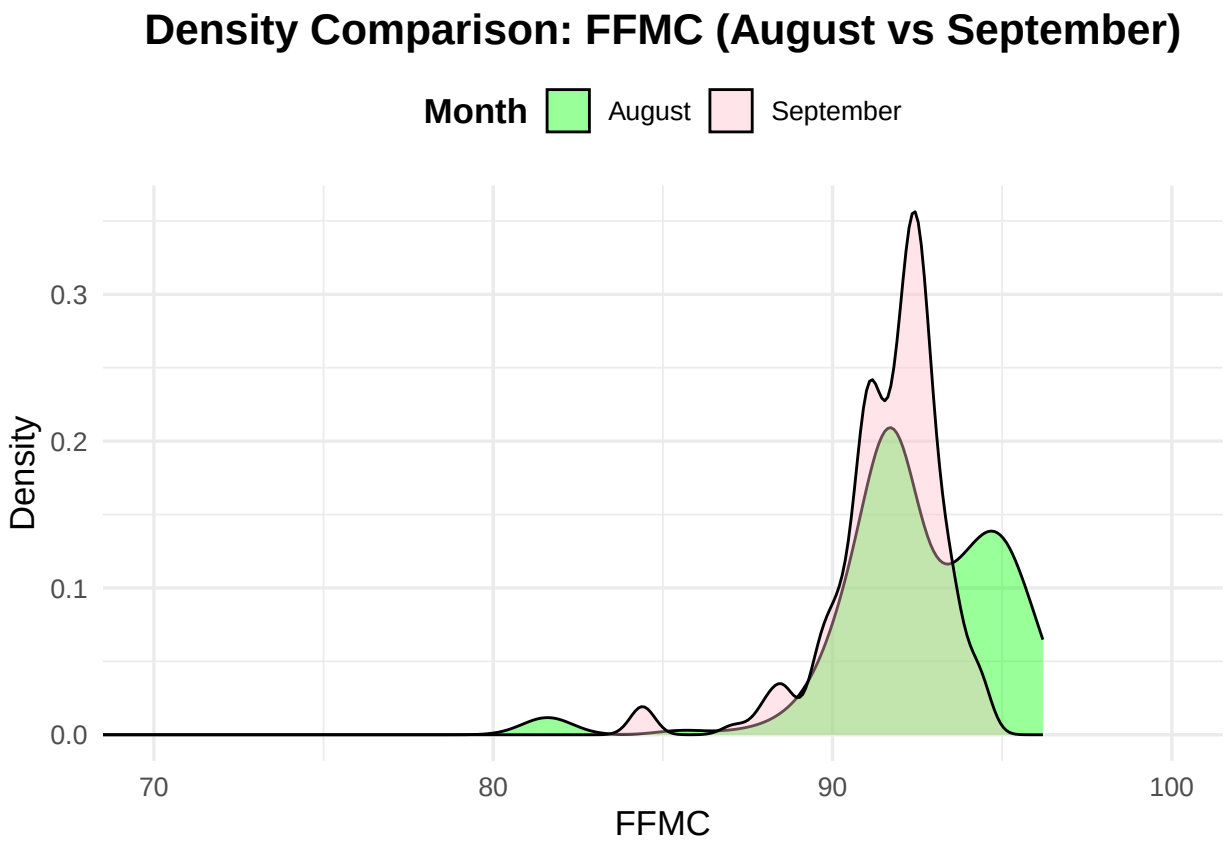



- H0: Variance of DC (Fire) >= Variance of DC (No Fire)
- H1: Variance of DC (Fire) < Variance of DC (No Fire)

```
##
##  Mood two-sample test of scale
##
## data:  df[df$fire == "Yes", "DC"] and df[df$fire == "No", "DC"]
## Z = -0.39851, p-value = 0.3451
## alternative hypothesis: less
```

- p-value = 0.3451 : Fail to reject H0
- **Conclusion:** DC variability is similar regardless of fire occurrence (p = 0.345); drought level matters more than variability.

**FFMC Variability: August vs September**

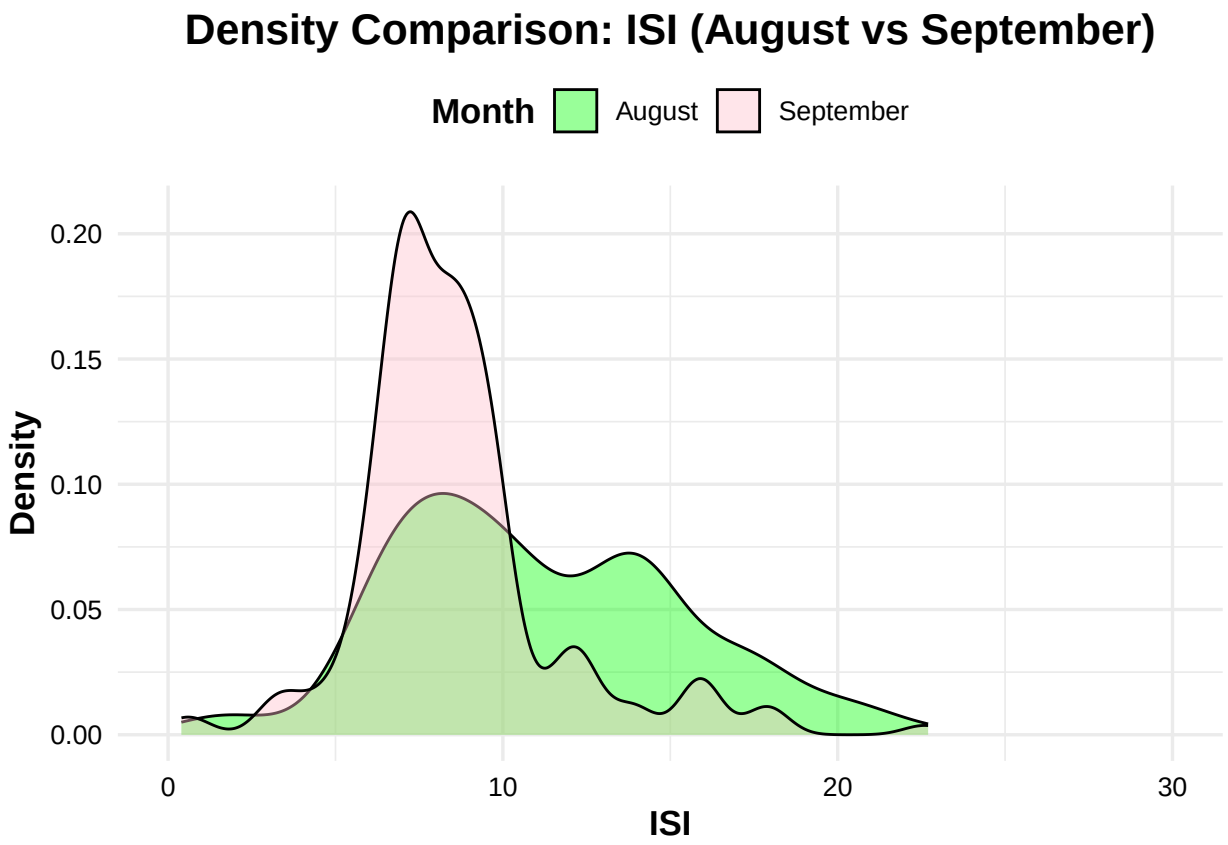



- H0: Variance of FFMC (August) <= Variance of FFMC (September)
- H1: Variance of FFMC (August) > Variance of FFMC (September)

```
##
##  Mood two-sample test of scale
##
## data:  aug_data$FFMC and sept_data$FFMC
## Z = 4.2292, p-value = 1.173e-05
## alternative hypothesis: greater
```

- p-value = 1.173e-05 : Reject H0
- **Conclusion:** August shows greater FFMC variability ($p < 0.001$), creating unpredictable ignition conditions.

**ISI Variability: August vs September**




- H0: Variance of ISI (August) <= Variance of ISI (September)
- H1: Variance of ISI (August) > Variance of ISI (September)

```
##
##  Mood two-sample test of scale
##
## data:  aug_data$ISI and sept_data$ISI
## Z = 2.9844, p-value = 0.001421
## alternative hypothesis: greater
```

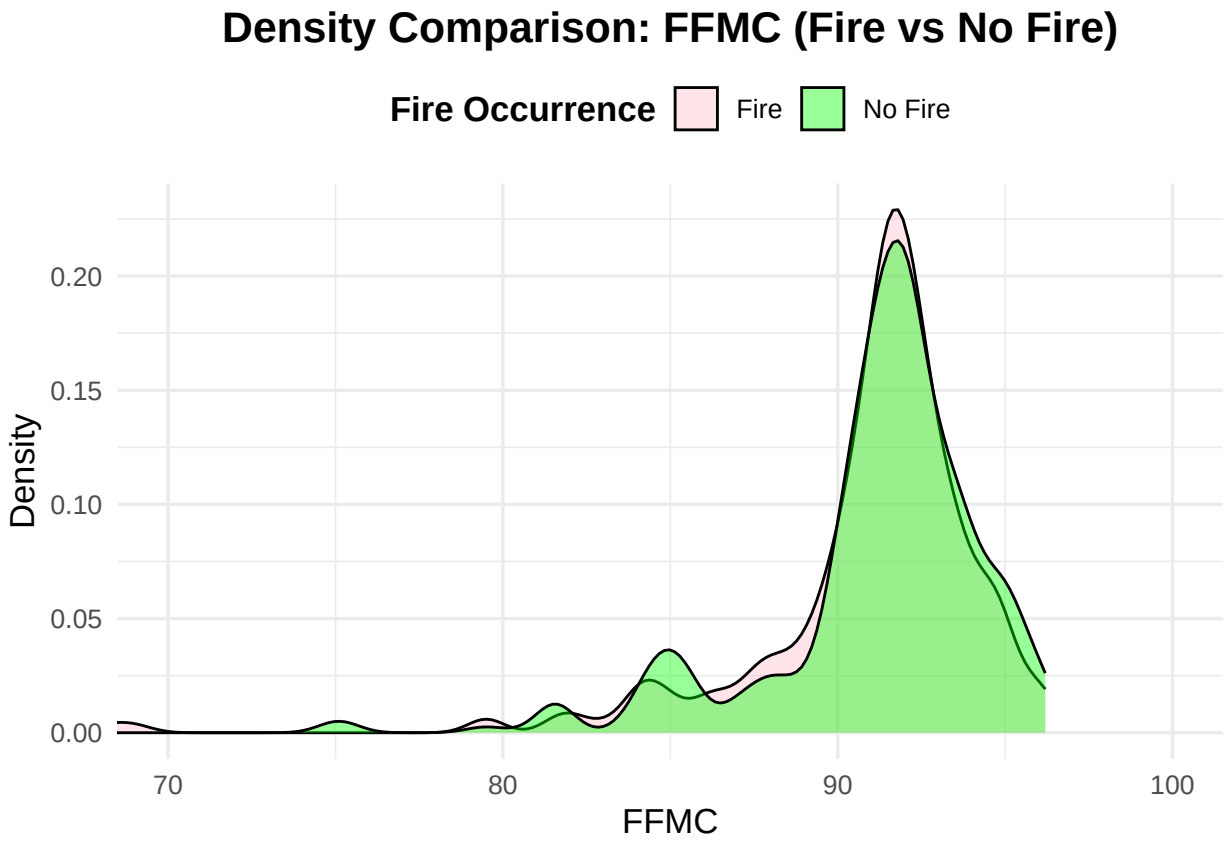
- p-value = 0.001421 : Reject H0
- **Conclusion:** ISI exhibits higher variability in August (p = 0.001), indicating erratic fire spread potential.

**5.2.3 Distribution Tests (Two-Sample KS Test)** The two-sample K–S test compares the entire distributions of two samples to assess if they come from the same population. Unlike Wilcoxon or Mood's tests, which focus on medians or variances, K–S detects any distributional difference (shape, spread, skewness, etc.). Here, it compares distributions of weather indices between:

- Fire vs No-Fire conditions, and
- August vs September.

A non-significant result suggests the overall distributional shape is similar across conditions.

**FFMC: Fire vs No Fire**




- H0: FFMC distributions are identical
- H1: FFMC distributions differ

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  df[df$fire == "Yes", "FFMC"] and df[df$fire == "No", "FFMC"]
## D = 0.05776, p-value = 0.7826
## alternative hypothesis: two-sided
```

- p-value = 0.7826 : Fail to reject H0
- **Conclusion:** FFMC distributions are indistinguishable (p = 0.783); fine fuel moisture alone doesn't differentiate fire days.

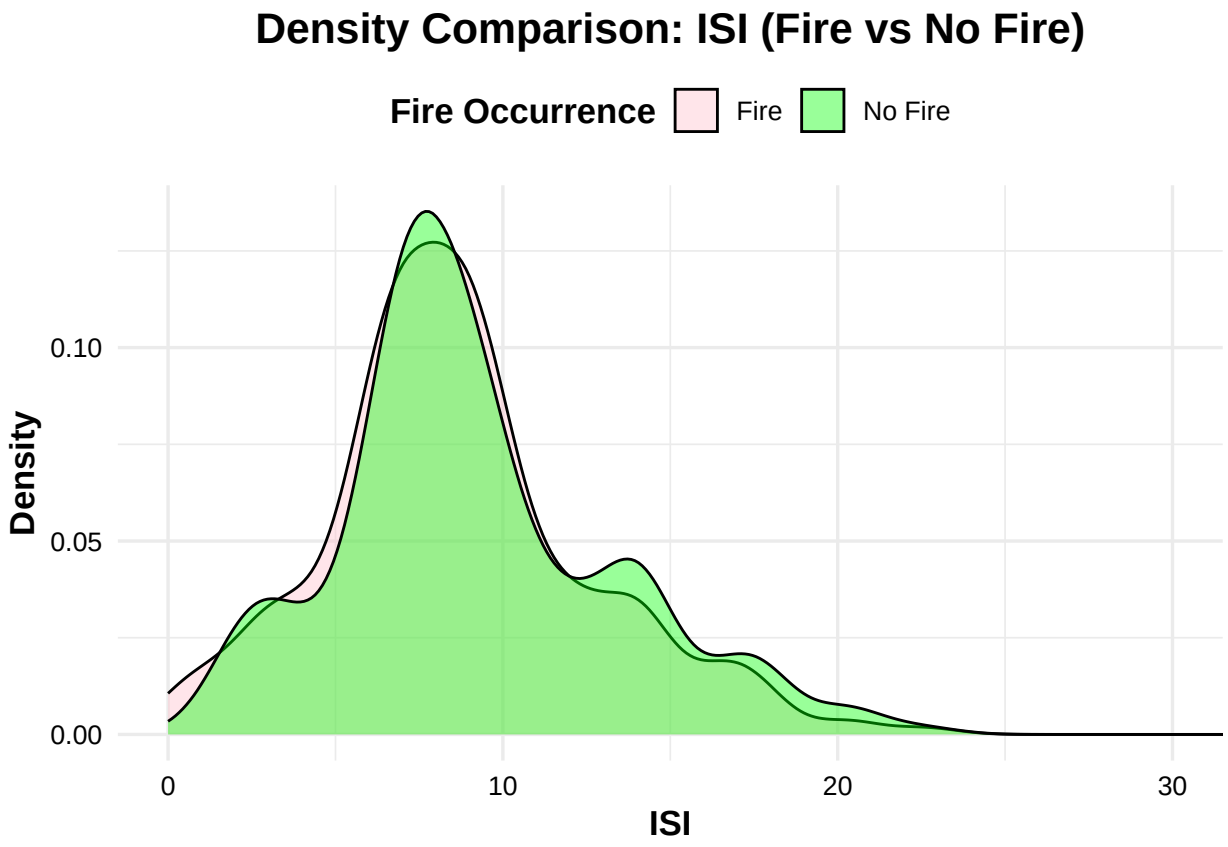
**RH: Fire vs No Fire**

- H0: RH distributions are identical
- H1: RH distributions differ

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  df[df$fire == "Yes", "RH"] and df[df$fire == "No", "RH"]
## D = 0.05746, p-value = 0.788
## alternative hypothesis: two-sided
```

- p-value = 0.788 : Fail to reject H0
- **Conclusion:** RH distributions show no significant difference (p = 0.788).

**ISI: Fire vs No Fire**

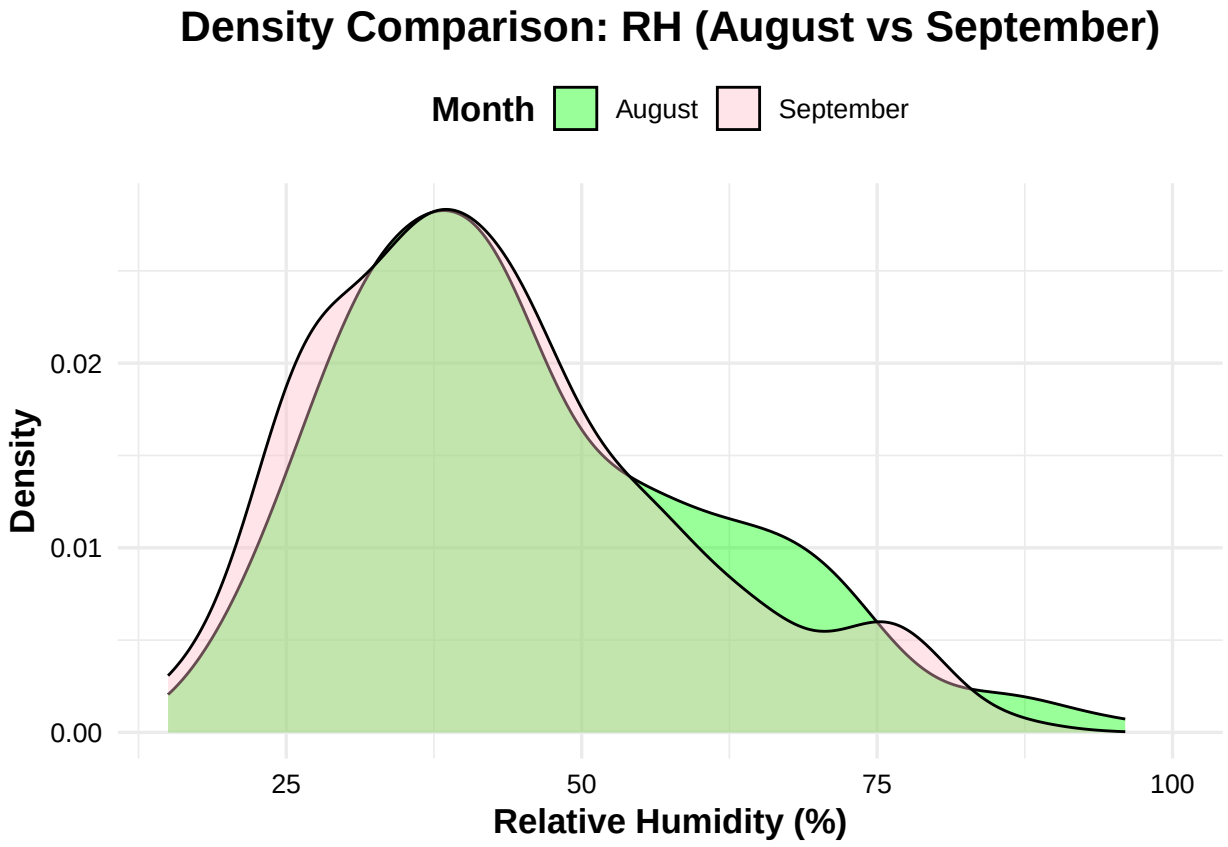


- H0: The distribution of ISI is identical for days with and without burned area
- H1: The distributions differ

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  df[df$fire == "Yes", "ISI"] and df[df$fire == "No", "ISI"]
## D = 0.060489, p-value = 0.7327
## alternative hypothesis: two-sided
```

- p-value = 0.7327 : Fail to reject H0
- **Conclusion:** ISI distributions are similar across fire conditions (p = 0.733).

**RH: August vs September**




- H0: The distribution of relative humidity is identical in August and September
- H1: The distributions differ

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
```

```
##
## data:  aug_data$RH and sept_data$RH
## D = 0.10099, p-value = 0.3249
## alternative hypothesis: two-sided
```

- p-value = 0.3249 : Fail to reject H0
- **Conclusion:** RH patterns remain stable between months (p = 0.325).

**Temperature: August vs September**

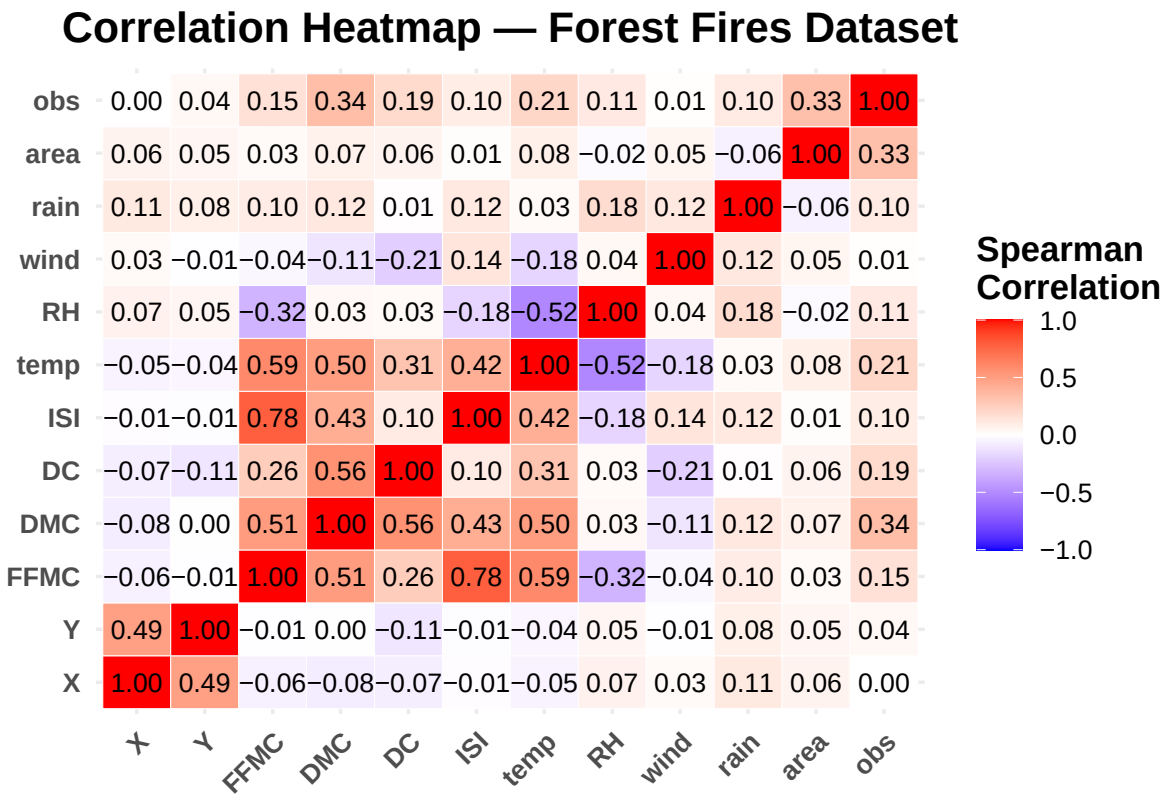
- H0: Temperature distributions are identical
- H1: August temperatures are stochastically lower

```
## Warning in ks.test.default(aug_data$temp, sept_data$temp, alternative =
## "less"): p-value will be approximate in the presence of ties
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  aug_data$temp and sept_data$temp
## D^- = 0.22118, p-value = 0.0001669
## alternative hypothesis: the CDF of x lies below that of y
```

- p-value = 0.0001669 : Reject H0
- **Conclusion:** August temperature distribution is stochastically lower than September ($p < 0.001$), despite higher median.

**5.2.4 Spearman Correlation** Spearman's test measures the strength and direction of a monotonic relationship between two continuous variables without assuming normality. In the forest fire dataset, it checks if weather variables (FFMC, DMC, DC, ISI, temp, RH, wind, rain) are significantly correlated with the burned area (extent of fire). A non-significant correlation indicates that no single variable independently explains burned area variation; instead, fire behavior is likely influenced by multiple interacting factors.



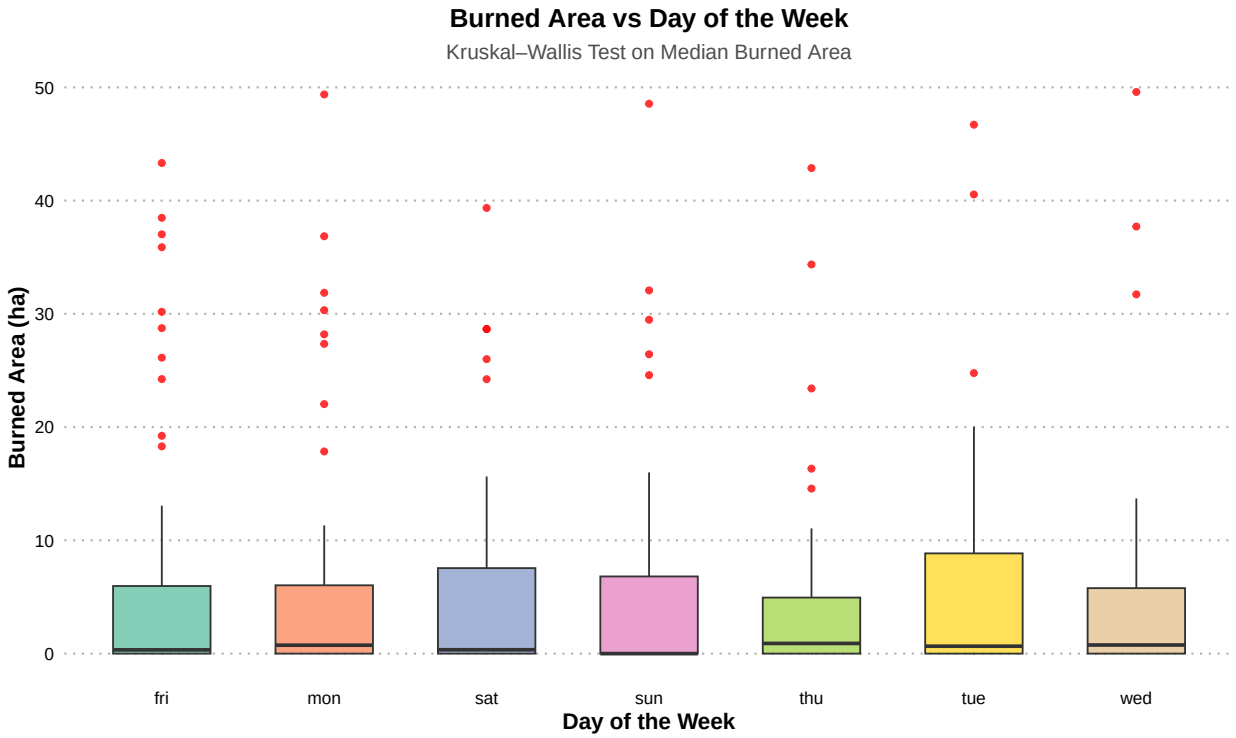

- H0: No monotonic relationship between weather variables and burned area
- H1: Significant monotonic relationship exists
- **Conclusion:** All p-values $> 0.05$ indicate weak monotonic relationships between individual weather indices and burned area extent when fires occur, suggesting complex multivariate interactions govern fire severity.
- FFMC vs area: Weak negative correlation ( = -0.055), not significant (p = 0.364).
- DMC vs area: Very weak positive correlation ( = 0.0076), not significant (p = 0.9005).
- DC vs area: Negligible positive correlation ( = 0.0024), not significant (p = 0.9685).
- ISI vs area: Weak negative correlation ( = -0.110), not significant (p = 0.071).
- Temp vs area: Weak negative correlation ( = -0.061), not significant (p = 0.3148).
- RH vs area: Weak negative correlation ( = -0.053), not significant (p = 0.3896).
- Wind vs area: Weak positive correlation ( = 0.079), not significant (p = 0.1984).
- Rain vs area: Negligible negative correlation ( = -0.0103), not significant (p = 0.8662).

**5.2.5 Kruskal-Wallis Test** The Kruskal–Wallis test is a nonparametric alternative to one-way ANOVA. It evaluates whether the mean values of a continuous variable differ across more than two groups. In this dataset, it tests if the mean burned area varies across days of the week. A non-significant result means fire severity does not depend on the weekday, suggesting that human weekly activity cycles (e.g., workdays vs weekends) do not influence burned area extent.

- H0: Mean burnt area is equal across all weekdays
- H1: At least one weekday has different mean burnt area

```
##
##  Kruskal-Wallis rank sum test
##
## data:  area by day
## Kruskal-Wallis chi-squared = 1.1977, df = 6, p-value = 0.977
```

- p-value $> 0.05$ : Fail to reject H0




- **Conclusion:** No significant difference in median burned area across weekdays ($p > 0.05$), indicating fire severity shows no weekly temporal pattern.

## 6. Conclusion

This nonparametric analysis of the Forest Fires dataset reveals critical insights into temporal, meteorological, and environmental factors influencing forest fire occurrence and severity in Montesinho Natural Park, Portugal.

**Key Findings Summary**

- **Runs Test for Randomness:** All meteorological variables (temperature, RH, wind, burned area, rain) exhibited significant non-randomness ($p < 0.05$), confirming systematic temporal patterns rather than random fluctuations, enabling predictive fire risk forecasting.
- **Chi-Square Goodness-of-Fit Test:** Fire occurrence showed highly significant seasonal variation ($p < 2.2e\text{-}16$), with concentration in August and September, revealing strong seasonal fire risk patterns driven by summer and early autumn climatic conditions.
- **One-Sample Kolmogorov-Smirnov Test:** Temperature follows normal distribution ($p = 0.139$), while burned area rejects shifted exponential distribution ($p < 2.2e\text{-}16$). Monthly minimum RH follows Gumbel distribution ($p = 0.596$), validating extreme value theory for drought analysis.
- **Wilcoxon Rank-Sum Test (Location):** Days with burned area exhibit significantly higher median temperatures ($p = 0.007$) and DMC values ($p = 0.037$), confirming temperature and fuel moisture as critical fire risk indicators. September shows higher DC than August ($p < 2.2e\text{-}16$), while August exhibits stronger winds ($p = 0.0005$).
- **Mood's Test (Scale):** DC variability is similar between fire and no-fire days ($p = 0.345$), indicating drought level matters more than variability. August shows significantly greater FFMC ($p = 1.17e\text{-}05$) and ISI variability ($p = 0.001$) than September, creating unpredictable ignition and erratic fire spread conditions.
- **Two-Sample Kolmogorov-Smirnov Test:** FFMC, RH, ISI, and wind distributions remain similar between fire and no-fire conditions (all $p > 0.05$), suggesting these variables alone do not fundamentally alter fire likelihood. Temperature distribution differs between August and September ($p = 0.0002$) despite similar medians.
- **Spearman Correlation:** All weather indices showed weak and non-significant correlations with burned area when fires occur (all $p > 0.05$), indicating fire severity is governed by complex multivariate interactions rather than single dominant factors, requiring integrated predictive models.
- **Kruskal-Wallis Test:** No significant difference in median burned area across weekdays ($p > 0.05$), indicating fire severity shows no weekly temporal pattern and that natural meteorological cycles dominate over human activity rhythms.

### Implications and Future Directions

- **Key Implications:** (1) Temperature and DMC monitoring effectively identifies high-risk fire days; (2) August requires focus on volatile wind-driven risks while September demands drought stress management; (3) Multivariate models are essential as single-variable approaches fail; (4) Extreme value analysis enables worst-case scenario preparation; (5) Non-random temporal patterns enable short-term forecasting.
- **Future Research:** Machine learning approaches should capture complex multivariate interactions, threshold effects where condition combinations trigger extreme behavior should be investigated, and spatial analysis techniques must account for topographic and vegetation influences.
- **Final Remarks:** Forest fire behavior in Montesinho Natural Park is governed by complex temporal patterns, seasonal variations, and multivariate interactions. While individual weather variables show predictable patterns and significant seasonal differences, no single factor explains fire severity. Effective fire management requires integrated monitoring systems considering multiple environmental factors simultaneously, with particular attention to temperature, fuel moisture, seasonal drought accumulation, and wind conditions during peak fire months (August and September).

## 7. Appendix

```
library(tseries)
library(ggplot2)
library(tidyr)
library(dplyr)
library(stats)
library(evd)
library(fitdistrplus)
library(corrplot)

df <- read.csv('forestfires.csv')
attach(df)
random_sample <- df[sample(1:nrow(df), 7, replace=FALSE), ]
random_sample

summary(df)

plot_data <- df %>%
  dplyr::select(temp, RH, wind, area, rain) %>%
  mutate(obs = 1:n()) %>%
  tidyr::pivot_longer(cols = c(temp, RH, wind, area, rain),
```

```
                      names_to = "Variable",
                      values_to = "Value")

ggplot(plot_data, aes(x = obs, y = Value, color = Variable)) +
  geom_line(size = 0.9) +
  facet_wrap(~Variable, scales = "free_y", ncol = 1) +
  theme_minimal(base_size = 13) +
  labs(
    title = "Forest Fire Dataset Variables Over Observations",
    x = "Observation Index",
    y = "Value"
  ) +
  theme(
    legend.position = "none",
    plot.title = element_text(hjust = 0.5, face = "bold"),
    strip.text = element_text(face = "bold")
  )

df['fire'] <- ifelse(df$area == 0, 'No', 'Yes')

# Create monthly subsets
sept_data <- df[df$month == 'sep', ]
aug_data <- df[df$month == 'aug', ]

df$obs <- 1:nrow(df)

ggplot(df, aes(x = obs, y = temp)) +
  geom_line(color = "darkgreen", linewidth = 0.8, alpha = 0.7) +
  geom_smooth(method = "loess", se = FALSE, color = "orange", linewidth = 1) +
  labs(
    title = "Temperature Trend Over Observations",
    x = "Observation Index",
    y = "Temperature (\u00B0C)"
  ) +
  theme_minimal(base_size = 13) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))


temp_binary <- ifelse(temp > median(temp, na.rm = TRUE), 1, 0)
runs.test(as.factor(temp_binary))

ggplot(df, aes(x = obs, y = RH)) +
  geom_line(color = "darkgreen", linewidth = 0.8, alpha = 0.7) +
  geom_smooth(method = "loess", se = FALSE, color = "orange", linewidth = 1) +
```

```
  labs(
    title = "Relative Humidity Trend",
    x = "Observation Index",
    y = "RH"
  ) +
  theme_minimal(base_size = 13) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))

RH_binary <- ifelse(RH > median(RH, na.rm = TRUE), 1, 0)
runs.test(as.factor(RH_binary))

ggplot(df, aes(x = obs, y = wind)) +
  geom_line(color = "darkgreen", linewidth = 0.8, alpha = 0.7) +
  geom_smooth(method = "loess", se = FALSE, color = "orange", linewidth = 1) +
  labs(
    title = "Wind Trend Over Observations",
    x = "Observation Index",
    y = "Wind"
  ) +
  theme_minimal(base_size = 13) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))

wind_binary <- ifelse(wind > median(wind, na.rm = TRUE), 1, 0)
runs.test(as.factor(wind_binary))

area_binary <- ifelse(df$area > median(df$area, na.rm = TRUE), 1, 0)
runs.test(as.factor(area_binary))

rain_binary <- ifelse(rain > 0, 1, 0)
runs.test(as.factor(rain_binary))

area_pos <- df$area[df$area > 0]

# Estimate shift (minimum observed area)
x0 <- min(area_pos)

# Estimate rate parameter for shifted exponential
lambda_hat <- 1 / (mean(area_pos) - x0)

# Define shifted exponential CDF
pexp_shifted <- function(x) {
  ifelse(x < x0, 0, pexp(x - x0, rate = lambda_hat))
}
ggplot(data.frame(area_pos), aes(x = area_pos)) +
```

```r
  geom_histogram(aes(y = ..density..), bins = 30, fill = "lightgreen", color = "black") +
  stat_function(fun = function(x) dexp(x - x0, rate = lambda_hat),
                color = "red", size = 1.2) +
  labs(title = "Shifted Exponential Fit to Burnt Area",
       x = "Burnt Area (ha)", y = "Density") +
  theme_minimal(base_size = 13) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))

month_table <- table(month)
chisq.test(month_table)

ggplot(data.frame(month = names(month_table), freq = as.numeric(month_table)),
       aes(x = month, y = freq, fill = month)) +
  geom_bar(stat = "identity", color = "black") +
  theme_minimal(base_size = 13) +
  labs(title = "Distribution of Fires Across Months",
       x = "Month", y = "Frequency") +
  theme(legend.position = "none",
        plot.title = element_text(hjust = 0.5, face = "bold"))

area_pos <- df$area[df$area > 0]
x0 <- min(area_pos)
lambda_hat <- 1 / (mean(area_pos) - x0)
pexp_shifted <- function(x) {
  ifelse(x < x0, 0, pexp(x - x0, rate = lambda_hat))
}
ks.test(area_pos, pexp_shifted)


ggplot(df, aes(x = temp)) +
  geom_histogram(aes(y = ..density..), bins = 25, fill = "skyblue", color = "black") +
  stat_function(fun = dnorm,
                args = list(mean = mean(temp, na.rm = TRUE),
                            sd = sd(temp, na.rm = TRUE)),
                color = "red", size = 1.2) +
  theme_minimal(base_size = 13) +
  labs(
    title = "K-S Test: Temperature vs Normal Distribution",
    x = "Temperature (\u00B0C)",
    y = "Density"
  ) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```

```r
ks.test(temp, "pnorm", mean(temp, na.rm = TRUE), sd(temp, na.rm = TRUE))

RH_monthly_min <- df %>%
  group_by(month) %>%
  summarise(min_RH = min(RH, na.rm = TRUE))

fit_gumbel_RH <- fgev(RH_monthly_min$min_RH, shape = 0)
mu_RH <- fit_gumbel_RH$estimate["loc"]
sigma_RH <- fit_gumbel_RH$estimate["scale"]

pgumbel <- function(x, mu, sigma) exp(-exp(-(x - mu) / sigma))
ks.test(RH_monthly_min$min_RH, pgumbel, mu_RH, sigma_RH)

ggplot(RH_monthly_min, aes(x = min_RH)) +
  geom_histogram(aes(y = ..density..),
                 bins = 10, fill = "lightblue", color = "black") +
  stat_function(fun = function(x) dgev(x, loc = mu_RH, scale = sigma_RH, shape = 0),
                color = "red", size = 1.2) +
  labs(title = "Extreme Value (Gumbel) Fit - Monthly Minimum Relative Humidity",
       x = "Minimum RH (%)",
       y = "Density") +
  theme_minimal(base_size = 13) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))

area_monthly_max <- df %>%
  group_by(month) %>%
  summarise(max_area = max(area, na.rm = TRUE))

fit_gumbel_area <- fgev(area_monthly_max$max_area, shape = 0)
mu <- fit_gumbel_area$estimate["loc"]
sigma <- fit_gumbel_area$estimate["scale"]

ks.test(area_monthly_max$max_area, pgumbel, mu, sigma)

ggplot(df, aes(x = temp, fill = fire)) +
  geom_density(alpha = 0.4, color = "black") +
  scale_fill_manual(values = c("Yes" = "green", "No" = "pink")) +
  labs(title = "Density Comparison: Temperature (Fire vs No Fire)",
       x = "Temperature", y = "Density") +
  theme_minimal(base_size = 13) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))

wilcox.test(df[df$fire == 'Yes', 'temp'], df[df$fire == 'No', 'temp'],
            alternative = 'greater')
```

```r
ggplot(df, aes(x = DMC, fill = fire)) +
  geom_density(alpha = 0.4, color = "black") +
  scale_fill_manual(values = c("Yes" = "green", "No" = "pink")) +
  labs(title = "Density Comparison: DMC (Fire vs No Fire)",
       x = "DMC",
       y = "Density",
       fill = "Fire Occurrence") +
  theme_minimal(base_size = 13) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    legend.title = element_text(face = "bold"),
    legend.position = "top"
  )


wilcox.test(df[df$fire == 'Yes', 'DMC'], df[df$fire == 'No', 'DMC'],
            alternative = 'greater')

sept_data <- df[df$month == 'sep', ]
aug_data <- df[df$month == 'aug', ]

wilcox.test(aug_data$DC, sept_data$DC, alternative = 'less')

ggplot(df %>% filter(month %in% c("aug", "sep")),
       aes(x = wind, fill = month)) +
  geom_density(alpha = 0.4, color = "black") +
  scale_fill_manual(values = c("aug" = "green", "sep" = "pink"),
                    labels = c("August", "September")) +
  coord_cartesian(xlim = range(df$wind, na.rm = TRUE)) +
  labs(
    title = "Density Comparison: Wind (August vs September)",
    x = "Wind Speed",
    y = "Density",
    fill = "Month"
  ) +
  theme_minimal(base_size = 13) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    legend.title = element_text(face = "bold"),
    legend.position = "top",
    axis.title.x = element_text(face = "bold"),
    axis.title.y = element_text(face = "bold"),
    axis.text = element_text(color = "black")
  )
```

```r
wilcox.test(aug_data$wind, sept_data$wind, alternative = 'greater')

wilcox.test(aug_data$temp, sept_data$temp, alternative = 'greater')


ggplot(df, aes(x = DC, fill = fire)) +
  geom_density(alpha = 0.4, color = "black") +
  scale_fill_manual(values = c("Yes" = "green", "No" = "pink")) +
  labs(title = "Density Comparison: DC (Fire vs No Fire)",
       x = "DC",
       y = "Density",
       fill = "Fire Occurrence") +
  theme_minimal(base_size = 13) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    legend.title = element_text(face = "bold"),
    legend.position = "top"
  )


mood.test(df[df$fire == 'Yes', 'DC'], df[df$fire == 'No', 'DC'],
          alternative = 'less')

ggplot(df %>% filter(month %in% c("aug", "sep")),
       aes(x = FFMC, fill = month)) +
  geom_density(alpha = 0.4, color = "black") +
  scale_fill_manual(values = c("aug" = "green", "sep" = "pink"),
                    labels = c("August", "September")) +
  coord_cartesian(xlim = c(70,100)) +
  labs(title = "Density Comparison: FFMC (August vs September)",
       x = "FFMC",
       y = "Density",
       fill = "Month") +
  theme_minimal(base_size = 13) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    legend.title = element_text(face = "bold"),
    legend.position = "top"
  )



mood.test(aug_data$FFMC, sept_data$FFMC, alternative = 'greater')
```

```r
ggplot(df %>% filter(month %in% c("aug", "sep")),
       aes(x = ISI, fill = month)) +
  geom_density(alpha = 0.4, color = "black") +
  scale_fill_manual(values = c("aug" = "green", "sep" = "pink"),
                    labels = c("August", "September")) +
  coord_cartesian(xlim = c(0,30)) +
  labs(
    title = "Density Comparison: ISI (August vs September)",
    x = "ISI",
    y = "Density",
    fill = "Month"
  ) +
  theme_minimal(base_size = 13) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    legend.title = element_text(face = "bold"),
    legend.position = "top",
    axis.title.x = element_text(face = "bold"),
    axis.title.y = element_text(face = "bold"),
    axis.text = element_text(color = "black")
  )


mood.test(aug_data$ISI, sept_data$ISI, alternative = 'greater')

ggplot(df, aes(x = FFMC, fill = fire)) +
  geom_density(alpha = 0.4, color = "black") +
  scale_fill_manual(values = c("Yes" = "green", "No" = "pink"),
                    labels = c("Fire", "No Fire")) +
  coord_cartesian(xlim = c(70, 100)) +   # <-- Focus on main data range
  labs(title = "Density Comparison: FFMC (Fire vs No Fire)",
       x = "FFMC",
       y = "Density",
       fill = "Fire Occurrence") +
  theme_minimal(base_size = 13) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    legend.title = element_text(face = "bold"),
    legend.position = "top"
  )


ks.test(df[df$fire == 'Yes', 'FFMC'], df[df$fire == 'No', 'FFMC'])
```

```r
ks.test(df[df$fire == 'Yes', 'RH'], df[df$fire == 'No', 'RH'])

ggplot(df, aes(x = ISI, fill = fire)) +
  geom_density(alpha = 0.4, color = "black") +
  scale_fill_manual(values = c("Yes" = "green", "No" = "pink"),
                    labels = c("Fire", "No Fire")) +
  coord_cartesian(xlim = c(0,30)) +
  labs(
    title = "Density Comparison: ISI (Fire vs No Fire)",
    x = "ISI",
    y = "Density",
    fill = "Fire Occurrence"
  ) +
  theme_minimal(base_size = 13) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    legend.title = element_text(face = "bold"),
    legend.position = "top",
    axis.title.x = element_text(face = "bold"),
    axis.title.y = element_text(face = "bold"),
    axis.text = element_text(color = "black")
  )


ks.test(df[df$fire == 'Yes', 'ISI'], df[df$fire == 'No', 'ISI'])

ggplot(df %>% filter(month %in% c("aug", "sep")),
       aes(x = RH, fill = month)) +
  geom_density(alpha = 0.4, color = "black") +
  scale_fill_manual(values = c("aug" = "green", "sep" = "pink"),
                    labels = c("August", "September")) +
  coord_cartesian(xlim = range(df$RH, na.rm = TRUE)) +
  labs(
    title = "Density Comparison: RH (August vs September)",
    x = "Relative Humidity (%)",
    y = "Density",
    fill = "Month"
  ) +
  theme_minimal(base_size = 13) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    legend.title = element_text(face = "bold"),
    legend.position = "top",
    axis.title.x = element_text(face = "bold"),
```

```
    axis.title.y = element_text(face = "bold"),
    axis.text = element_text(color = "black")
  )


ks.test(aug_data$RH, sept_data$RH)

ks.test(aug_data$temp, sept_data$temp, alternative = 'less')

# Select numeric columns
num_vars <- df %>%
  select_if(is.numeric)

# Compute Spearman correlation matrix
corr_matrix <- cor(num_vars, method = "spearman", use = "complete.obs")

# Convert to long format for ggplot
corr_data <- as.data.frame(as.table(corr_matrix))

# Create heatmap with values
ggplot(corr_data, aes(Var1, Var2, fill = Freq)) +
  geom_tile(color = "white") +
  geom_text(aes(label = sprintf("%.2f", Freq)), color = "black", size = 3.5) +
  scale_fill_gradient2(
    low = "blue", mid = "white", high = "red", midpoint = 0,
    limits = c(-1, 1),
    name = "Spearman\nCorrelation"
  ) +
  labs(
    title = "Correlation Heatmap - Forest Fires Dataset",
    x = "", y = ""
  ) +
  theme_minimal(base_size = 13) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"),
    axis.text.y = element_text(face = "bold"),
    plot.title = element_text(hjust = 0.5, face = "bold"),
    legend.title = element_text(face = "bold"),
    legend.position = "right"
  )



data_z <- df[df$area != 0, ]
```

```r
X <- c("FFMC", "DMC", "DC", "ISI", "temp", "RH", "wind", "rain")

for (x in X) {
  print(paste(x, "vs area"))
  r <- cor.test(data_z[[x]], data_z$area, method = "spearman")
  print(r)
}

kruskal.test(area ~ day, data = df)

ggplot(df, aes(x = day, y = area, fill = day)) +
  geom_boxplot(outlier.color = "red", alpha = 0.8, width = 0.6) +
  scale_fill_brewer(palette = "Set2") +
  coord_cartesian(ylim = c(0, quantile(df$area, 0.95, na.rm = TRUE))) +  # focus on mai
  labs(
    title = "Burned Area vs Day of the Week",
    subtitle = "Kruskal-Wallis Test on Median Burned Area",
    x = "Day of the Week",
    y = "Burned Area (ha)"
  ) +
  theme_minimal(base_size = 13) +
  theme(
    legend.position = "none",
    plot.title = element_text(hjust = 0.5, face = "bold", size = 15),
    plot.subtitle = element_text(hjust = 0.5, size = 12, color = "gray30"),
    axis.title.x = element_text(face = "bold"),
    axis.title.y = element_text(face = "bold"),
    axis.text = element_text(color = "black"),
    panel.grid.minor = element_blank(),
    panel.grid.major.x = element_blank(),
    panel.grid.major.y = element_line(linetype = "dotted", color = "gray70")
  )
```