COSC 757 Assignment 1

Algerian Forest Fires

Ann Pham
Department of Marketing
Towson University
Towson, MD
npham6@students.towson.edu

Abstract

The study applies exploratory data analysis to understand the Algerian Forest Fires dataset's characteristics, including descriptive statistics and data visualizations. The dataset's continuous attributes underwent preprocessing techniques such as min-max normalization, z-scores, and decimal scaling, providing insights into various normalization methods. Additionally, a continuous variable was discretized using equal-width and equal-frequency binning methods, demonstrating the versatility of data transformation techniques.

The study also investigates non-normally distributed variables through natural log, square root, and inverse square root transformations. Based on the EDA findings, regression analysis will be conducted to answer if we can predict the Fire Weather Index (FWI) based on weather-related variables such as Temperature, Relative Humidity (RH), Wind Speed (Ws), Rainfall (Rain). The results of this study provide valuable insights to understand Algerian Forest Fires further.

Introduction

Wildfires, uncontrolled fires that burn in wildland are a global phenomenon. They can occur in various ecosystems, including forests, grasslands, and savannas[2]. These fires can have significant impacts on the environment and human health. Understanding the factors influencing these fires is vital for preventing and managing them effectively.

This paper focuses on Algerian Forest Fires Dataset[1] from the UCI Machine Learning Repository. The dataset contains observations of forest fires in two regions of Algeria: Bejaia and Sidi Bel-Abbes, from June to September 2012. Previous studies have conducted exploratory data analysis and implemented machine learning algorithms on this dataset.

The objective of this study is to further explore and analyze the dataset using various statistical techniques in R. The study is divided into three main parts: Exploratory Data Analysis (EDA), Data Preprocessing, and Regression Analysis.

This paper aims to provide valuable insights into the Algerian forest fires dataset and demonstrate the effectiveness of various data analysis techniques in understanding and predicting outcomes from complex datasets.

Dataset information

I combines data from two regions in Algeria: Sidi-Bel Abbes and Bejaia. The cleaning process is simple. In Excel, I merged 2 datasets to create the ‘region’ attribute, then remove the extra row store attribute name. In RStudio, from the merge dataset, I remove all extra whitespaces from attributes name, and trim all extra space from 'classes' attribute. Later, covert attributes to appropriate types. Finally, I remove all rows with N/A values.

The clean dataset has 243 entries and 15 attributes. I focused on meteorological observations during the summer of 2012, from June to September, when wildfire incidents are most frequent. The year 2012 had the highest recorded fire occurrences between 2007 and 2018.

The dataset includes date variables and key weather factors: temperature, relative humidity (RH), and wind speed (Ws) and Rain that significantly influence wildfires. Additionally, there are also components of the FWI system included in the dataset which are Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), Buildup Index (BUI) and Fire Weather Index (FWI).

The output attribute (‘classes’) is categorized as 'fire' and 'not fire'. Among the 244 instances, 138 were classified as 'fire' and 106 as 'not fire'. The details of the dataset are summarized in Table 1.

Table 1. Attribute information

Day	1 to 31
Month	June to September
Year	2012
Temperature	in Celsius degrees, ranged from 22 to 42
RH	Relative Humidity in %, ranged from 21 to 90
Ws	Wind speed in km/h, ranged from 6 to 29
Rain	total rain in mm, ranged from 0 to 16.8
FFMC	Fine Fuel Moisture Code index from the FWI system, ranged from 28.6 to 92.5

[1] Algerian Forest Fires Dataset. (2019). UCI Machine Learning Repository. https://doi.org/10.24432/C5KW4N.
[2] Wildfires. National Geographic. Wildfires (nationalgeographic.org)

DMC	Duff Moisture Code index from the FWI system, ranged from 1.1 to 65.9
DC	Drought Code index from the FWI system, ranged from 7 to 220.4
ISI	Initial Spread Index index from the FWI system, ranged from 0 to 18.5
BUI	Buildup Index index from the FWI system, ranged from 1.1 to 68
FWI	Fire Weather Index, ranged from 0 to 31.1
Classes	two classes, namely “fire” and “not fire”
Region	categorized as ‘Sidi-Bel Abbes’ and ‘Bejaia’, created after merging two dataset to identify 2 regions

1. Exploratory data analysis

1.1 Distribution of attributes

1.1.1 Weather data observations

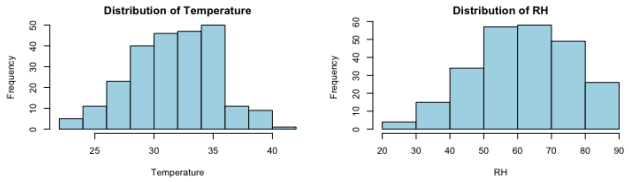


Fig. 1. Histogram of Temperature, RH

During the study period, Temperature is centered around 25-35 degrees, with occasional higher temperatures. Temperature is generally warm.

Most RH observations are concentrated between 50-80%, indicating moderately high to high humidity levels. There is a long left tail extending down to 20%, showing some low RH instances occur. Lower RH is associated with increased fire risk.

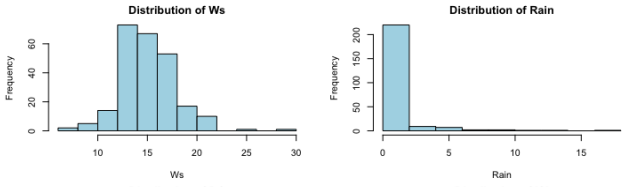


Fig. 2. Histogram of Ws, Rain

Wind speed (Ws) is right-skewed, with most values centered around 10-20km/h and a peak around 15 km/h. Higher wind speeds can accelerate fire spread. Rain amounts are also right-skewed, with most values near 0 mm. Lack of rain contributes to fire risk. The plot is not distributed normally, indicates rainfall events were infrequent.

1.1.2 FWI Components

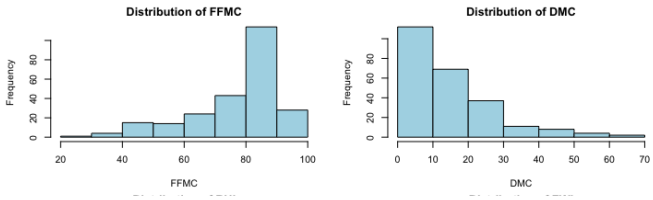


Fig. 3. Histogram of FFMC, DMC

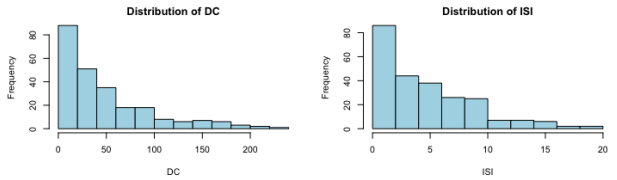


Fig. 4. Histogram of DC, ISI

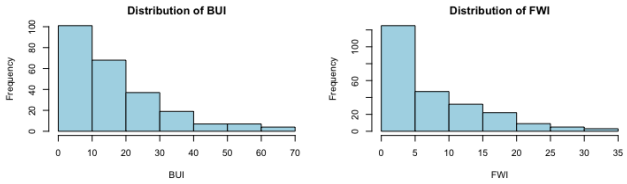


Fig. 5. Histogram of BUI, FWI

Except for Fine Fuel Moisture Code (FFMC), other FWI components all have right-skewed distributions, with higher values indicating higher fire risk.

Most Duff Moisture Code (DMC) values are below 20, with a long right tail up to 70. Higher DMC values indicate dry soil and organic layers, elevating fire risk.

Most Drought Code (DC) values are below 50, with a long value clustered on lower end. Elevated DC values indicate prolonged dry conditions, increasing the potential for intense and long-lasting fires.

Most Initial Spread Index (ISI) values are below 10. Buildup Index (BUI) is distributed somewhat uniformly between 0-30, pointing to variation in buildup index. Higher values suggest increased potential for severe fires.

Fire Weather Index (FWI) shows that most days have low fire danger (<5) and fewer with high danger.

FFMC has a left-skewed distribution, with a peak frequency around 80 and a long left tail. Low FFMC values suggest dry, flammable conditions, increasing the likelihood of fire ignition and rapid spread.

1.1.3 Fire and region

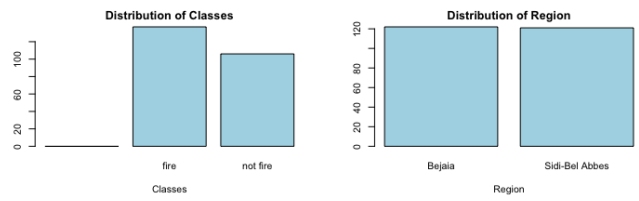


Fig. 6. Histogram of Fire, Region

Classes are relatively evenly split between fire and no fire conditions. Regions too, are evenly split between Bejaia and Sidi-Bel Abbes.

1.2 Explore month vs other attributes

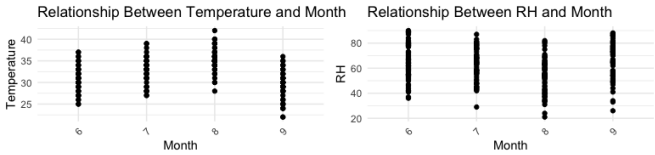


Fig. 7. Histogram of Month vs Temperature and Month vs RH

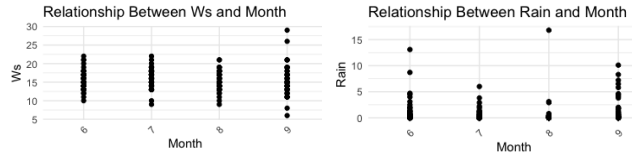


Fig. 8. Histogram of Month vs Ws and Month vs Rain

Fig. 9. Histogram of Month vs FFMC and Month vs DMC

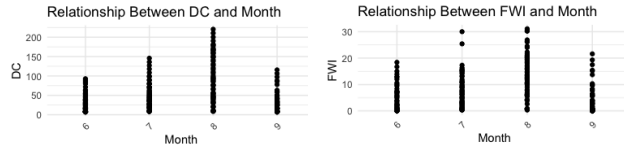


Fig. 10. Histogram of Month vs DC and Month vs FWI

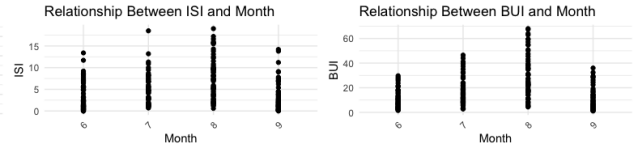


Fig. 11. Histogram of Month vs ISI and Month vs BUI

These plots show a clear seasonal trend, as Temperature, FFMC, DMC, DC, FWI, ISI and BUI all peak in July and August, while Rain has the lowest amount in these 2 months. These seasonal patterns significantly impact the risk of wildfires in Algeria.

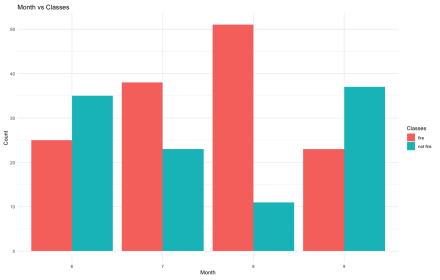


Fig. 12. Histogram of Month vs Classes

Combine to the findings from the previous plots of weather and FWI indicators, it makes sense why July and especially August has significant high level of wildfire.

1.3 Explore FWI vs other attributes

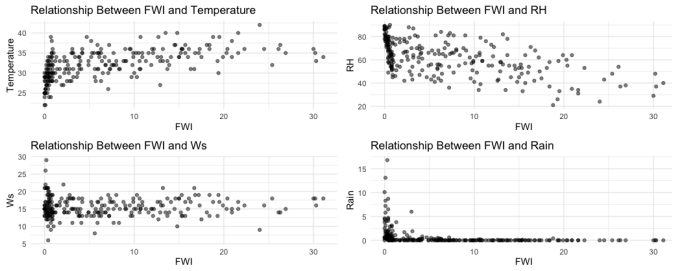


Fig. 13. Histogram of FWI vs weather factors (Temperature, RH, Ws, Rain)

FWI increases with Temperature, showing a clear positive correlation, while it decreases as RH increases. Therefore, higher temperatures and lower RH drive higher fire risk. In Ws plot, there is no clear correlation with FWI. Wind speed does not appear strongly tied to overall fire risk. For rain plot, more rain is associated with lower fire risk.

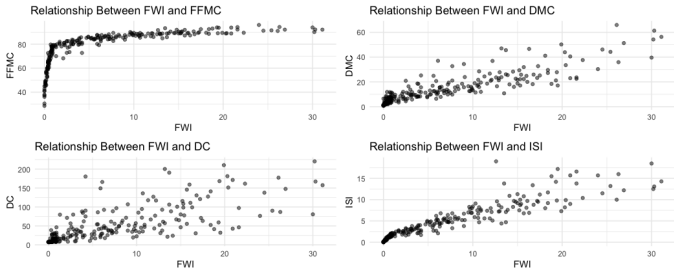


Fig. 14-1. Histogram of FWI vs FWI factor (FFMC, DMC, DC, ISI)

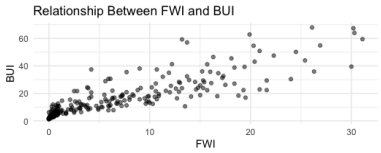


Fig. 14-2. Histogram of FWI vs FWI factor (BUI)

FFMC, DMC, DC, ISI and BUI plots all show a positive correlation with FWI.

2. Data preprocessing

2.1 Normalization

2.1.1 Min-max normalization

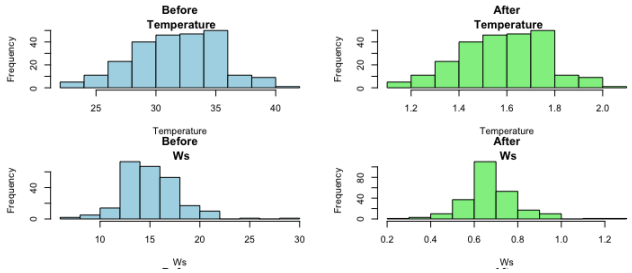


Fig. 15. Min-Max Normalization of Temperature and Ws

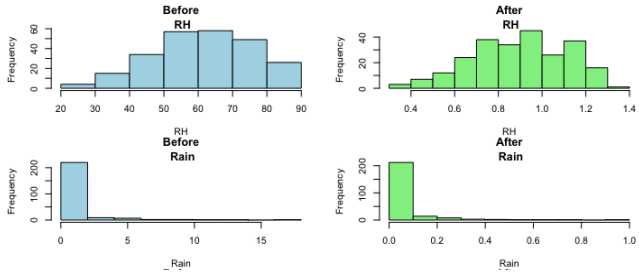


Fig. 16. Min-Max Normalization of RH and Rain

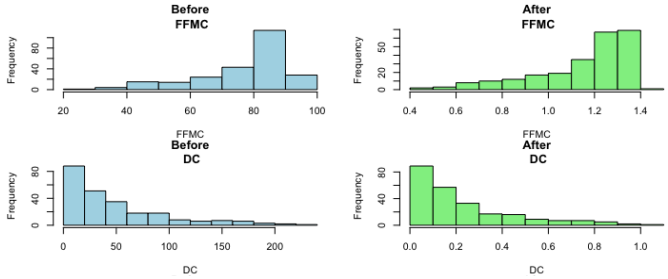


Fig. 17. Min-Max Normalization of FFMC and DC

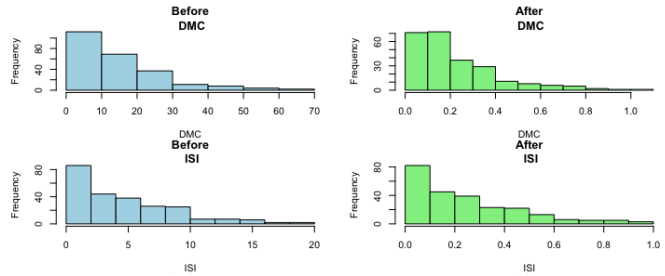
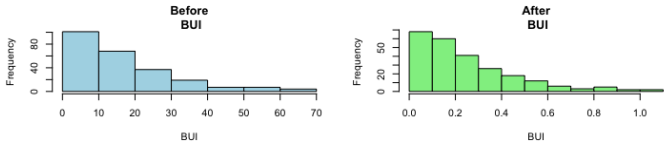


Fig. 18. Min-Max Normalization of DMC and ISI



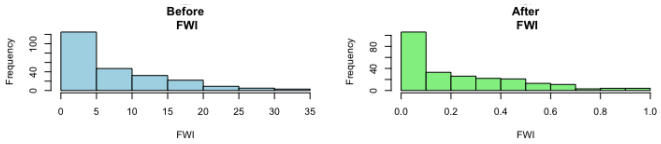


Fig. 19. Min-Max Normalization of BUI and FWI

These plots show the distribution of each variable before and after min-max normalization. The shape of the distributions remained generally the same after normalization, and attributes are rescaled to 0-1 range with the exception of Temperature (spanning from 0.2-2.0)

2.1.2 Z-scores normalization

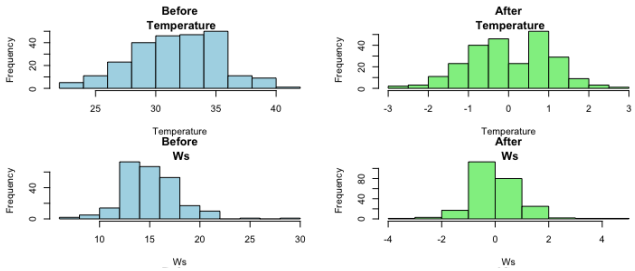


Fig. 20. Z-scores Normalization of Temperature and Ws

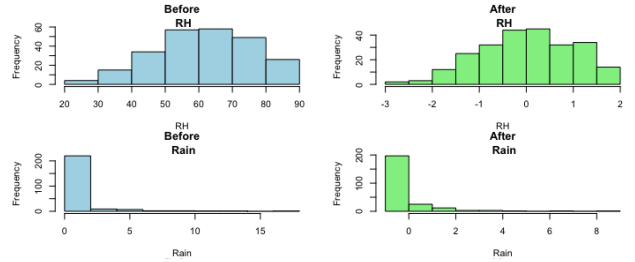


Fig. 21. Z-scores Normalization of RH and Rain

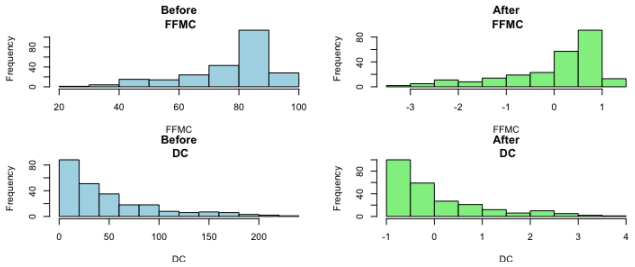


Fig. 22. Z-scores Normalization of FFMC and DC

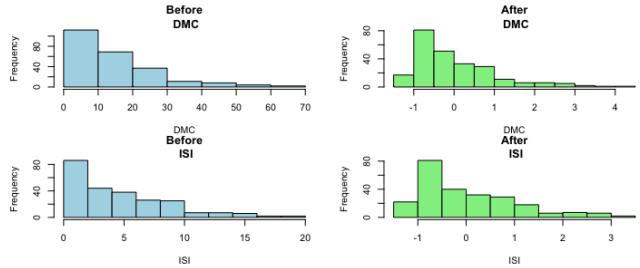


Fig. 23. Z-scores Normalization of DMC and ISI

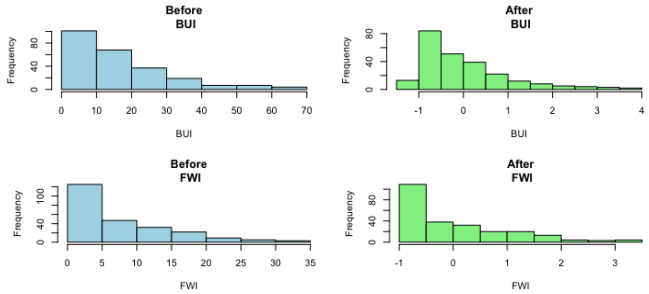


Fig. 24. Z-scores Normalization of FWI and BUI

For z-score, the shape of the distributions also remained generally the same after normalization, and attributes are rescaled to -3-+3 range.

2.1.3 Decimal scaling normalization

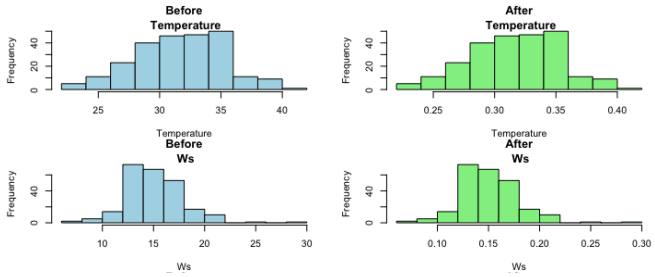


Fig. 25. Z-scores Normalization of Temperature and Ws

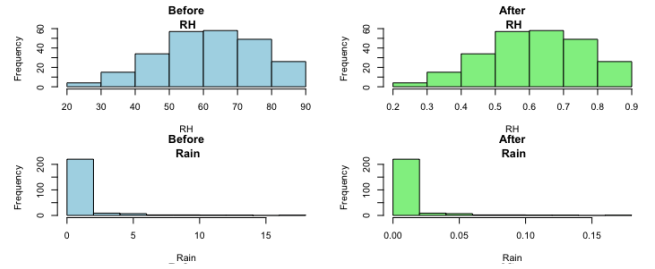


Fig. 26. Decimal scaling Normalization of Temperature and Ws

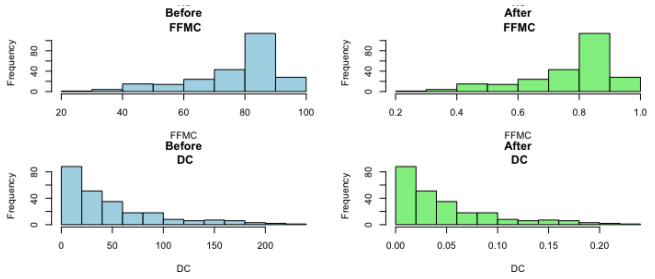


Fig. 27. Decimal scaling Normalization of FFMC and DC

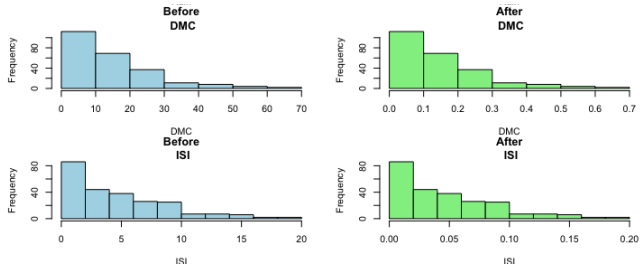


Fig. 28. Decimal scaling Normalization of DMC and ISI

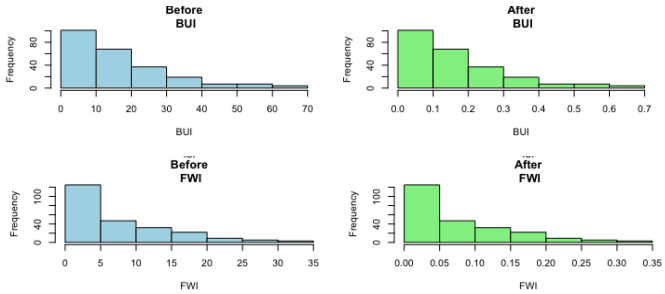


Fig. 29. Decimal scaling Normalization of FWI and BUI

Decimal scaling seems like the best normalization method for this dataset, as all attributes are rescaled to 0-1 range, and the shape of the distributions also remained the same after normalization.

2.2 Binning

2.2.1 Equal Width Binning for temperature

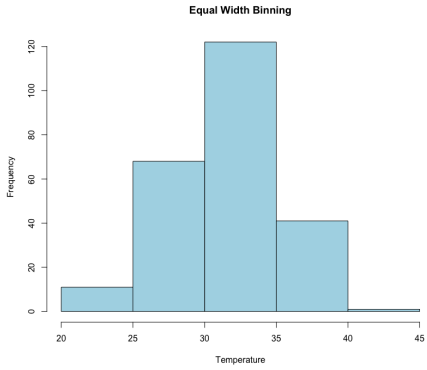


Fig. 30. Equal Width Binning for temperature

The attribute I choose to bin is Temperature, a continuous variable. With equal width binning, temperature has been divided into bins of equal width (5°C intervals). The first bin (20-25°C) contains around 20-30 samples, 25-30°C bin contains the most samples contains around 60, 30-35°C peaks at 120 samples, 35-40°C bin contains around 40 samples, and very few samples in the highest 40-45°C bin. Equal Width Binning still keeps the original shape of the plot, but it does not add additional insight for us on Temperature attribute.

2.2.2 Equal Frequency Binning for temperature

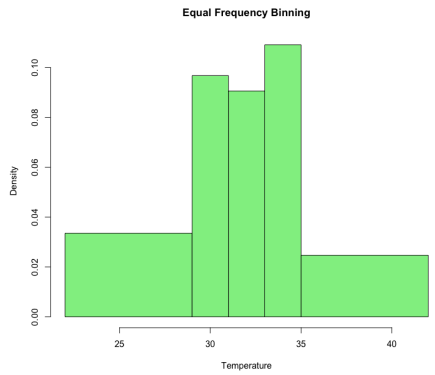


Fig. 31. Equal Frequency Binning for temperature

Equal Frequency Binning divided temperature into 5 bins, however, the bins do not all have equal densities, indicates it failed to conduct Equal Frequency Binning here.

2.3 Transformation for rain

In my EDA step, I observed that the original rain data was highly right skewed, indicating a non-normal distribution. Therefore, the natural log, square root, and inverse square root transformations will be conducted to achieve normality.

2.3.1 Natural log transformation for rain

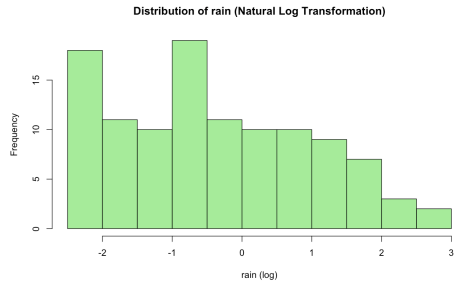


Fig. 32. Natural log transformation for rain

Comparing the original rain vs the transformation plot, natural log method is effective in making the distribution more normal. The distribution now appears much more symmetrical and bell-shaped.

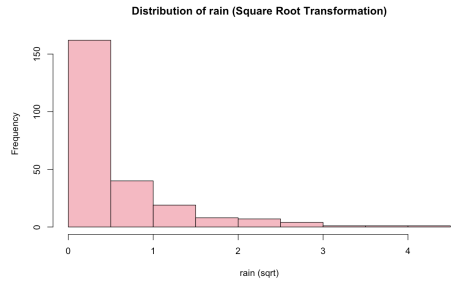


Fig. 33. Square root transformation for rain

The square root transformation do not make the distribution of rain normal, as it still keeps the original shapes of the rain plot.

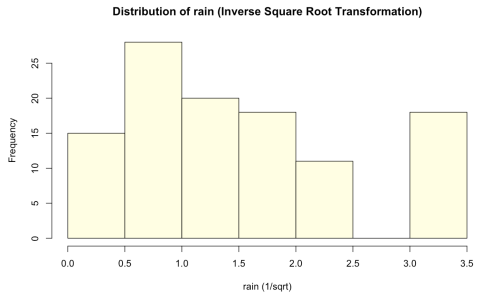


Fig. 34. Inverse square root transformation for rain

Inverse square root transformation also works in helping normalize rain attribute and make the distribution more symmetrical.

In general, as seen from the plots, natural log transformation and inverse square root all work to transform the highly skewed rain attribute into a more normalized one.

3. Regression Analysis

Based on EDA step, I perform regression analysis for this question: Can we predict the Fire Weather Index (FWI) based on weather-related variables such as Temperature, Relative Humidity (RH), Wind Speed (Ws), Rainfall (Rain)?

I employed three regression models, Linear Regression, Random Forest Regression, and Gradient Boosting Regression and compared to assess their effectiveness in predicting FWI. The models are trained on an 80% subset of the dataset, and their predictions are evaluated on the remaining 20% of the data.

3.1 Linear regression

Null Hypothesis: There is no significant relationship between the weather-related variables and FWI

Alternative Hypothesis: There is a significant relationship between the weather-related variables and FWI

Table 2. Results of the linear regression

Predictor	*b*	*b* 95% CI [LL, UL]	*beta*	*beta* 95% CI [LL, UL]	sr^2	sr^2 95% CI [LL, UL]	*r*	Fit
(Intercept)	-11.53	[-24.70, 1.65]						
Temperature	0.67**	[0.38, 0.96]	0.33	[0.19, 0.47]	.06	[.01, .11]	.57**	
RH	-0.20**	[-0.26, -0.13]	-0.39	[-0.53, -0.26]	.09	[.03, .15]	-.58**	

Ws	0.61 **	[0.34, 0.88]	0.24	[0.13, 0.35]	.05	[.01, .10]	.06	
Rain	-0.61 **	[-1.01, -0.20]	-0.16	[-0.27 , -0.05]	.02	[-.01 , .06]	-.30 **	
								R^2 = .478**
								95% CI[.37 ,.55]

Note. A significant *b*-weight indicates the beta-weight and semi-partial correlation are also significant. *b* represents unstandardized regression weights. *beta* indicates the standardized regression weights. sr^2 represents the semi-partial correlation squared. *r* represents the zero-order correlation. *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.* indicates $p < .05$. ** indicates $p < .01$.

According to linear regression on the train data, there is a significant relationship between the weather-related variables and FWI ($F(4,189)=43.3$, $p < .001$, $R^2 = .467$). Temperature exhibited a positive relationship with FWI ($t=4.625$, $p<.001$), indicating that as Temperature increases, FWI tends to increase. Conversely, RH showed a negative relationship with FWI ($t=-5.701$, $p<.001$), suggesting that higher RH values are associated with lower FWI. Wind Speed (Ws) also demonstrated a positive relationship with FWI ($t=4.423$, $p<.001$), while Rainfall (Rain) exhibited a negative relationship ($t=-2.962$, $p=.003$).

3.2 Gradient Boosting Regression

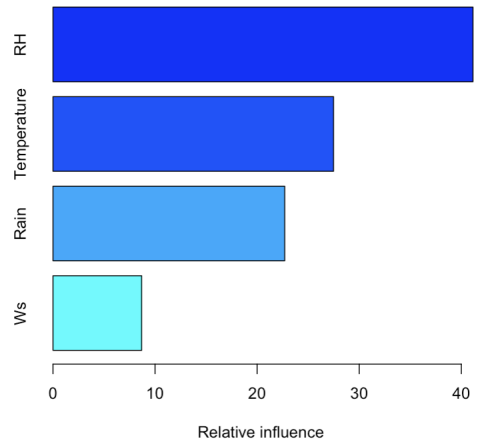


Fig. 35. Gradient Boosting Regression

The Gradient Boosting Regression model was specified with 100 trees. Running the model, I observe that: RH demonstrated the highest relative importance (44.06%), indicating its substantial influence on the model's predictive power. Temperature has the second highest relative importance (25.79%), and Rain ranks the third (24.05%). Ws only correspond with 6.09% importance.

Therefore, combine with the previous linear regression, RH has a significant role in predicting fire weather conditions and a higher RH values were associated with lower FWI.

3.3 RMSE

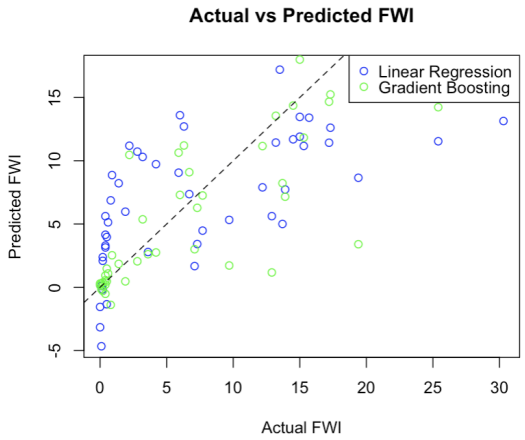


Fig. 36. Visualization of actual vs predicted of Linear Regression and Gradient Boosting Regression

Calculate RMSE, Gradient Boosting models have lower RMSE compared to Linear Regression (5.84 vs 4.67), suggests that Gradient Boosting provides more accurate predictions of the Fire Weather Index based on the given set of weather-related variables.

Conclusion

This study, we delved into the intricate patterns of the Algerian Forest Fires dataset. Our EDA revealed seasonal trends in fire risk and relationship between attributes and fire risk. Regression analyses indicate the influence of humidity in mitigating fire risk.

In conclusion, this study significantly contributes to our understanding of Algerian forest fires. There are still some limit, such as the sample size, with only 244 instances, it's challenging to reduce error. A larger sample size can give a more reliable result for my research.

While this study provides a comprehensive analysis, further research can explore additional variables and more complex machine learning models to enhance predictive accuracy. By continually refining the analyses, I believe we can strengthen our ability to predict and mitigate the risk of forest fires, and save our environment and communities.

Reference

[1] Algerian Forest Fires Dataset. (2019). UCI Machine Learning Repository. https://doi.org/10.24432/C5KW4N.
[2] Wildfires. National Geographic. Wildfires (nationalgeographic.org)