Understanding Causes of Variation in area of wildfires in the northeast region of Portugal

# Abstract:

The area that a wildfire burns differs for every fire. Scientists, government officials, and fire rescuers need to make decisions about fire safety. They need to know what areas are at higher risk of lighting on fire and what conditions are more likely to cause fires. This project begins the process of helping to identify and predict what factors contribute to wildfires. This is not a full analysis. In order to accurately predict when, where, and under what conditions wildfires begin, a more intense process of modeling would need to be undertaken. In the future, time series would need to be accounted for and another type of modeling (ex. Multiple Linear Regression) would need to be performed to account for the interdependence of variables.

# Authors:

Joseph Arnesen, [josepharnesen@gmail.com](mailto:josepharnesen@gmail.com)

**LAST UPDATED:** Thursday, December 17, 2020

Contents

[Abstract: 1](#_Toc59129206)

[Authors: 1](#_Toc59129207)

[Initial Data Collection Report 3](#_Toc59129208)

[Data Source 3](#_Toc59129209)

[Data Description Report 4](#_Toc59129210)

[Dictionary 4](#_Toc59129211)

[Univariate Properties 5](#_Toc59129212)

[Univariate Visualizations 6](#_Toc59129213)

[Data Exploration Report 10](#_Toc59129214)

[Numeric to numeric analysis 10](#_Toc59129215)

[FFMC 12](#_Toc59129216)

[DMC 13](#_Toc59129217)

[DC 14](#_Toc59129218)

[ISI 15](#_Toc59129219)

[Temperature 16](#_Toc59129220)

[RH (Relative Humidity) 17](#_Toc59129221)

[Wind 18](#_Toc59129222)

[Categorical to Numeric Analysis 19](#_Toc59129223)

[X 20](#_Toc59129224)

[Y 21](#_Toc59129225)

[Month 22](#_Toc59129226)

[Day 23](#_Toc59129227)

[Data Quality Report 24](#_Toc59129228)

[Missing Data 24](#_Toc59129229)

[Skewness 24](#_Toc59129230)

[Outliers 24](#_Toc59129231)

[Conclusion 25](#_Toc59129232)

# Initial Data Collection Report

## Data Source

* Forest Fires
  + <https://archive.ics.uci.edu/ml/datasets.php?format=&task=reg&att=num&area=phys&numAtt=&numIns=100to1000&type=&sort=nameUp&view=table>

# Data Description Report

## Dictionary

* X
  + X-axis spatial coordinate within the Montesinho park map: 1 to 9
* Y
  + Y-axis spatial coordinate within the Montesinho park map: 2 to 9
* Month
  + Month of the year: ‘jan’ to ‘dec’
* Day
  + Day of the week: ‘mon’ to ‘sun’
* FFMC
  + FFMC index from the FWI system: 18.7 to 93.20
  + FFMC - This index classifies the moisture content of litter and other cured fine fuels, like needles or twigs less than 1 cm in diameter. FFMC is representative of the top litter layer 1-2 cm deep and has a short-term memory, only reflecting weather conditions that have occurred over the past three days. \*
* DMC
  + DMC index from the FWI system: 1.1 to 291.3
  + DMC - This index indicates the moisture content of loosely-compacted organic layers with a depth of 5-10 cm. DMC fuels have a slower drying rate than FFMC fuels and DMC may be used in predicting the probability of fire ignition by lightning. \*
* DC
  + DC index from the FWI system: 0.0 to 56.10
  + DC - This third moisture index reflects the moisture content of compact organic layers, 10-20 cm deep. DC is indicative of long-term moisture conditions and deep burning fires, being related to mop-up and patrol difficulties. \*
* ISI
  + ISI index from the FWI system: 0.0 to 56.10
  + ISI - This index combines FFMC and wind speed, being a good indicator for fire spread. Moreover, if the topography is known, the slope effect on fire spread may be converted into a wind speed equivalent and added to the real wind speed. \*
* Temp
  + Temperature in Celsius degrees: 2.2 to 33.30
* RH
  + Relative humidity in %: 15.0 to 100
* Wind
  + Wind speed in km/h: 0.4 to 9.40
* Rain
  + Outside rain in mm/m2: 0.0 to 6.4
* Area
  + The burned area of the forest (in ha): 0.00 to 1090.84

\* These descriptions were taken from this website: <http://www.eumetrain.org/data/3/30/navmenu.php?tab=3&page=6.0.0#:~:text=DMC%20%2D%20This%20index%20indicates%20the,of%20fire%20ignition%20by%20lightning>.

## Univariate Properties

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **Count** | **1st Quartile** | **2nd Quartile** | **3rd Quartile** | **Mean** | **Median** | **Mode** | **Min** | **Max** | **StDev** | **Kurtosis** | **Skewness** |
| X | 517 |  |  |  | Inf | 4 | 4 | 1 | 9 |  | -1.17 | 0.036 |
| Y | 517 |  |  |  | Inf | 4 | 4 | 2 | 9 |  | 1.42 | 0.417 |
| Month | 517 |  |  |  |  |  | Aug | Apr | Sep |  |  |  |
| Day | 517 |  |  |  |  |  | Sun | Fri | Wed |  |  |  |
| FFMC | 517 | 90.2 | 91.6 | 92.9 | 90.6 | 91.6 | 91.6 | 18.7 | 96.2 | 5.5 | 67.0 | -6.57 |
| DMC | 517 | 68.6 | 108.3 | 142.4 | 110.8 | 108.3 | 99 | 1.1 | 291.3 | 64.0 | 0.204 | 0.547 |
| DC | 517 | 437.7 | 664.2 | 713.9 | 547.9 | 664.2 | 745.3 | 7.9 | 860.6 | 248.0 | -0.245 | -1.100 |
| ISI | 517 | 6.5 | 8.4 | 10.8 | 9.02 | 8.4 | 9.6 | 0 | 56.1 | 4.5 | 21.45 | 2.536 |
| Temp | 517 | 15.5 | 19.3 | 22.8 | 18.9 | 19.3 | 17.4 | 2.2 | 33.3 | 5.8 | 0.136 | -0.331 |
| RH | 517 | 33.0 | 42 | 53 | 44.3 | 42 | 27 | 15 | 100 | 16.3 | 0.438 | 0.862 |
| Wind | 517 | 2.7 | 4 | 4.9 | 4.01 | 4 | 2.2 | 0.4 | 9.4 | 1.79 | 0.054 | 0.571 |
| Area | 517 | 0.0 | 0.52 | 6.57 | 12.8 | 0.52 | 0 | 0 | 1090.8 | 63.6 | 194.1 | 12.84 |

# Univariate Visualizations

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

# Data Exploration Report

This part of the report is split into two parts—the numeric to numeric analysis and the categorical to numeric analysis. In the process of performing the analysis we discovered that the “rain” variable was highly skewed. After looking at the raw data, we discovered that all but 8 of the values for “rain” were zero. Because it rained so little we determined that it was not worth it to include rain in our analysis. So for the rest of the analysis, rain is not included.

## Numeric to numeric analysis

In our numeric to numeric analysis we ran functions to determine the statistics and print visualizations of the data before and after the data was cleaned. After cleaning the data, we discovered that our label “area” had a high volume of zeros. We decided that since we wanted to determine where fires actually start we would drop all the rows where area was zero. We then ran the functions again. The following data shows the r values and the p-values for all of the numeric features.

Unfortunately, the data did not show that any of these features except for perhaps “temp” (from the analysis before cleaning) had a significant affect on our label—“area.” As was mentioned at the beginning of the report a more intense analysis should be performed (probably MLR) in order to better determine which variables actually affect the area of forest fires.

Correlation analysis before cleaning

r p-value

FFMC 0.040 0.363

DMC 0.073 0.097

DC 0.049 0.262

ISI 0.008 0.851

temp 0.098 0.026

RH -0.076 0.086

wind 0.012 0.780

area 1.000 0.000

Correlation analysis after cleaning

r p-value

FFMC 0.046 0.292

DMC 0.068 0.122

DC 0.079 0.074

ISI 0.041 0.355

temp 0.074 0.095

RH -0.044 0.322

wind 0.063 0.155

area 1.000 0.000

Correlation analysis after dropping zeros in the label “area”

r p-value

FFMC -0.029 0.633

DMC 0.045 0.463

DC 0.002 0.976

ISI -0.085 0.162

temp -0.003 0.964

RH -0.059 0.337

wind 0.049 0.427

area 1.000 0.000

### FFMC

For these first three features, I thought that they would be significant because they measure the moisture in the ground, but according the analysis they were not significant

Chart, scatter chart

Description automatically generated

Before Cleaning

Chart, scatter chart

Description automatically generated

After Cleaning

Chart, scatter chart

Description automatically generated

After dropping zeros in area

### DMC

Chart, scatter chart

Description automatically generated

Before Cleaning

Chart, scatter chart

Description automatically generated

After Cleaning

Chart, scatter chart

Description automatically generated

After dropping zeros in area

### DC

Chart, scatter chart

Description automatically generated

Before Cleaning

Chart, scatter chart

Description automatically generated

After Cleaning

Chart, scatter chart

Description automatically generated

After dropping zeros in area

### ISI

I also thought that this feature would affect the area, but again according to the analysis, that is not true.

Chart

Description automatically generated

Before Cleaning

Chart, scatter chart

Description automatically generated

After Cleaning

Chart, scatter chart

Description automatically generated

After dropping zeros in area

### Temperature

Temperature was the only feature that showed some significance (p-value before cleaning < 0.05 and p-value after cleaning is close to 0.5). This makes sense because the hotter the weather, the more likely it would be that a fire would start.

Chart, scatter chart

Description automatically generated

Before Cleaning

Chart, scatter chart

Description automatically generated

After Cleaning

Chart, scatter chart

Description automatically generated

After dropping zeros in area

### RH (Relative Humidity)

I thought humidity would have a greater affect, but according to the data, it is not quite significant.

Chart, scatter chart

Description automatically generated

Before Cleaning

Chart, scatter chart

Description automatically generated

After Cleaning

Chart, scatter chart

Description automatically generated

After dropping zeros in area

### Wind

Interestingly, wind had no affect on the area of the fire. I thought that wind would be a major factor, but according to the data, it doesn’t show that wind had a significant affect.

Chart, scatter chart

Description automatically generated

Before Cleaning

Chart, scatter chart

Description automatically generated

After Cleaning

Chart, scatter chart

Description automatically generated

After dropping zeros in area

## Categorical to Numeric Analysis

For the categorical to numeric analysis, not only did we include the “month” and “day” features, but we also analyzed the “X” and “Y” features using the analysis. Since all of these features included more than two groups we performed ANOVA tests on all of them and then determined which t-tests between groups were significant.

Again, we were surprised that there were not very many significant differences between groups. However, we should note again, that since this data is heavily influenced by time that further analysis should be performed using time series analyzes.

The following show the bar charts for the four categorical variables that we have. Again, each feature has three graphs—one for before data cleaning, another for after the data cleaning process, and a final for after taking the zeros out of the label “area.”

### X

The groups that were significant in these two features (“X” and “Y”) are not as valuable unless you know the exact areas that they represent. However, if you did know which areas they represent you could do further research on what the difference between areas is.

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, bar chart

Description automatically generated

### Y

Chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart, bar chart

Description automatically generated

### Month

Interestingly, there is a significant difference between March and December and between January and December. This is very curious, and would require more research to understand why this occurred.

Chart

Description automatically generated

Chart

Description automatically generated

Chart, bar chart

Description automatically generated

### Day

There are no significant differences between days of the week.

Chart, box and whisker chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

# Data Quality Report

## Missing Data

There was no missing data in this data set, so we do not need to delete any case/features, replace missing values, or impute missing values.

## Skewness

There were several variables that were highly skewed. Both “FFMC” and “DC” had skewness scores < -1. In addition “ISI” and “area” both had positive skewness scores > 1. We used typical techniques to correct for skewness (square root, cubed root, log, for positive skewness and exponentials for negative skewness). However, we found that binning all features that were skewed corrected all of the skewness to scores between -1 and 1. So we decided to use the binned version of those features in our second analysis.

## Outliers

The following table shows the outliers after we had corrected for skewness:

min count below max count above

FFMC 0.215053 1.0 82.879725 0.0

DMC -81.267106 0.0 303.011787 0.0

DC 0.290540 1.0 83.322612 1.0

ISI -0.191410 0.0 83.487348 1.0

temp 1.469292 0.0 36.309044 0.0

RH -4.664207 0.0 93.240609 5.0

wind -1.357356 0.0 9.392559 4.0

area -53.566203 0.0 108.494636 0.0

We decided to use the Turkey method to find our outliers since the original data was very skewed. Since there were so few outliers, we decided to not replace any of the outliers.

# Conclusion

The most significant take away I had from this analysis is that there is much more that needs to be done to truly analyze where and what conditions affect higher areas of the forest burning. As previously mentioned, time series definitely needs to play an important role in future analysis. Likewise, a more thorough analysis using some sort of multi variable model is also needed.