Programming Language
Python 3.7.3

Dataset Information [1]
Dataset link: https://archive.ics.uci.edu/ml/datasets/Forest+Fires
Title: Forest Fires
Created by: Paulo Cortez and Anibal Morais (Univ. Minho) @ 2007

Statistical Analysis
The purpose of this analysis is to understand how the weather, season and other parameters would influence the likelihood of a forest fire.

Data are collected from Montesinho natural park in Portugal from January 2000 to December 2003. Each time a forest fire occurs, the corresponding features of the fire will be recorded[1]. There are total 13 attributes and 517 observations with no missing value, example dataset is presented in Figure 1. The meaning of each attribute is provided in the Appendix. To simplify the plotting, the data string in the month and day column is changed to numeric values. The example data and corresponding data type is summarized in Figure 2.

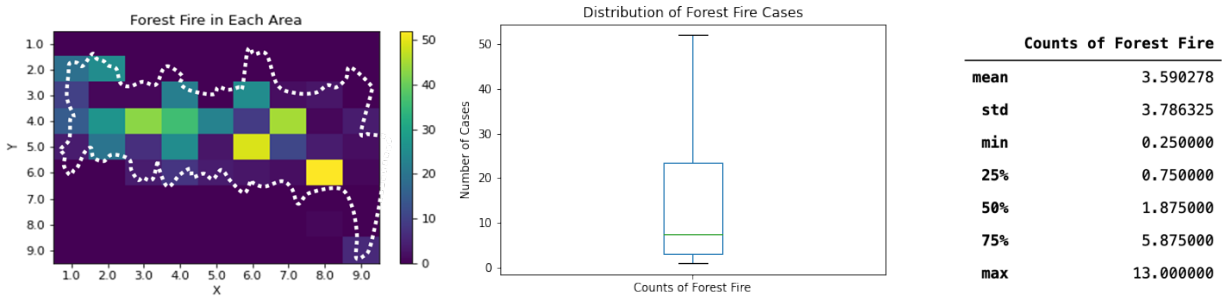
	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
0	7	5	mar	fri	86.2	26.2	94.3	5.1	8.2	51	6.7	0.0	0.0
1	7	4	oct	tue	90.6	35.4	669.1	6.7	18.0	33	0.9	0.0	0.0
2	7	4	oct	sat	90.6	43.7	686.9	6.7	14.6	33	1.3	0.0	0.0
3	8	6	mar	fri	91.7	33.3	77.5	9.0	8.3	97	4.0	0.2	0.0
4	8	6	mar	sun	89.3	51.3	102.2	9.6	11.4	99	1.8	0.0	0.0

Figure 1 The first 5 rows of example data.

	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
0	7	5	3	5	86.2	26.2	94.3	5.1	8.2	51	6.7	0.0	0.0
1	7	4	10	2	90.6	35.4	669.1	6.7	18.0	33	0.9	0.0	0.0
2	7	4	10	6	90.6	43.7	686.9	6.7	14.6	33	1.3	0.0	0.0
3	8	6	3	5	91.7	33.3	77.5	9.0	8.3	97	4.0	0.2	0.0
4	8	6	3	7	89.3	51.3	102.2	9.6	11.4	99	1.8	0.0	0.0

```
#   Column  Non-Null Count  Dtype
---  ------  --------------  -----
0   X       517 non-null    int64
1   Y       517 non-null    int64
2   month   517 non-null    int64
3   day     517 non-null    int64
4   FFMC    517 non-null    float64
5   DMC     517 non-null    float64
6   DC      517 non-null    float64
7   ISI     517 non-null    float64
8   temp    517 non-null    float64
9   RH      517 non-null    int64
10  wind    517 non-null    float64
11  rain    517 non-null    float64
12  area    517 non-null    float64
```

Figure 2 Processes data and corresponding data type.



	Counts of Forest Fire
mean	3.590278
std	3.786325
min	0.250000
25%	0.750000
50%	1.875000
75%	5.875000
max	13.000000

Figure 3 Left: Number of forest fire cases at the location specified by X and Y within 4 years with white dots representing the border of Montesinho natural park. Note the boundary is not accurate and should be used for guiding the eye only. Middle: Summary of the total number of cases that happened in different areas within 4 years. Right: Summary of cases which happens annually at different locations.

In order to understand which area is suffered from forest fire the most. The statistics of the number of cases in different areas is summarized in Figure 3 right. The data shows that there are approximately 3.59 cases in each location annually. However, the distribution of the cases is highly inhomogeneous, most areas have 1 to 6 cases annually, while some area has more than 10 cases annually. As shown in Figure 3 left, most cases locate at y from 4 to 6 and x from 6 to 8, especially the area where x=8, y=6.

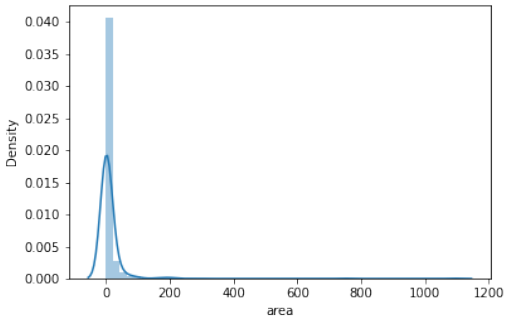


Figure 4 The area burnt in each fire.

According to Figure 4, generally, a small area is burnt during a forest fire.

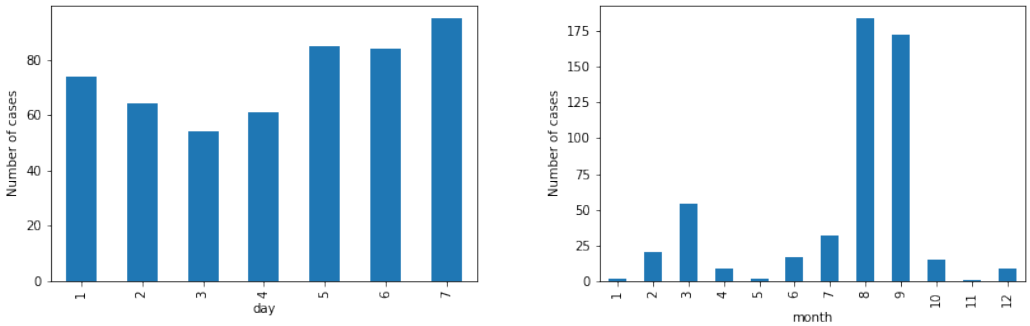


Figure 5 Left: Number of cases on the day of the week (1 means Monday) within the 4 years under research. Right: Number of cases of each month within the 4 years under research.

Human interaction or weather of a year is highly likely to play a role in the forest fire. The data type is categorical, thus a bar plot is chosen to illustrate the data. As shown in Figure 5, from the cases on the day of the week and month of a year, there is no significant difference in the number of cases between weekdays and weekends, even though cases that happened on weekends are

slightly higher than on weekdays. Nevertheless, there is a strong correlation between the month of a year and the likelihood of forest fire. The majority of fire cases occurred on in August and September.

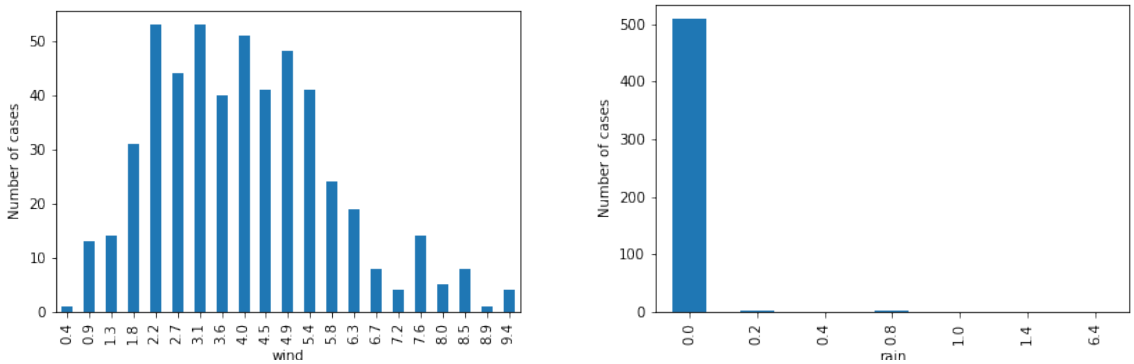


Figure 6 Left: Correlation of wind speed and number of cases within the 4 years under research. Right: Correlation of outside rain amount and number of cases of each month within the 4 years under research.

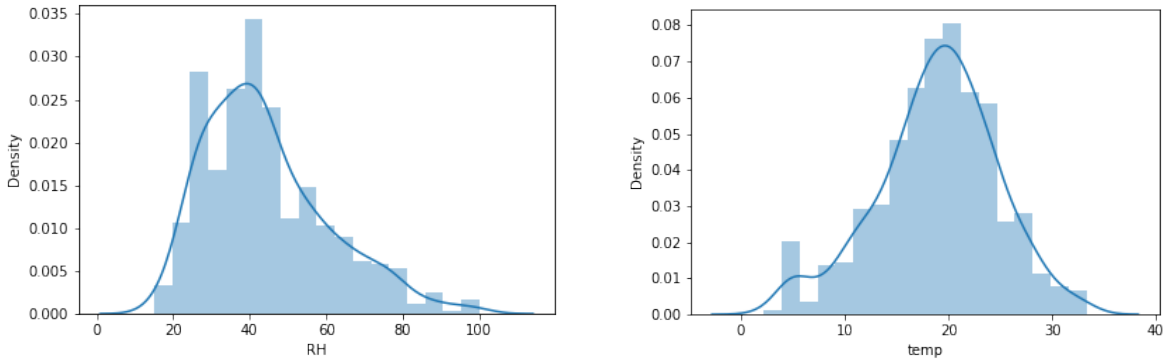


Figure 7 Left: Correlation of relative humidity and number of cases within the 4 years under research. Right: Correlation of temperature and number of cases of each month within the 4 years under research.

The fire is expected to be highly correlated with weather. The data is continuous for wind speed, rain, temperature and relative humidity. However, there are limited outputs in wind speed and amount of rain, thus I use bar plots (Figure 6) to see the distribution of the results. When the wind speed is within 1.8 to 5.8 km/h, the fire is likely to happen. In addition, forest fire has an extremely sensitive response to outside rain, such that a 0.2 mm/m^2 increase of outside rain will result in a significant change. Figure 7 shows the correlation of the number of cases with relative humidity and temperature. Out of the expectation, the fire is likely to occur on a relative humid (30<HR<60) day instead of on a super dry day. Besides, a fire has a higher chance to happen on a warm day but not on an extremely hot day.

The conflict between the intuition and analysis results may indicate the limitation of the current analysis method - only one feature is selected to study the impact. For example, high temperature will facilitate the evaporation of water from plants and ground, which further lead to higher relative humidity. One needs to consider both factors to see predict the likelihood of forest fire. Therefore, new parameters that include multiple features need to be used.

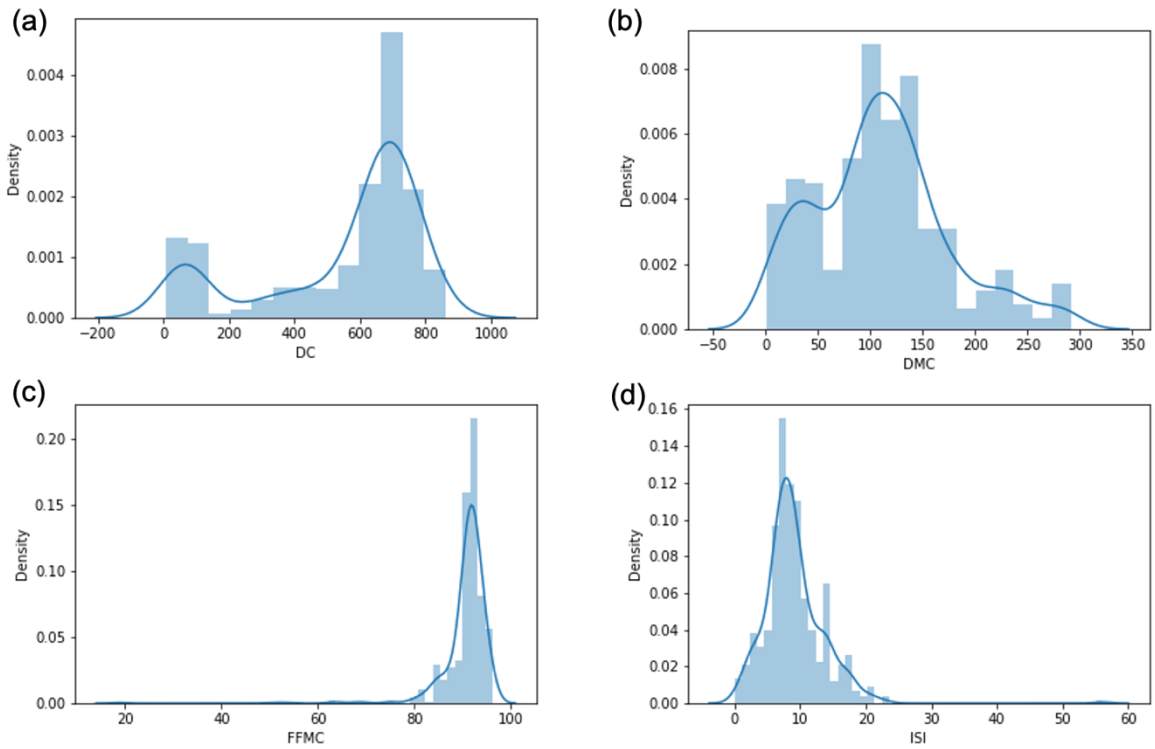


Figure 8 Correlation of number of cases with (a) DC, (b) DMC, (c)FFMC, (d)ISI.

	FFMC	DMC	DC	ISI
count	517.000000	517.000000	517.000000	517.000000
mean	90.644681	110.872340	547.940039	9.021663
std	5.520111	64.046482	248.066192	4.559477
min	18.700000	1.100000	7.900000	0.000000
25%	90.200000	68.600000	437.700000	6.500000
50%	91.600000	108.300000	664.200000	8.400000
75%	92.900000	142.400000	713.900000	10.800000
max	96.200000	291.300000	860.600000	56.100000

Figure 9 Statistical summarization of various index plotted in Figure 8.

Drought Code (DC), Duff Moisture Code (DMC), Fine Fuel Moisture Code (FFMC), and Initial Spread Index (ISI) index are indexes that incorporate multiple existing features, for more detail, please refer to Appendix and Ref.[1] and [2]. The current summarization will lead naïve conclusion that drought condition, high fuel moisture of forest litter fuels, the middle level of the fuel moisture of decomposed organic material will likely to induce forest fire. But, on the bright side, the spread potential is relatively low. To further improve the prediction, the model needs to incorporate those relative features in a more efficient way. The model also needs to be further tested with accuracy and stability.

Appendix

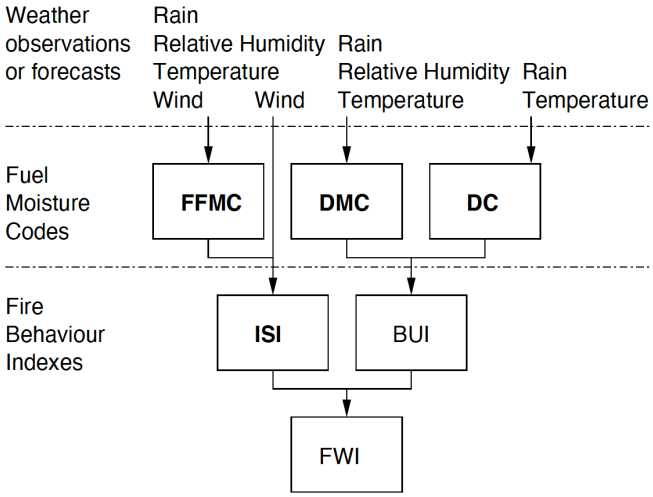


Figure 10 The Fire Weather Index structure taken from Ref [1]

The meaning of each attributes is summarized as follow [1]:
1. X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
2. Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
3. month - month of the year: "jan" to "dec"
4. day - day of the week: "mon" to "sun"
5. FFMC – Fine Fuel Moisture Code (FFMC) index from the forest Fire Weather Index (FWI) system: 18.7 to 96.20
6. DMC - Duff Moisture Code (DMC) index from the FWI system: 1.1 to 291.3
7. DC - Drought Code (DC) index from the FWI system: 7.9 to 860.6
8. ISI - Initial Spread Index (ISI) index from the FWI system: 0.0 to 56.10
9. temp – temperature (temp) in Celsius degrees: 2.2 to 33.30
10. RH - relative humidity (RH) in %: 15.0 to 100
11. wind - wind speed in km/h: 0.40 to 9.40
12. rain - outside rain in mm/m^2 : 0.0 to 6.4
13. area - the burned area of the forest (in ha): 0.00 to 1090.84 (this output variable is very skewed towards 0.0, thus it may make sense to model with the logarithm transform).

Meaning of the abbreviation taken from Ref. [2]:
The Fine Fuel Moisture Code (FFMC) represents fuel moisture of forest litter fuels under the shade of a forest canopy. It is intended to represent moisture conditions for shaded litter fuels, the equivalent of 16-hour timelag. **It ranges from 0-101**. Subtracting the FFMC value from 100 can provide an estimate for the equivalent (approximately 10h) fuel moisture content, most accurate when FFMC values are roughly above 80.
The Duff Moisture Code (DMC) represents fuel moisture of decomposed organic material underneath the litter. System designers suggest that it is represents moisture conditions for the equivalent of 15-day (or 360 hr) timelag fuels. It is unitless and open ended. It may provide insight to live fuel moisture stress.
The Drought Code (DC), much like the Keetch-Byrum Drought Index, represents drying deep into the soil. It approximates moisture conditions for the equivalent of 53-day (1272 hour) timelag

fuels. It is unitless, with a maximum value of 1000. Extreme drought conditions have produced DC values near 800.

The Initial Spread Index (ISI) is analogous to the NFDRS Spread Component (SC). It integrates fuel moisture for fine dead fuels and surface windspeed to estimate a spread potential. ISI is a key input for fire behavior predictions in the FBP system. It is unitless and open ended.

Reference:

[1] P. Cortez and A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data. In J. Neves, M. F. Santos and J. Machado Eds., New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimarães, Portugal, pp. 512-523, 2007. APPIA, ISBN-13 978-989-95618-0-9. Available at: http://www3.dsi.uminho.pt/pcortez/fires.pdf

[2] National wide fire coordinating group
https://www.nwcg.gov/publications/pms437/cffdrs/fire-weather-index-system .