Homework- Initial Analysis on Forest Fire Dataset

Alison Jing Huang

4/15/2018

### 1. Brief Dataset Glimpse

##### The dataset contains 13 different variables, with **X**, **Y**, **MONTH** and **DAY** being categorical, and the remaing 9 attributes being continuous. This multivariable dataset is suitable for setting up a predictive model and using Machine Learning methods to train datasets.

1. **X**: x-axis coordinate (from 1 to 9). **It indicates one of the 9 sub-areas**.
2. **Y**: y-axis coordinate (from 1 to 9). **It indicates one of the 9 sub-areas obtained from the division of the area of study along the Y axis. All the areas have the same size**.
3. **MONTH**: **Month of the year (from 1 to 12)**
4. **DAY**: **Day of the week (from 1 to 7)**
5. **FFMC**: Fine Moisture Code (from 18.7 to 96.20) - **moisture content of surface litter**
6. **DMC**: Duff Moisture Code (from 1.1 to 291.3) - **rating for average moisture content of loosely connected organic layers**
7. **DC**: Drought Code (from 7.9 to 860.6) - **moisture content of deep, compact, organic layers**
8. **ISI**: Initial Spread Index (from 0 to 56.10) - **rate of fire spreading at its beginning**
9. **TEMP**: **Temperature(Celsius) (from 2.2 to 33.30)**
10. **RH**: **Relative humidity(%) (from 15.0 to 100)**
11. **WIND**: **Wind speed(km hr-1) (from 0.40 to 9.40)**
12. **RAIN**: **Rain(mm) (from 0.0 to 6.4)**
13. **BURNED AREA**: **Total burned area(ha) (from 0 to 1090.84)**

##### Below shows the first six rows of the forest fire dataset.

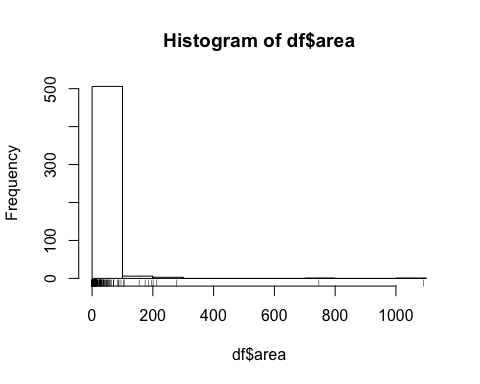
## # A tibble: 6 x 13  
## X Y month day FFMC DMC DC ISI temp RH wind rain  
## <int> <int> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <int> <dbl> <dbl>  
## 1 7 5 mar fri 86.2 26.2 94.3 5.10 8.20 51 6.70 0   
## 2 7 4 oct tue 90.6 35.4 669 6.70 18.0 33 0.900 0   
## 3 7 4 oct sat 90.6 43.7 687 6.70 14.6 33 1.30 0   
## 4 8 6 mar fri 91.7 33.3 77.5 9.00 8.30 97 4.00 0.200  
## 5 8 6 mar sun 89.3 51.3 102 9.60 11.4 99 1.80 0   
## 6 8 6 aug sun 92.3 85.3 488 14.7 22.2 29 5.40 0   
## # ... with 1 more variable: area <dbl>

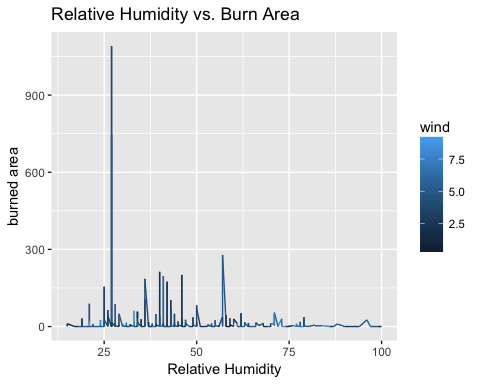
#### Summary

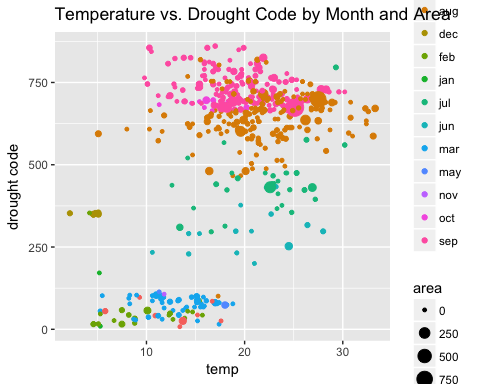
## X Y month day   
## Min. :1.000 Min. :2.0 Length:517 Length:517   
## 1st Qu.:3.000 1st Qu.:4.0 Class :character Class :character   
## Median :4.000 Median :4.0 Mode :character Mode :character   
## Mean :4.669 Mean :4.3   
## 3rd Qu.:7.000 3rd Qu.:5.0   
## Max. :9.000 Max. :9.0   
## FFMC DMC DC 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

### 2. Exploratory Data Analysis and Visualizations

#### We can use ggplot2 to better visualize the data, see below sample histogram, relationship between humidity and burn area, as well as the correlations between temperature, drought code, area and month:







#### 2A. Factor Analysis using KMO Test

The next step is then to carefully examine the data variables and determine which ones are most/relatively more important given a number of potential causes. **Factor Analysis** method will play a vital role in this step. Factor analysis is the most widely used multivariate technique to desribe variability among observed, correlated variables in terms of potentially lower number of unobserved variables.It is a statisical method for dimension reduction.

Factor analysis requires numeric input, the dataset needs to be cleaned/transformed - any character types would be converted to numeric type. Hence, we convert “**month**” and “**day**” to numbers as shown in below code:

df$month <- as.numeric(df$month)  
df$day <- as.numeric(df$day)

This however will give both variables as **NA** values. A second attempt to transform the dataset is then carried out in below code.

url <- "https://archive.ics.uci.edu/ml/machine-learning-databases/forest-fires/forestfires.csv"  
df1 <- read.csv(url)  
fires$month <- as.numeric((fires$month))  
fires$day <- as.numeric(fires$day)  
library(dplyr)  
head(fires)

**The result is shown as below:**

## X Y month day FFMC DMC DC ISI temp RH wind rain area  
## 1 7 5 8 1 86.2 26.2 94.3 5.1 8.2 51 6.7 0.0 0  
## 2 7 4 11 6 90.6 35.4 669.1 6.7 18.0 33 0.9 0.0 0  
## 3 7 4 11 3 90.6 43.7 686.9 6.7 14.6 33 1.3 0.0 0  
## 4 8 6 8 1 91.7 33.3 77.5 9.0 8.3 97 4.0 0.2 0  
## 5 8 6 8 4 89.3 51.3 102.2 9.6 11.4 99 1.8 0.0 0  
## 6 8 6 2 4 92.3 85.3 488.0 14.7 22.2 29 5.4 0.0 0

In order to find the relevant variables, a **KMO** test is needed to answer the question.KMO stands for **Kaiser-Meyer-olkin** test. It’s a measure of the proportion of variance among variables that might be a common variances. **The lower the proportion, the more suited your data is to Factor Analysis**.

#### Checking adequacy of factor analysis

There are two major criteria to check the adequacy of the factor analysis to help identify more relevant variables.

**1. Criteria of sample size adequacy**: sample size of 300 and above is good, 500 and more is considered very good.In our dataset, the sample size is 517, which implies it is suitable for factor analysis.

**2. KMO’s sampling adequacy criteria with** MSA(individual measures of sampling adequacy of each variable)**:** The range of KMO is from 0.0 to 1.0 and if the calculated percentage is > 0.5, the variable is desired value. Variables with MSA being < 0.5 indicate that items do not belong to a group and may be removed from the factor analysis.

To successfully perform KMO test, a R package named Psych is installed and used with the following code:

library(psych)  
fires\_corr <- cor(fires)  
KMO(fires\_corr)

**The result shows that the overall MSA is 0.57 which is greater than 0.5 that is desired value.**

## Kaiser-Meyer-Olkin factor adequacy  
## Call: KMO(r = fires\_corr)  
## Overall MSA = 0.57  
## MSA for each item =   
## X Y month day FFMC DMC DC ISI temp RH wind rain   
## 0.51 0.50 0.27 0.66 0.72 0.59 0.58 0.67 0.63 0.41 0.52 0.44   
## area   
## 0.61

Based on the table shown above, we can eliminate **MONTH**, **RH (Relative Humidity)**,and **RAIN** and keep **X**, **Y**, **DAY**, **FFMC**, **DMC**, **DC**, **ISI**, **WIND** and **AREA** for further metric evalulation.