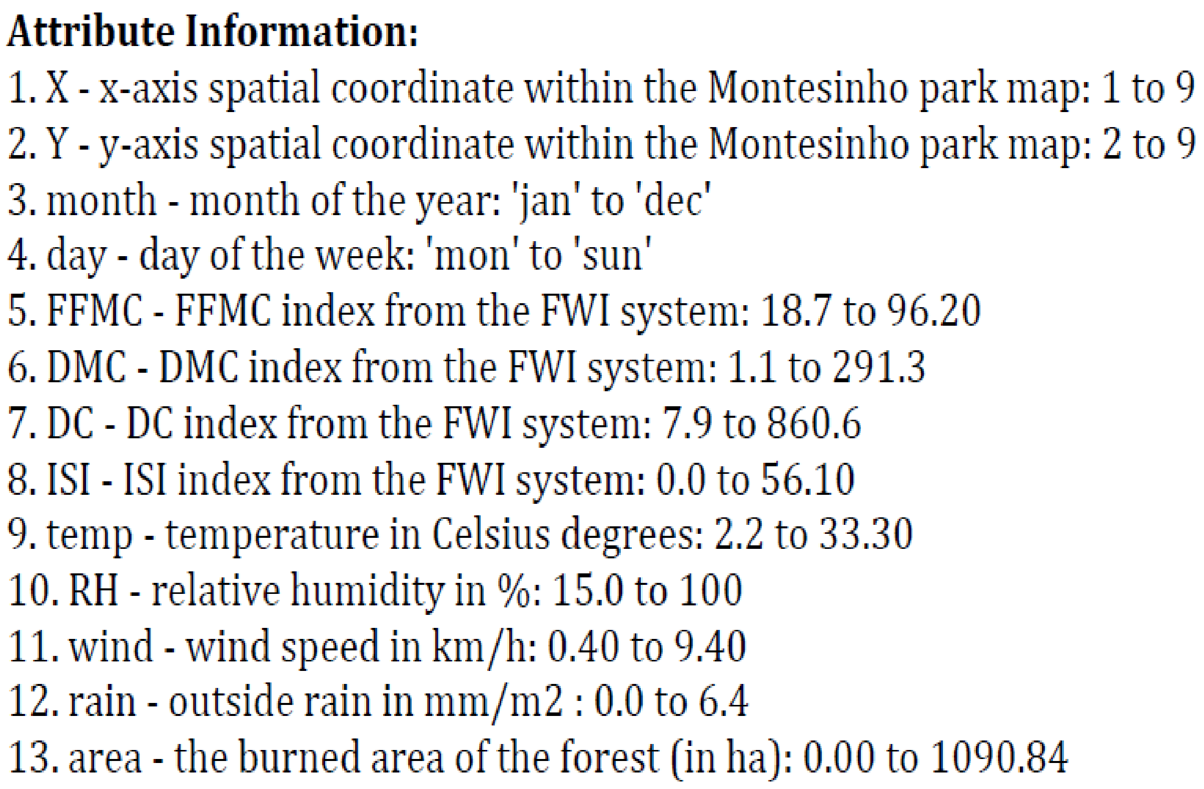
Practical DataScience
(Data Manipulation)
Assignment-2

Import the datafile “forestfires.csv” available at algorithmica github repository to answer following questions. Description of each attribute is given as follows:



Attribute Information:

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 - FFMC index from the FWI system: 18.7 to 96.20
6. DMC - DMC index from the FWI system: 1.1 to 291.3
7. DC - DC index from the FWI system: 7.9 to 860.6
8. ISI - ISI index from the FWI system: 0.0 to 56.10
9. temp - temperature in Celsius degrees: 2.2 to 33.30
10. RH - relative humidity in %: 15.0 to 100
11. wind - wind speed in km/h: 0.40 to 9.40
12. rain - outside rain in mm/m2 : 0.0 to 6.4
13. area - the burned area of the forest (in ha): 0.00 to 1090.84

Practical DataScience
(Data Manipulation)
Assignment-2

1. Compute the following:

   How many observations are there in the dataset?
   How many observations are there with a fire (i.e. *area*>0)?
   How many observations are there with rain (i.e. *rain*>0)?
   How many observations are there with both a fire and rain?

2. Show the columns *month, day, area* of all the observations.
   Show the columns *month, day, area* of the observations with a fire.

3. How large are the five largest fires (i.e. having largest *area*)?
   What are the corresponding *month, temp, RH, wind, rain, area*?

4. Reorder factor levels of *month* to be from Jan to Dec.
   Add one column to the data indicating whether a fire occurred for each observation ('TRUE' for *area*>0 and 'FALSE' for *area*==0).

5. What is the mean *area/wind/temp/RH* per *month*?
   How many observations are there in each month?
   How many observations are there with a fire in each month?
   What is the probability of a fire in each month?