**Case Study of forestfires using SVM**

**Problem Statement:**

classify the Size\_Categorie using SVM

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 96.20

DMC DMC index from the FWI system: 1.1 to 291.3

DC DC index from the FWI system: 7.9 to 860.6

ISI ISI index from the FWI system: 0.0 to 56.10

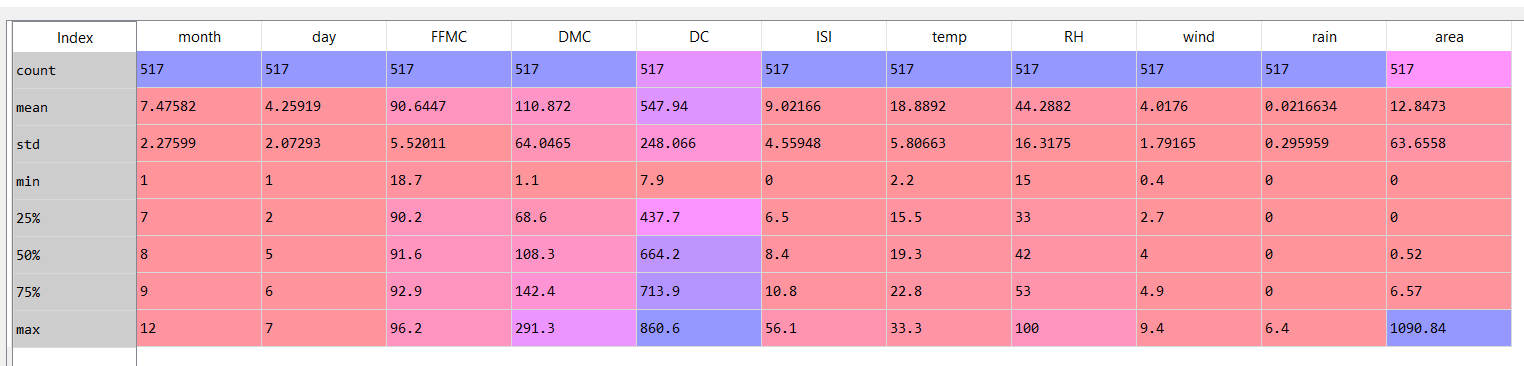
temp temperature in Celsius degrees: 2.2 to 33.30

RH relative humidity in %: 15.0 to 100

wind wind speed in km/h: 0.40 to 9.40

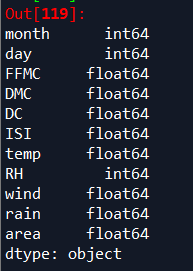
rain outside rain in mm/m2 : 0.0 to 6.4

Size\_Categorie the burned area of the forest ( S**Exploratory data analysis:**

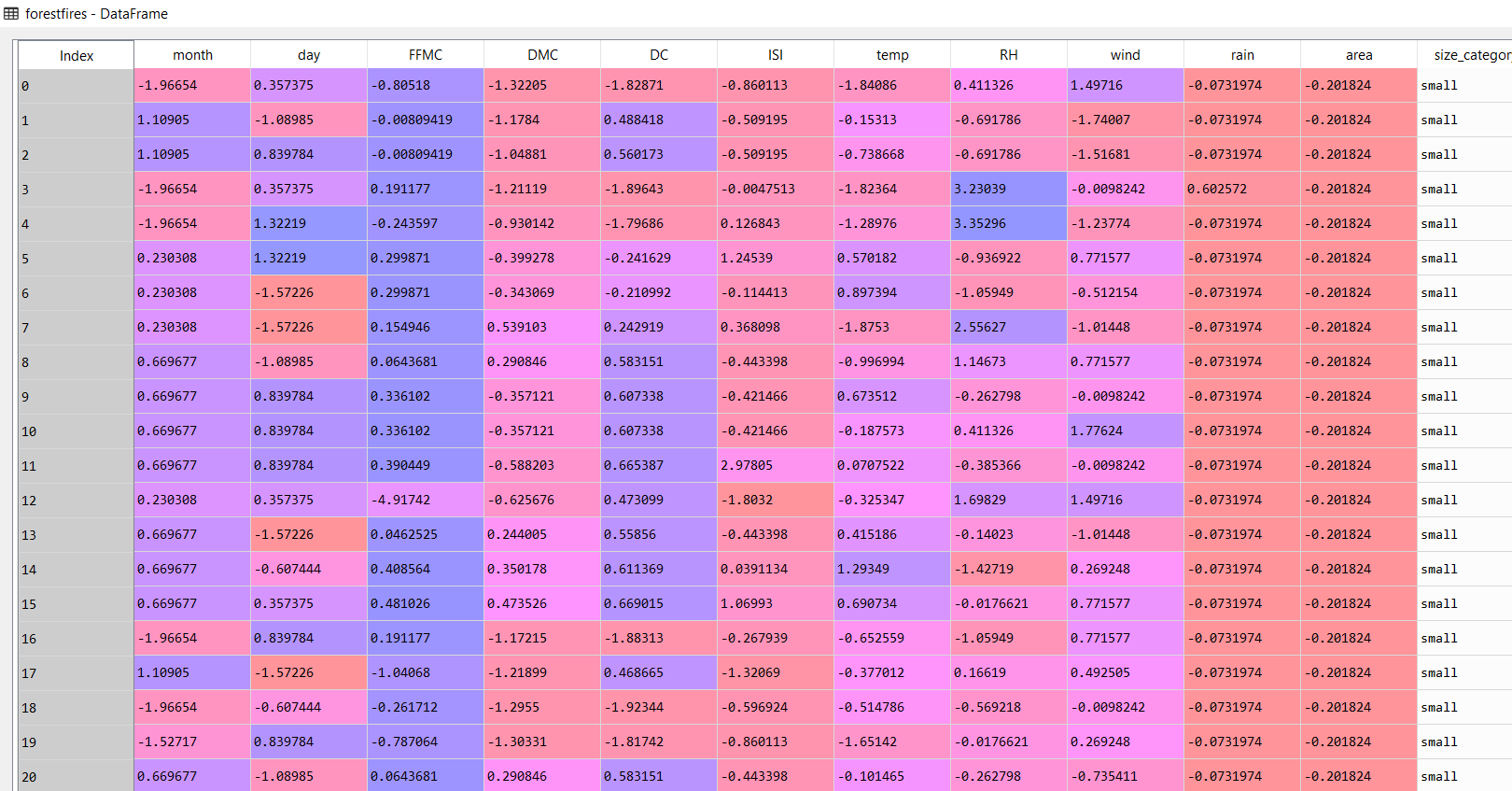


The above is the descriptive statistics of the forestfire dataset. There are 517 observations and 11 variables in the data frame. The missing variable in the above descriptive analysis is the “State” variables. Which is a categorical variable.

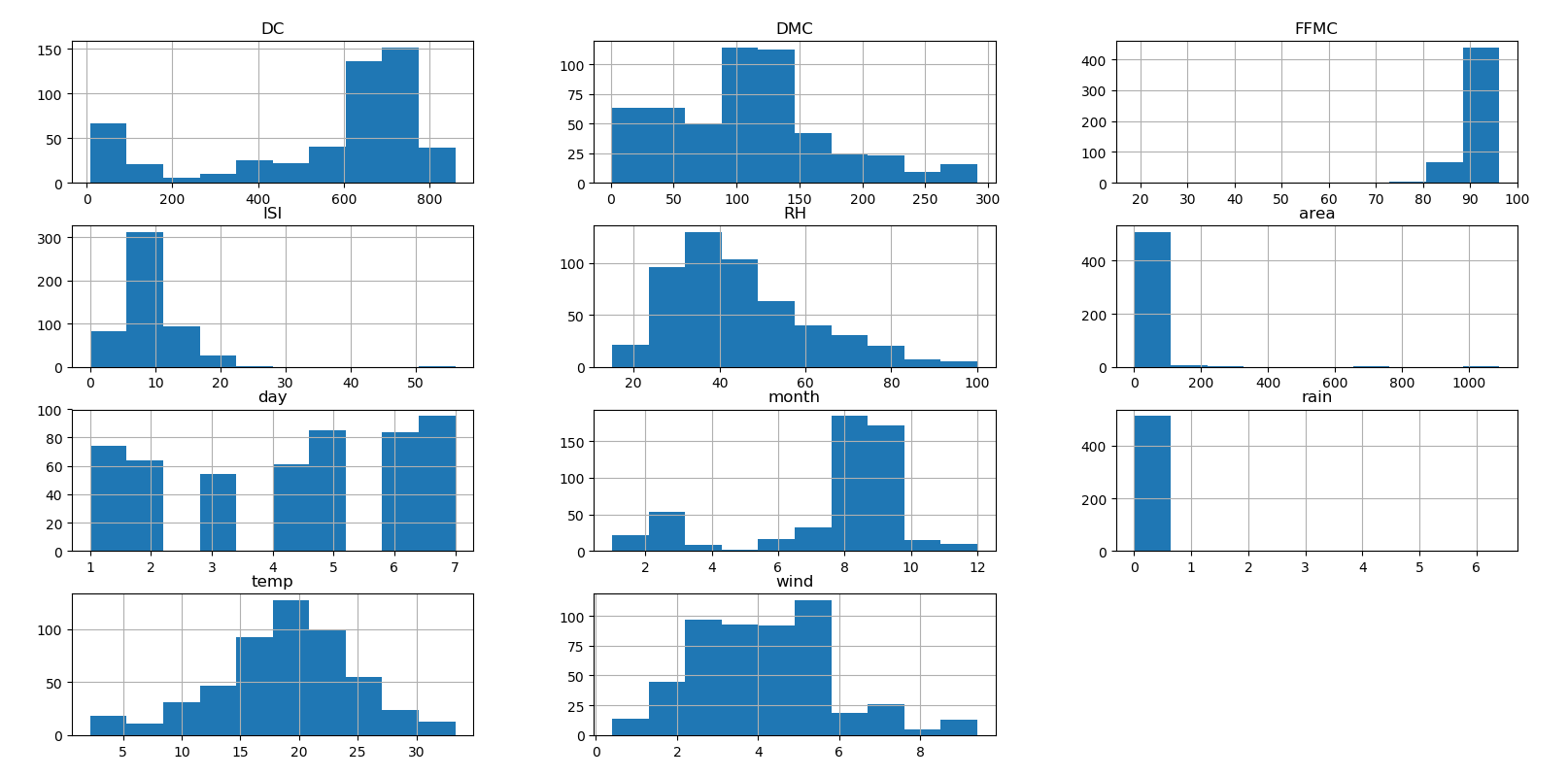
The following are the variable types:



Due to a large difference in the ranges of the values of al the features, we normalize the dataset as follows:

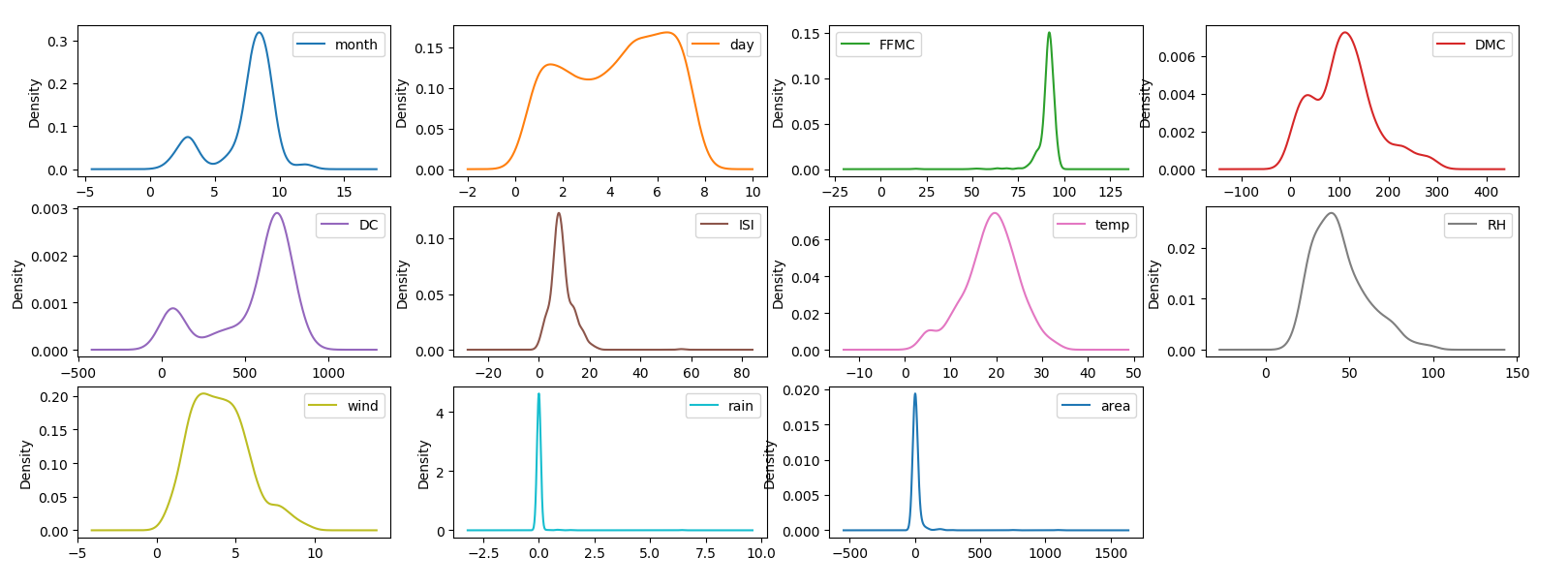


The following are the histograms of the dataset:

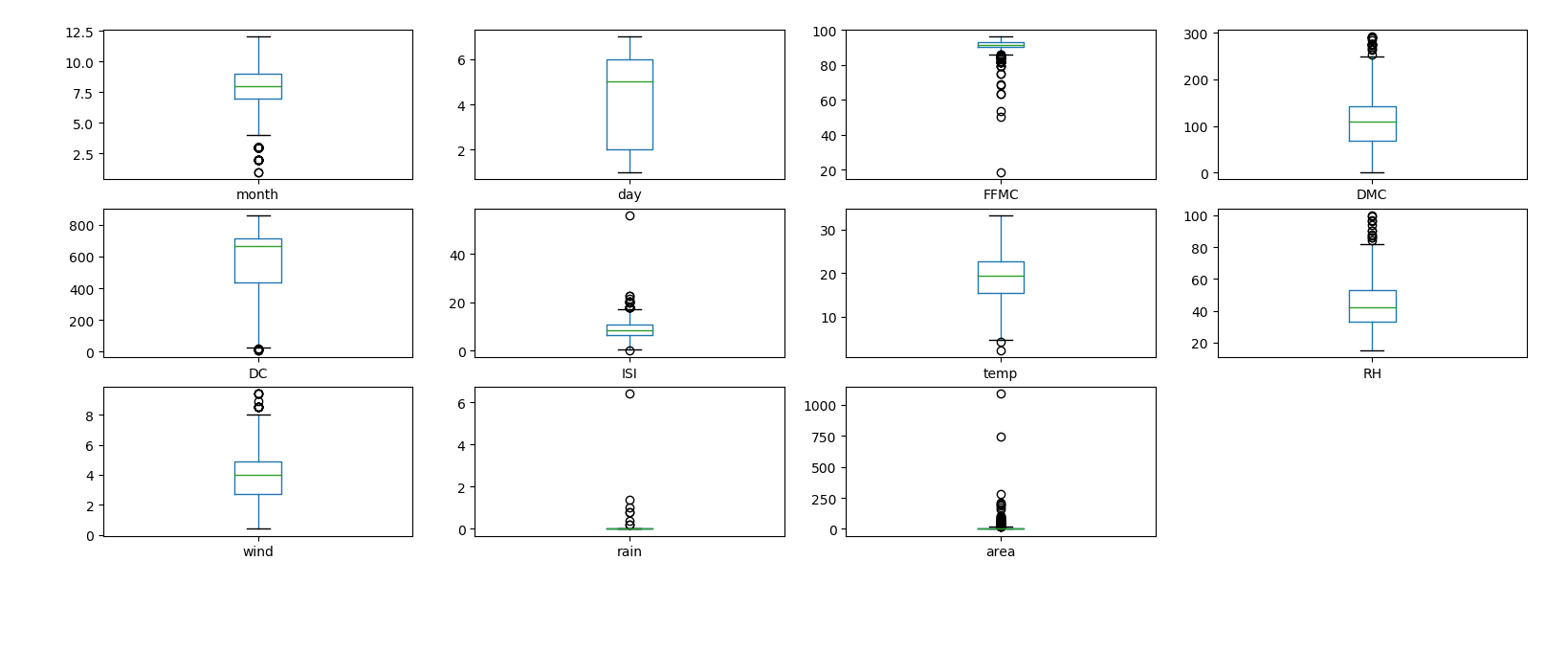


The variables ‘temp’ has near Gaussian distribution or near normal distribution. The rest of the histograms are either positively or negatively skewed.

The following is the continuous density function:

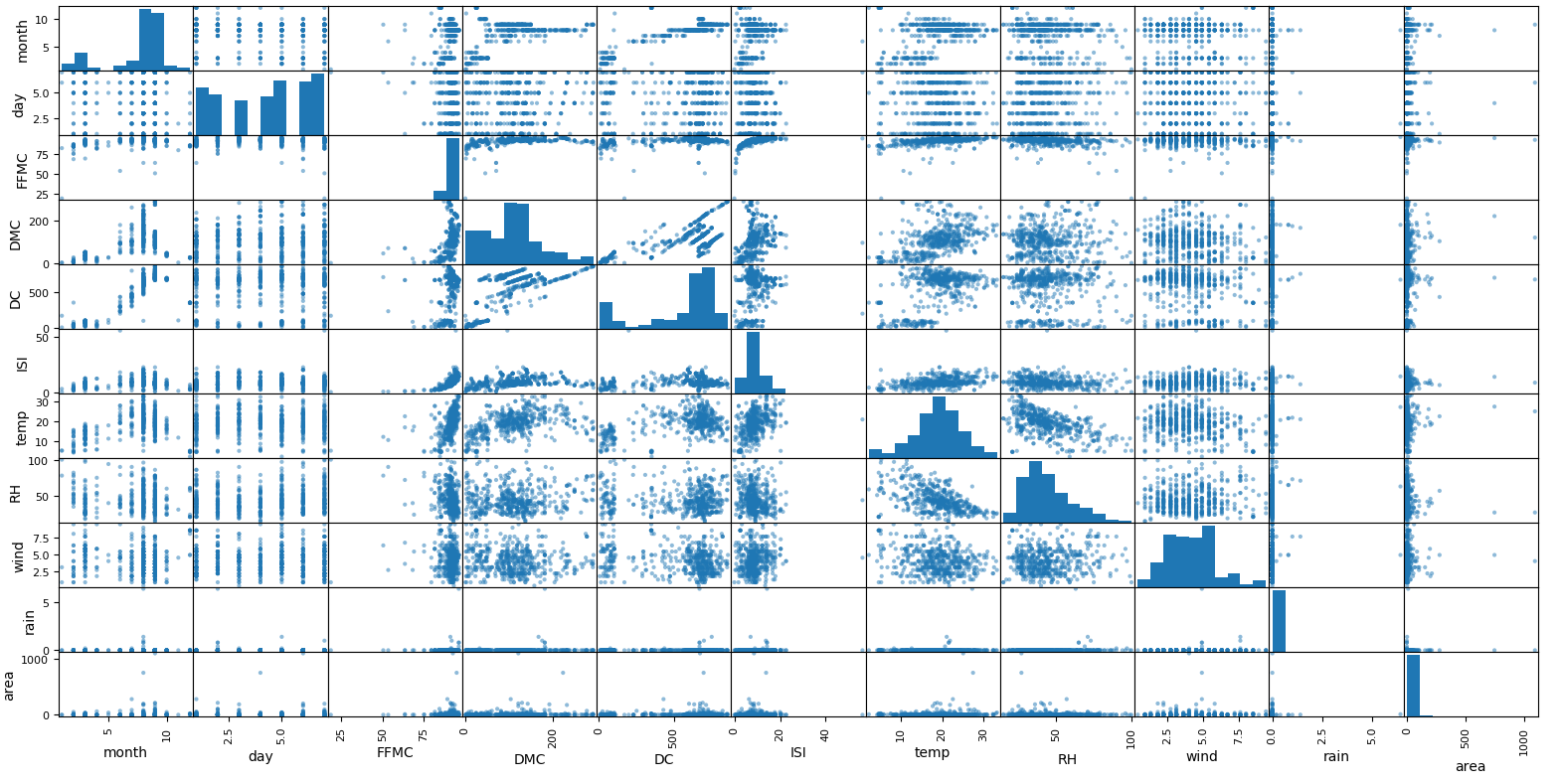


The following are the boxplots of the variables

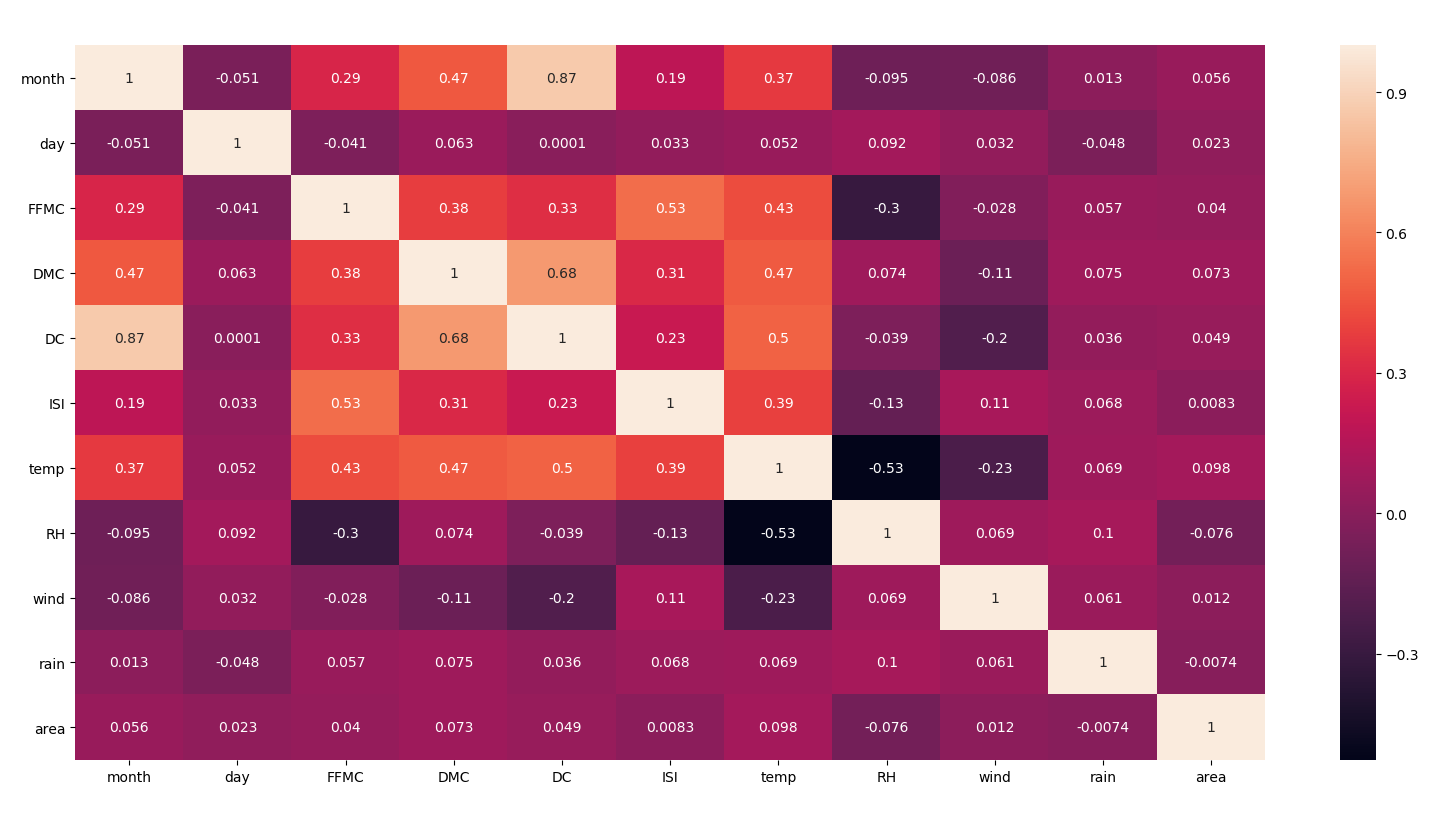


The variables FFMC, month, DMC, DC, ISI, RH, wind, rain and area have statistical outliers

The following is the pair plot or the scatter matrix of the variables:



The heat map of the correlation between different variables

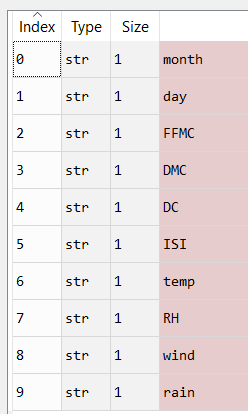


From the above heat map we can estimate the following:

* High negative correlation between month and DC
* The correlation between rest of the independent variables are not significant enough to consider Multi Co linearity problem in the given dataframe.

Building the sequential neural network model:

The following variables are set as the predictors:



We feed the data through the Support Vector Classifier SVC imported from the sklearn package

With the help of different kernels we will estimate the accuracy and select the best model

**Result of Kernel = linear**



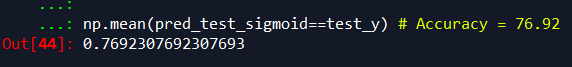
**Result of Kernel = poly**



**Result of Kernel = rbf**



**Result of Kernel = sigmoid**



Of the above model the model with the kernel linear provides the highest accuracy with **91%**