Project 1 - DS-600

## Introduction

Every year, there are many forest fires occur worldwide. To protect the natural wildlife, numerous attempts has been made to prevent forests from getting fire. In this project, we will be outlining a summary in about a forest fire that occurred in a region at Portugal. The data is collected by using metheorological and other data. For more information about the dataset, this [link](http://www3.dsi.uminho.pt/pcortez/forestfires/forestfires-names.txt) can be followed.

The following is the dataset information.

# reading   
forestfirest = read.csv('forestfires.csv')  
head(forestfirest)

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

## Data Cleaning

With the dataset, we will first be investigating the variables’ data type and general shape of occurance. We will be examining data in terms of the following objectives:

1. Variable data types
2. Empty or NA values

Below, are presenting the general outline of the dataset.

str(forestfirest)

## 'data.frame': 517 obs. of 13 variables:  
## $ X : int 7 7 7 8 8 8 8 8 8 7 ...  
## $ Y : int 5 4 4 6 6 6 6 6 6 5 ...  
## $ month: Factor w/ 12 levels "apr","aug","dec",..: 8 11 11 8 8 2 2 2 12 12 ...  
## $ day : Factor w/ 7 levels "fri","mon","sat",..: 1 6 3 1 4 4 2 2 6 3 ...  
## $ FFMC : num 86.2 NA NA 91.7 89.3 92.3 92.3 91.5 91 92.5 ...  
## $ DMC : num 26.2 35.4 43.7 33.3 51.3 ...  
## $ DC : num 94.3 669.1 686.9 77.5 102.2 ...  
## $ ISI : num 5.1 6.7 6.7 9 9.6 14.7 8.5 10.7 7 7.1 ...  
## $ temp : num 8.2 18 14.6 8.3 11.4 22.2 24.1 8 13.1 22.8 ...  
## $ RH : int 51 33 33 97 99 29 27 86 63 40 ...  
## $ wind : num 6.7 0.9 1.3 4 1.8 5.4 3.1 2.2 5.4 4 ...  
## $ rain : num 0 0 0 0.2 0 0 0 0 0 0 ...  
## $ area : num 0 0 0 0 0 0 0 0 0 0 ...

We observed that FFMC variable has some missing values. Therefore, we decided to update values with replacing NAs with mean of it.

#replacing empty data with 0  
#forestfirest$FFMC = str\_replace(forestfirest$FFMC,"","0")  
  
# finding indices of na  
ind = which(is.na(forestfirest$FFMC))  
  
#replacing empty data with 0  
forestfirest$FFMC[ind]=0  
  
#replacing empty data with 0  
forestfirest$FFMC[ind]=mean(forestfirest$FFMC)

TODO: 1. use dplyr and tidyr package. 2. use a few functions in that packages 3. lecture notes, find sth from that notes that we can use on our dataset