Report 3

Dataset: 'Algerian Forest Fire'

Dataset Source: https://archive.ics.uci.edu/

Introduction: The Algerian Forest Fire dataset meets the assignment requirements with more than 10 attributes and 100 instances. I have selected this dataset because it primarily concerns my country and I have seen the tremendous damage caused to the citizens, so I intend to explore the key influencing factors triggering these recurrent fires.

Data Description: The data included in this multivariate dataset concerns two regions in Algeria called Bejaia and Sidi Bel-Abbes that were affected by the forest fires. There are 12 variables described as follows:

Date: (DD/MM/YYYY)

Temp: temperature noon (temperature max) in Celsius degrees: 22 to 42

RH: Relative Humidity in %: 21 to 90

Ws: Wind speed in km/h: 6 to 29

Rain: Total day in mm: 0 to 16.8

FWI Components

(FFMC): Fine Fuel Moisture Code index from the FWI system: 28.6 to 92.5

(DMC): Duff Moisture Code index from the FWI system: 1.1 to 65.9

(DC): Drought Code index from the FWI system: 7 to 220.4

(ISI): Initial Spread Index index from the FWI system: 0 to 18.5

(BUI): Buildup index from the FWI system: 1.1 to 68

(FWI): Fire Weather Index Index: 0 to 31.1

Classes: two classes, namely 'Fire' and 'Not Fire'

Note: the dataset does not include missing values.

Context: The fire is studied through the Fire Weather Index (FWI) which is based on important subindices measurement organized into multiple layers in a given region. (FFMC) measures moisture content for the superior layer impacted by meteorological factors, (DMC) measures the moisture content of decomposed organics in a medium layer, (DC) dryness in a deeper organic layer, and (BUI) asses the amount of combustible fuel. All indices represent an assessment of potential fire danger. Hence, (ISI) is a rate for expecting a fire spread.

Data validation:

The subsequent table represents a checklist for data validation including cleaning if needed to prepare the data qualitatively.

Checkpoints	Validity (V/NV)	Number of issues	Observation	Handling	Result
Readability	NV	2	Regions data were organized subsequently.	Assigning correspondent data indexes to each region	Structured dataset

Type conversion	NV	1	Invalid Date Format	Merging day, month, year in a uniform date format	Date column
Missing data	V	0	-	-	-
Incoherent data	NV	1	Classes variable instances	Transform into Binary format (fire, not fire)	Coherent representation

Data exploration:

- Structure of dataset

```
'data.frame':   243 obs. of  13 variables:
 $ Temperature: int  29 29 26 25 27 31 33 30 25 28 ...
 $ RH         : int  57 61 82 89 77 67 54 73 88 79 ...
 $ Ws         : int  18 13 22 13 16 14 13 15 13 12 ...
 $ Rain       : num  0 1 13 2 0 0 0 0 0 0 ...
 $ FFMC       : num  65 64 47 28 64 82 88 86 52 73 ...
 $ DMC        : num  3 4 2 1 3 5 9 12 7 9 ...
 $ DC         : num  7 7 7 6 14 22 30 38 38 46 ...
 $ ISI        : num  1 1 0 0 1 3 6 5 0 1 ...
 $ BUI        : num  3 3 2 1 3 7 10 13 10 12 ...
 $ FWI        : num  0 0 0 0 0 2 7 7 0 0 ...
 $ Classes    : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 2 2 1 1 ...
 $ Region     : Factor w/ 2 levels "Bejaia","Sidi Bel-abbess": 1 1 1 1 1 1 1 1 1 1 ...
 $ date       : Date, format: "2012-06-01" "2012-06-02" "2012-06-03" "2012-06-04" ...
```

There are 2 factor variables and 10 numerical variables.

- Univariate relationships:

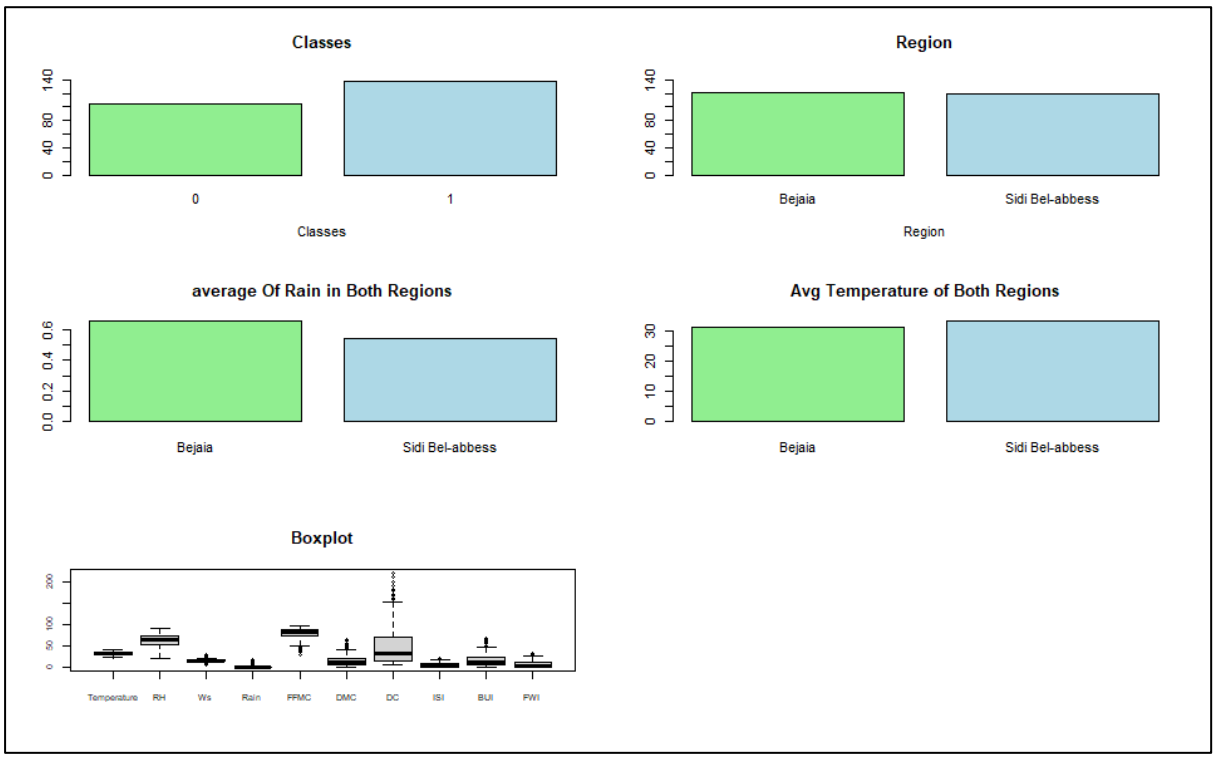

Figure 1. Univariate relationships

By visual inspection, Although the dataset records equal instances for both regions, we can see that the region Sidi Bel-abbes had more fires, with less average rain, and tends to have a quieter more temperature average than the Bejaia region. The boxplot indicates that (DC) includes considerable potential outliers.

- Bivariate relationships inspection:

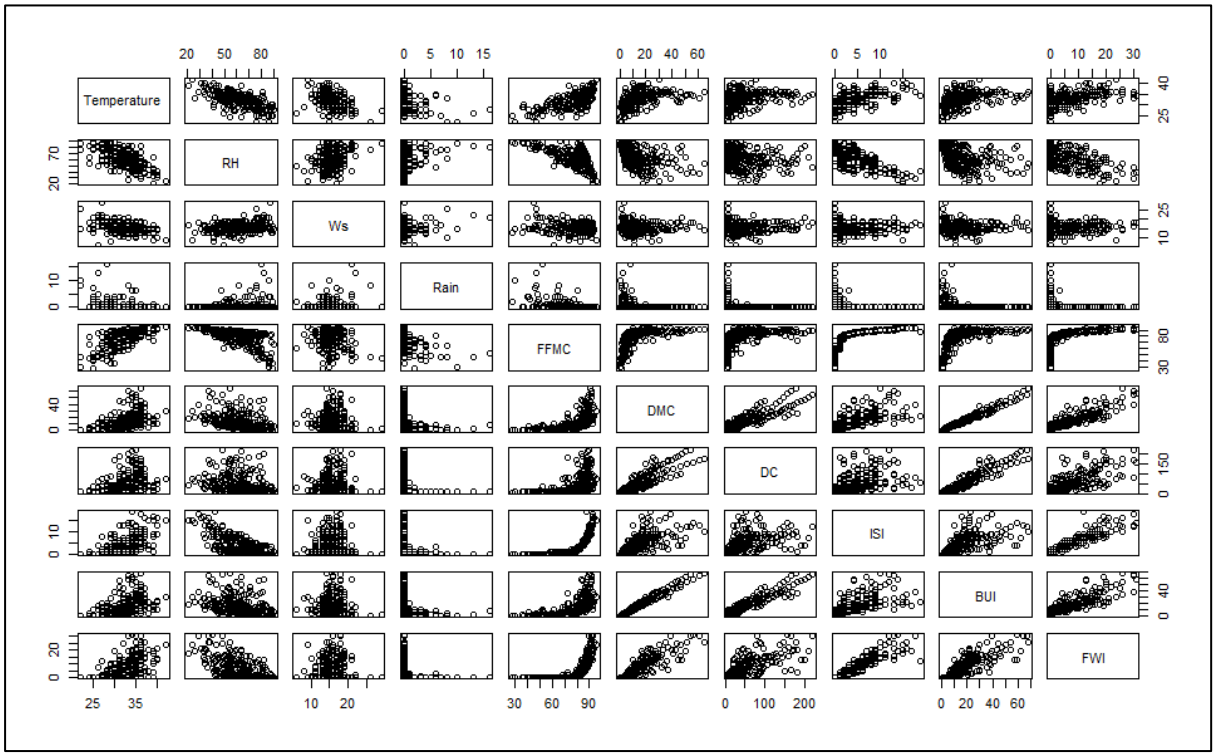

Figure 2. Pairs plot

The pairs plot illustrates important bivariate relations, where the Sub indices of FWI as explained in the context above, all show a positive variance between them (BUI, ISI, DC, DMC, FMC) which is reasonable that the indexes all increase together. Alternatively, we can see that temperature is positively related to FWI, but RH and FWI show negative interactions. So, a correlation inspection is needed for the analysis of these relations.

- Correlation inspection

The correlation matrix below emphasizes the observations mentioned in the bivariate relationship section, where you see that all the columns that include the FWI and its subindices have a strong positive correlation, while when it comes to variables like temperature, RH, rain, Ws against the fire indices variables, there is a clear negative correlation. This may indicate that the meteorological aspect is the influencer factor of fire incidents.

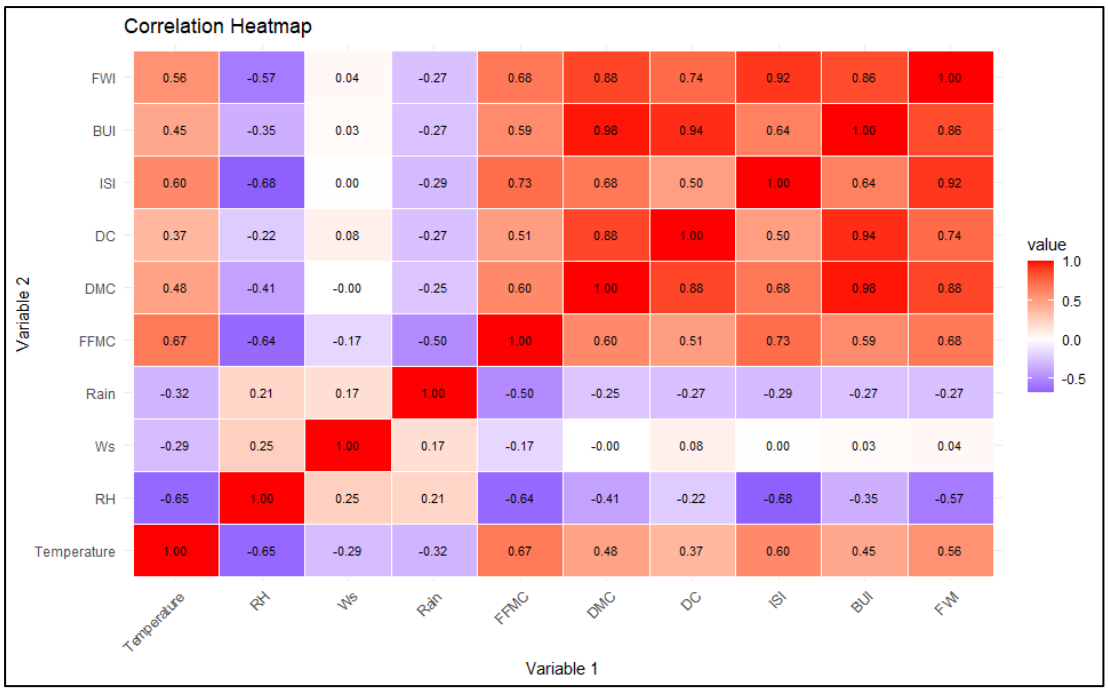

Figure 3. Correlation heatmap matrix

Since that 'temperature' has the highest correlation interacting with the fire indices (FFMC, DMC), and 'Rain' shows a negative correlation, the following figure visualizes the linear models for the mentioned interactions.

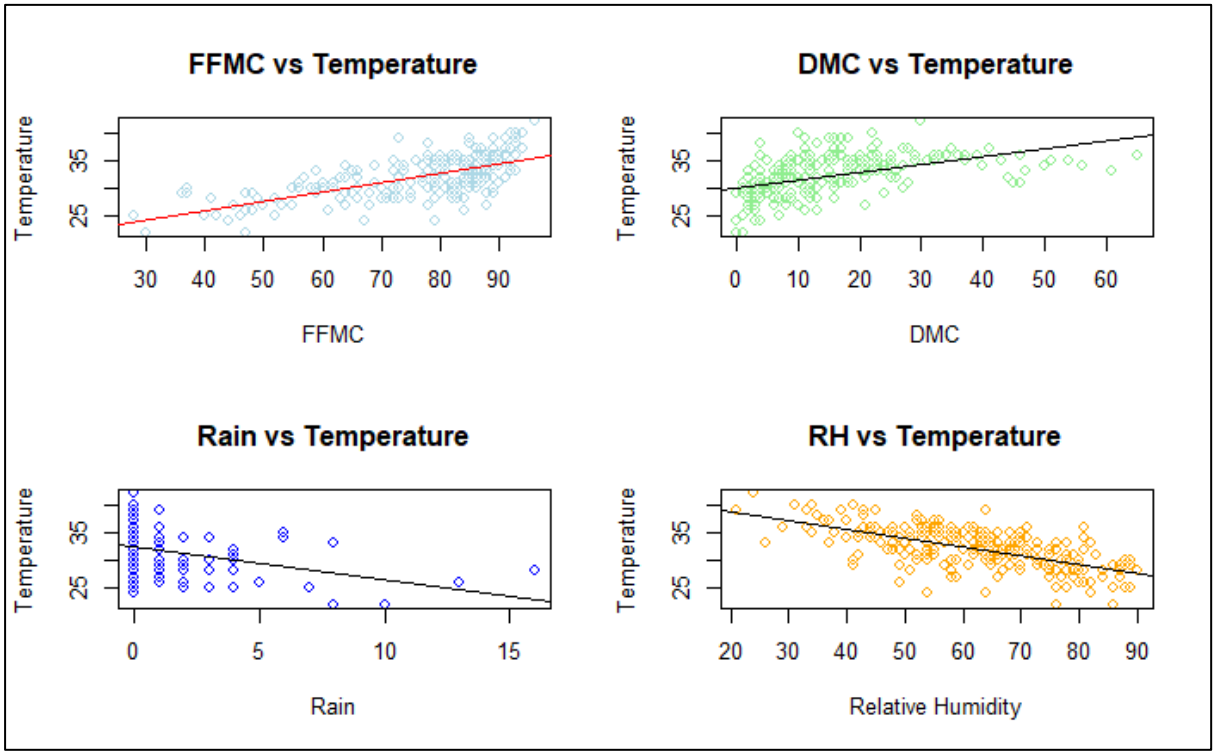

Figure 4. Linear models

Observations: the factors that decrease the temperature are rain and relative humidity, while the factor that increases the temperature are, a higher index of Duff Moisture, and a higher index of Fine Fuel Moisture.

Conclusion: Temperature is the principal factor in triggering forest fires, where the region of Sidi Bel-abbes had more fires than Bejaia, this can be explained by the quite higher temperature in addition to less average rain which implicates directly a decrease in the fire indices that causes by their turn the increase in temperature.

Code:

```
#Merging day, month, and year into a single date variable
df$date <- as.Date(with(df, paste(year, month, day, sep = '-')))
df <- df[, !names(df) %in% c('day', 'month', 'year')]
dim(df)
#Transforming the character 'fire' 'not fire' into binary representation
df$Classes <- ifelse(df$Classes == 'not fire', 0, 1)
```

```
par(mfrow=c(3,2))
for (i in 1:length(object_cols)) {
  barplot(table(df[[object_cols[i]]]), main=object_cols[i], col=c('lightgreen', 'lightblue'), xlab=object_cols[i], ylim=c(0, 150))
}
barplot(tapply(df$Rain, df$Region, mean), col=c('lightgreen', 'lightblue'), main='average Of Rain in Both Regions')
table_data <- table(df$Region, df$Classes)
# Define custom labels for colors
color_labels <- c('Bejaia', 'Sidi belabess')
# Histogram
# Barplot for average temperature in both regions
barplot(tapply(df$Temperature, df$Region, mean), col=c('lightgreen', 'lightblue'), main='Avg Temperature of Both Regions')
boxplot(df[, numeric_cols], main='Boxplot', cex.axis=0.7)
```

```
# Convert correlation matrix to long format
cor_long <- reshape2::melt(cor_matrix)

ggplot(cor_long, aes(Var1, Var2, fill = value, label = sprintf("%.2f", value))) +
  geom_tile(color = "white") +
  geom_text(size = 3, color = "black", show.legend = FALSE) +
  scale_fill_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0, na.value = "grey50") +
  labs(title = "Correlation Heatmap", x = "Variable 1", y = "Variable 2") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
par(mfrow=c(2,2))
# Regression Plots
plot(df$FFMC, df$Temperature, main='FFMC vs Temperature', col='lightblue', xlab = "FFMC",ylab = "Temperature")
abline(lm(df$Temperature ~ df$FFMC), col='red')

plot(df$DMC, df$Temperature, main='DMC vs Temperature', col='lightgreen', xlab = "DMC",ylab = "Temperature")
abline(lm(df$Temperature ~ df$DMC), col='black')

plot(df$Rain, df$Temperature, main='Rain vs Temperature', col='blue', xlab = "Rain",ylab = "Temperature")
abline(lm(df$Temperature ~ df$Rain), col='black')

plot(df$RH, df$Temperature, main='RH vs Temperature', col='orange', xlab = "Relative Humidity",ylab = "Temperature")
abline(lm(df$Temperature ~ df$RH), col='black')
```