Algerian forest fires dataset report

1. Introduction:

This report focus on Algerian Forest Fires Dataset from the UCI Machine Learning Repository. The dataset was collected during summer of 2012, from June to September. This dataset contains information about forest fires in two regions in Algeria, which are Bejaia and Sidi-Bel Abbas.

The objective is to apply the methods and techniques acquired from this module for data exploration.

This database was chosen since there are over 100 instances and more than 10 attributes in the dataset, beside that I am Algerian, and forest fires have been a serious problem in Algeria in recent decades.

2. Dataset description:

The data set have 244 instance and 14 variable: day, month, year, Temperature, relative humidity RH, wind speed Ws, Rain, and several fire danger indices (Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), Build-up Index (BUI) and Fire Weather Index (FWI)). The database also include another column "Classes" which indicates if a fire occurred or not. Additionally, it is divided on two sections: section one for data collected from Bejaia, and section two for data collected from Sidi-Bel Abbas.

The table below gives more information about the variables:

day	Integer from 1 to 31
month	Integer from 6 to 9
year	Integer 2012
Temperature	Integer, from 22 to 42, in Celsius degrees
RH	Integer, from 21 to 90, Relative Humidity in %
Ws	Integer, from 6 to 29, Wind speed in km/h
Rain	Numeric, from 0.0 to 16.80, total rain in mm
FFMC	Numeric, from 28.60 to 96.00
DMC	Numeric, from 0.70 to 65.90
CD	Numeric, from 7.0 to 220.4
ISI	Numeric, from 0.0 to 18.5
BUI	Numeric, from 1.1 to 68.0
FWI	Numeric, from 0.0 to 31.1
Classes	Character, fire or not fire

3. Data cleaning:

This dataset came one csv file but with two tables inside, and they were separated with sentences that indicate which region is this table. I read this csv file as lines and then I removed those separation and empty lines, and asses a new column called "region" to indicates from which region is the data, it takes 2 variable: Bejaia or Sidi-Bel Abbas. At the end, I got one database with 243 row and 15 column. After that, I transformed data type of DC and FWI from character to numeric. I checked how many missing data there was: I found only one in both DC and FWI, so I removed the rows where there was N/As.

4. Data exploration:

4.1. Weather data observation:

According to the histograms below in figure 1, the temperature distribution is centred between 25 to 35 degrees, overall temperature is high. For humidity, most observations are between 50-80%, which means there was high levels of humidity, with left tail for humidity less than 20% which is associated with high risk of fire.

For wind speed, most values are between 10-20km/h and a peak of 15 km/h approximately. High wind speed can accelerate fire spread. Most values of rain are close to 0 mm which indicates that during the 4 months there was a lack of rain, which contributes to fire risk.

4.2. Other attributes observation:

FFMC has long left tail with a peak around 80 of frequency. Low FFMC values indicate flammable conditions and dryness, which increase the probability of fire occurred. For all the other factors, the distributions are right skewed, which indicates dry conditions that would definitely increase the risk of fire.

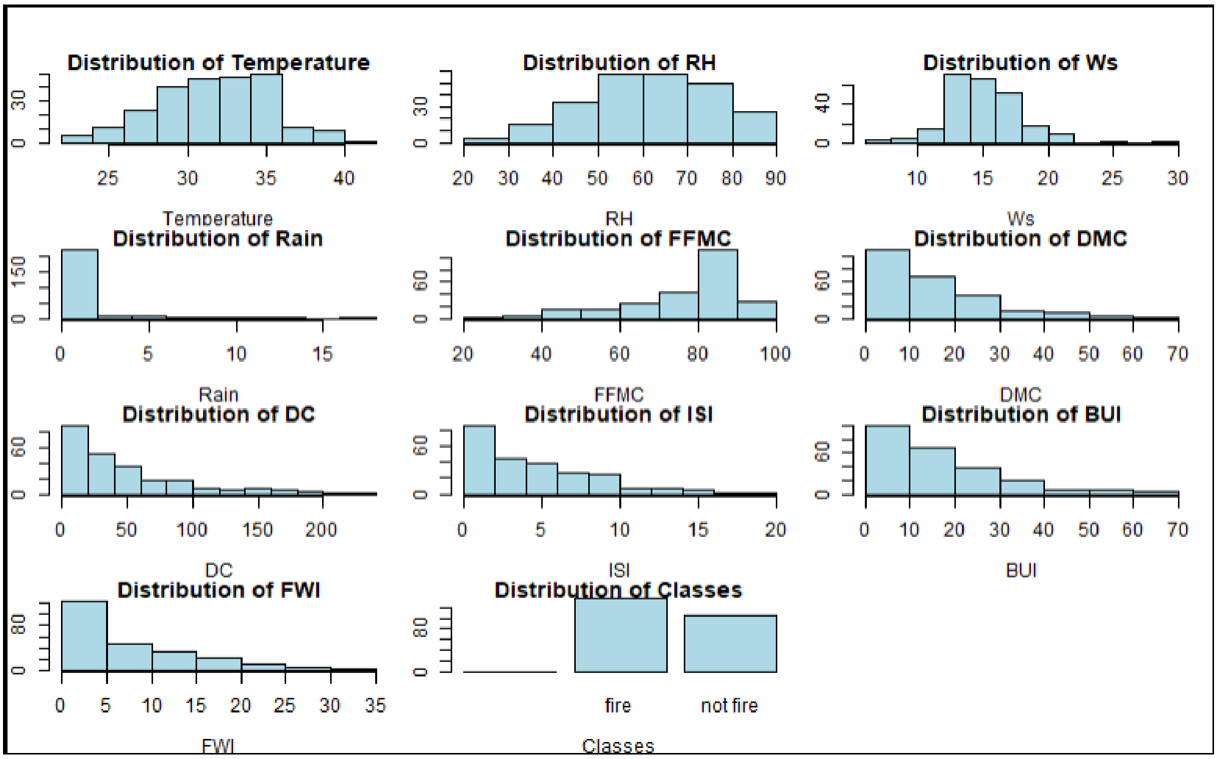
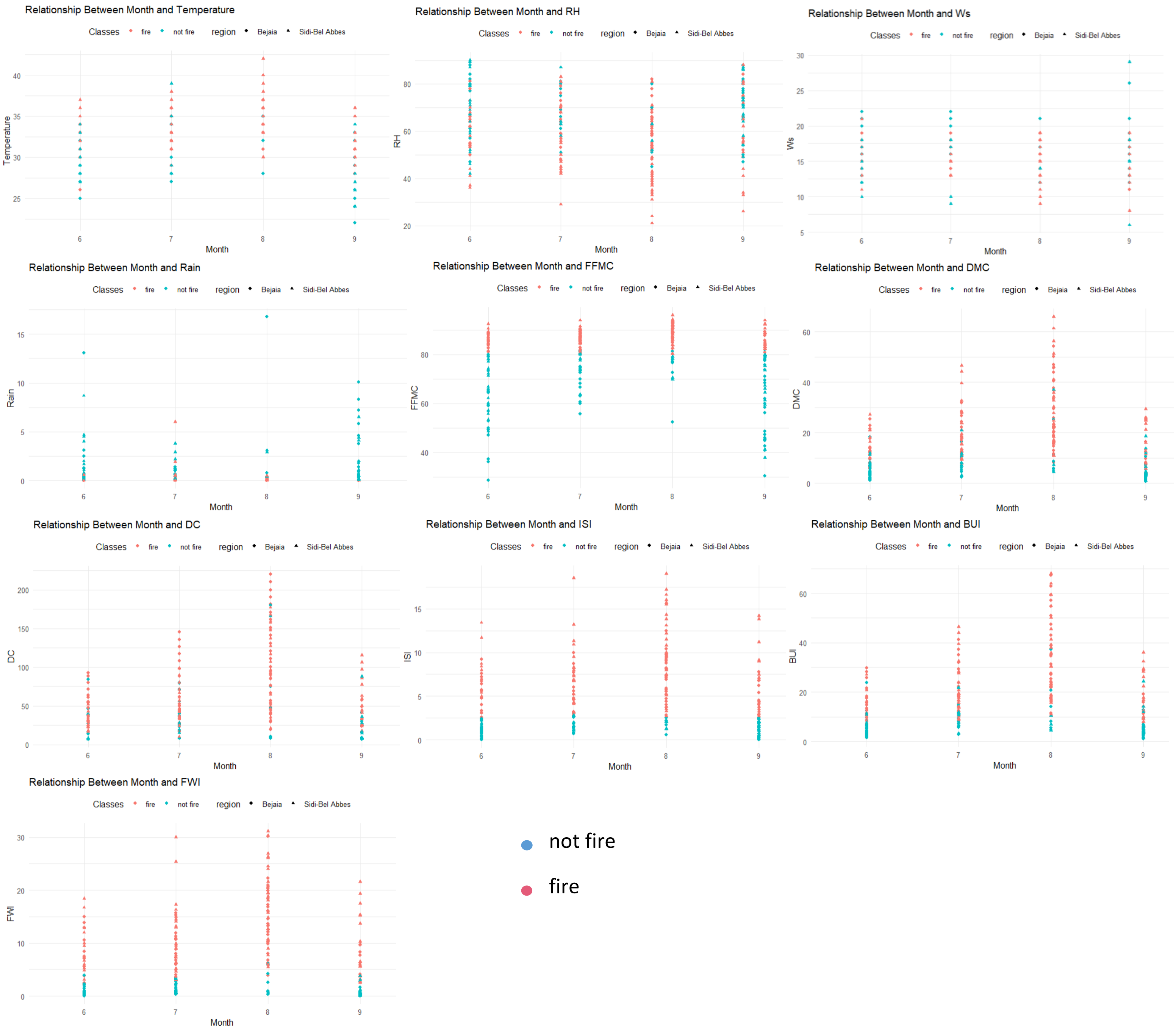


Figure1: distribution of attributes

4.3. Explore month vs other attributes:

The plots in figure 2 clearly show that temperature and other FWI components peaked in july and august, while the rain was the lowest in these two months which make the risk of fires higher. And figure 3 confirms that; it indecates that July and august have high number of fires.



Relationship Between Month and Temperature

Relationship Between Month and RH

Relationship Between Month and Ws

Relationship Between Month and Rain

Relationship Between Month and FFMC

Relationship Between Month and DMC

Relationship Between Month and DC

Relationship Between Month and ISI

Relationship Between Month and BUI

Relationship Between Month and FWI

Figure 2: relation between month and attributes

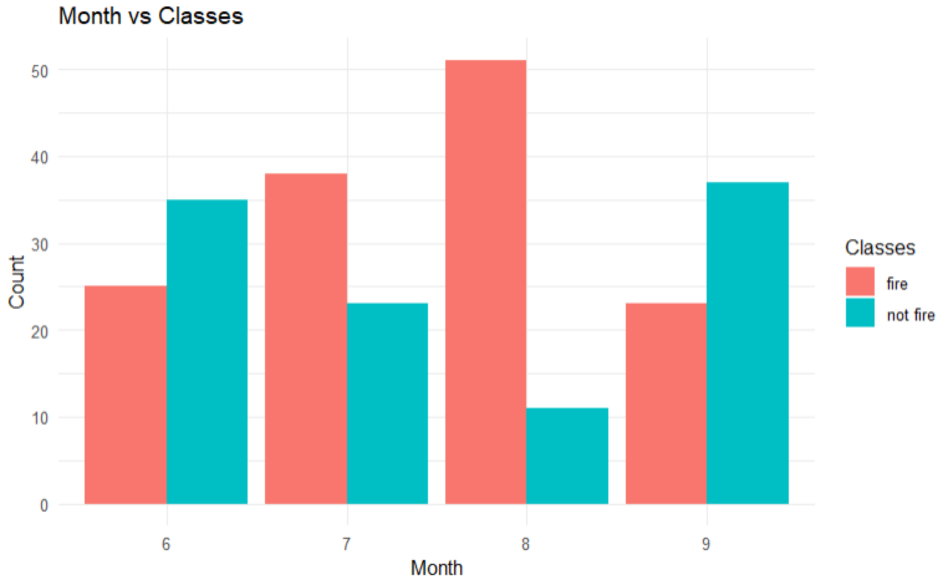


Figure 3: bar plot of number of fires in each month

4.4. Explore correlation between attributes:

As the pairs plot shows in figure 5, there is correlation between almost all the attributes, but the strongest and most significant ones are between temperature and humidity (RH), FFMC and ISC. The higher temperature goes, the higher FFMC and ISC goes, while RH will become less. And the higher FFMC goes, each of DMC, DC, ISI, and BUI goes also higher. Also there was a strong relation between the others factors such as FFMC, DMC, DC, ISI, BUI and FWI; ones any of them increase, the other one will also increase.

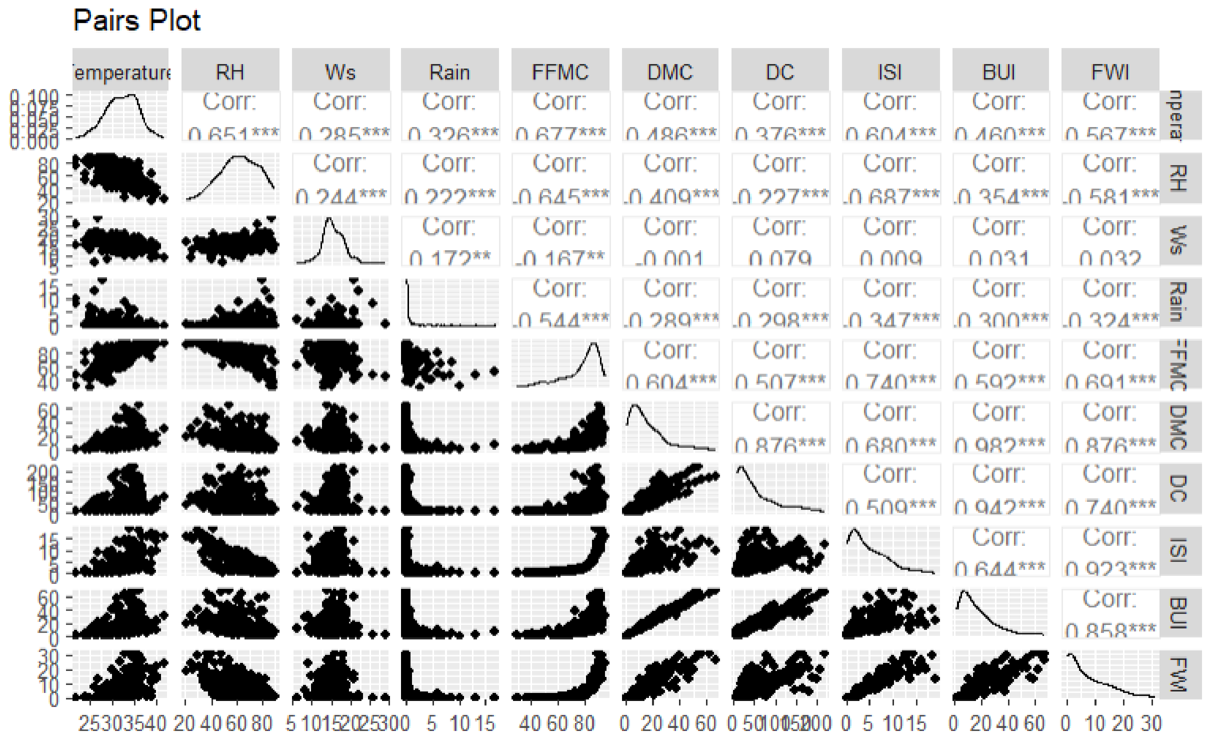


Figure 4: Pairs plot to demonstrate correlation between attributes

4.5. Time series:

Figure 5 below, shows the different attributes by time. For the variables DC, BUI, FWI and DMC in these plots, we can confirm again that there is a high relation between them, I can say that they are identic in their plots. For temperature, it was hot in general with a significant increase in temperature by time, to reach the peak in mid-august and start decreasing after that. For the winds speed, it was between 10 to 20 km/h from June to the beginning of September and start to increase after that. The quantity of rain was decreasing from June to July to be almost don't exist till September. FFMC was getting higher by time and decrease again in September. ISI was fluctuating all the time but the highest values were recorded between mid-July to mid-august. In the contrast, humidity was fluctuating but recorded the lowest levels in august.

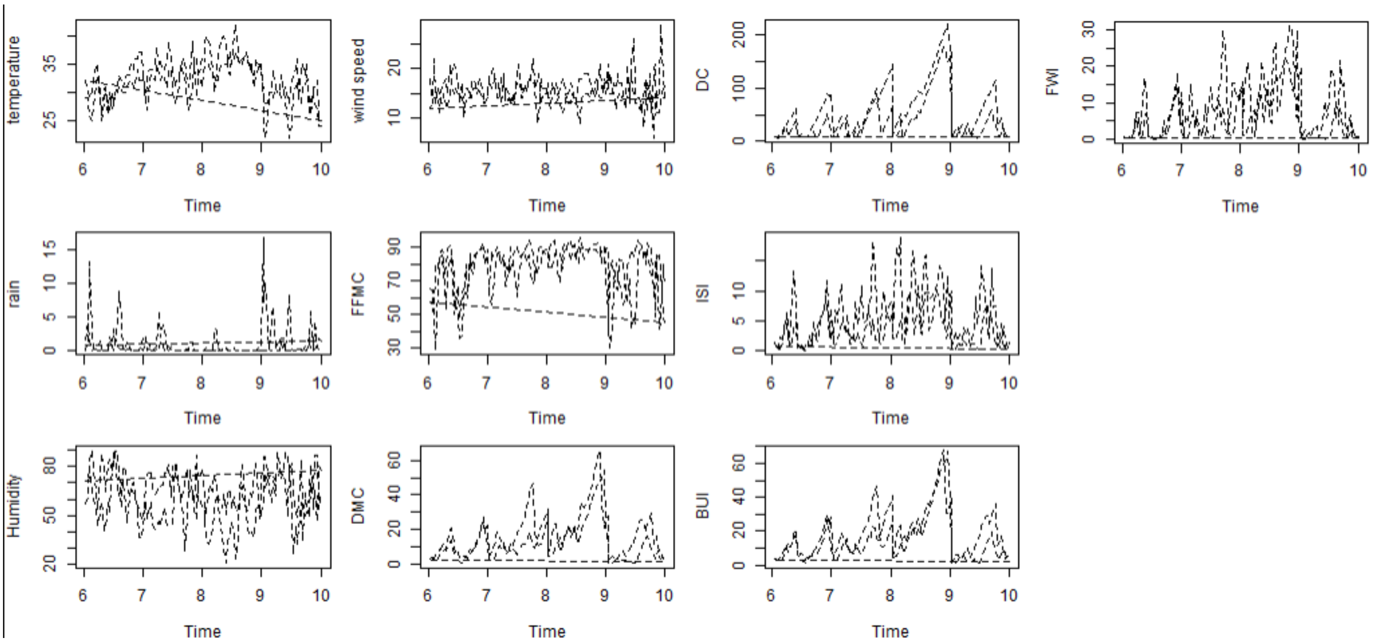


Figure 5: time series

4.6. Discussion:

From the previous visualizations, most fires happened between July and August. The temperature recorded was between 30 and 40 degrees which is very high temperature, and with lack of rain and low humidity, fire incidence such as Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), Build-up Index (BUI) and Fire Weather Index (FWI) increased. Consequently, more than 90 fire happened in this period, and it seems that the speed of the winds helped at speeding the fire.

5. Conclusion:

By employing some of the techniques covered in this module to explore the Algerian forest fires dataset, I can conclude that the exploration of data revealed seasonal trends in fire risk and relationship between attributes and fire risk.

References:

Abid, Faroudja. (2019). Algerian Forest Fires Dataset. UCI Machine Learning Repository. https://doi.org/10.24432/C5KW4N.

Example code:

```r
# libraries
library(tidyverse)
library(ggplot2)
library(dplyr)
library(GGally)
library(gridExtra)
lines= readLines("Algerian_forest_fires_dataset_UPDATE.csv")#read as lines
# cleaning
lines= lines[-c(1,125,126,127)] # remove unnecessary lines
data=read.csv(text=lines)
data$region=character(244) # new column region
data$region[1:122]="Bejaia";data$region[123:244]="Sidi-Bel Abbes"
str(data);summary(data)
# Convert variables to appropriate data types
data$DC= as.numeric(data$DC)
data$FWI= as.numeric(data$FWI)
data$Classes= as.factor(data$Classes)
data$region= as.factor(data$region)
summary(data)
# Check for missing values
na_count = colSums(is.na(data))
print(na_count) # 1 in DC 1 in FWI
# Remove rows with NA values in any column
data= na.omit(data)
# Distribution of attributes
# Loop through each attribute and create individual plots
par(mfrow=c(4, 3),mar=c(3.8, 2, 1, 1) ,oma=c(0, 0, 2, 0))
for (col in names(data[,-c(1,2,3,15)])) {
    # Check if the column contains numeric data
    if (is.numeric(data[[col]])) {
      # For numeric columns, create a histogram
      hist(data[[col]], main = paste("Distribution of", col), xlab = col, col = "lightblue", border = "black")
    } else {
      # For categorical columns, create a bar plot
      barplot(table(data[[col]]), main = paste("Distribution of", col), xlab = col, col = "lightblue", border = "black")}}
numeric_vars <- c('Temperature', 'RH', 'Ws', 'Rain', 'FFMC', 'DMC', 'DC', 'ISI', 'BUI', 'FWI')
# Loop through each numeric variable and create scatter plots
for (var in numeric_vars) {
  print(ggplot(data, aes(x = factor(month), y = !!sym(var))) +
    geom_point(aes(color = Classes,shape=region)) +
     labs(title = paste("Relationship Between Month and", var),x = "Month", y = var) +theme_minimal() +theme(legend.position = "top")) }
# Month vs classes
ggplot(data,aes(x=factor(month),fill=Classes))+geom_bar(position='dodge')+
  labs(title='Month vs Classes',x= 'Month', y = 'Count') + theme_minimal()
# pairs plot
ggpairs(data,columns = attributes,title = "Pairs Plot")
data$time= data$month + data$day/30
#time series plot
par(mfcol=c(3,3),mar=c(4,4,1,1))
  for (var in numeric_vars) {
    plot(data$time, data[[var]] ,type="l",lty=2, xlab="Time",ylab=var)}
```