



Collaborative Project Report

Semester	B.E. Semester VII – Computer Engineering
Subject	Big Data Analytics
Subject Professor In-charge	Prof. Pankaj Vanvari
Assisting Teachers	Dr. Umesh Kulkarni

Roll Number 	Name of Students 
21102A0003	Omkar Patil
21102A0005	Pranav Redij
21102A0006	Sahil Pokharkar
21102A0014	Deep Salunkhe

Name of the Project:

Data Visualization in R to get inference (Data set: Forest Fire)

Project Details:

1. Dataset Source

- The dataset originates from the UCI Machine Learning Repository and focuses on the Montesinho natural park, located in the Trás-os-Montes region of Portugal. It is widely used in machine learning studies for regression and classification problems.

2. Dataset Description

- The dataset records forest fire occurrences and related environmental factors. It aims to predict the burned area of forest fires based on meteorological data and seasonal information.

3. Attributes and Features

- The dataset consists of 517 instances (rows), each representing a forest fire event.
- There are 13 attributes (columns) in total, which include both categorical and numerical features:

Attribute Type	Description	
X	Integer	X-axis spatial coordinate within the Montesinho park map (1 to 9).
Y	Integer	Y-axis spatial coordinate within the Montesinho park map (2 to 9).
month	Categorical	Month of the year when the fire occurred (jan to dec).
day	Categorical	Day of the week when the fire occurred (mon to sun).
FFMC	Numeric	Fine Fuel Moisture Code (FFMC) index from the Canadian Forest Fire Weather Index (18.7 to 96.20).
DMC	Numeric	Duff Moisture Code (DMC) index from the Canadian Forest Fire Weather Index (1.1 to 291.3).
DC	Numeric	Drought Code (DC) index from the Canadian Forest Fire Weather Index (7.9 to 860.6).
ISI	Numeric	Initial Spread Index (ISI) from the Canadian Forest Fire Weather Index (0.0 to 56.10).
temp	Numeric	Temperature in degrees Celsius (2.2 to 33.30).
RH	Numeric	Relative humidity in percentage (15 to 100).
wind	Numeric	Wind speed in km/h (0.40 to 9.40).
rain	Numeric	Rainfall in mm/m ² (0.0 to 6.4).
area	Numeric	The burned area of the forest in hectares (ha) (0.00 to 1090.84). Note: Most values are 0.00 (no significant burned area).

4. Target Variable

- The area attribute represents the burned area (in hectares) and serves as the target variable for predictive modeling. Since many fires resulted in no significant burned area, a large portion of the data contains zero values for this attribute, making it a challenging regression problem.

5. Data Characteristics

- **Seasonality and Temporal Patterns:** The month and day features capture seasonal and weekly variations in fire occurrences.
- **Meteorological Factors:** Features such as temp (temperature), RH (relative humidity), wind (wind speed), and rain (rainfall) are important for understanding fire behavior and spread.
- **Fire Weather Indices:** The dataset includes Canadian Forest Fire Weather Index components (FFMC, DMC, DC, ISI), which are calculated from meteorological data and help assess fire potential.

6. Challenges with the Dataset

- **Imbalanced Target Variable:** Most entries have a burned area of zero, leading to an imbalanced dataset that poses challenges for modeling.
- **Outliers:** The presence of a few very large burned areas can skew the distribution and affect the performance of machine learning models.
- **Correlation Among Features:** Some features may exhibit strong correlations, such as temperature and wind speed, which could impact the predictive power of individual variables.

7. Potential Uses of the Dataset

- **Predictive Modeling:** The dataset can be used to develop regression models to predict the burned area based on environmental factors.
- **Clustering and Classification:** It can be employed for classifying the severity of fires or clustering similar fire events.
- **Environmental and Fire Management Studies:** Analysis of this data can provide insights into the factors driving forest fires and help develop strategies for prevention and control.

Inference:

1. Scatter Plot: Area Burned vs. Temperature

- **Inference:** The scatter plot indicates whether there's a relationship between temperature and the area burned. If there is a positive trend, it suggests that higher temperatures might correlate with larger burned areas. If the points are widely scattered without a clear pattern, it suggests temperature alone isn't a strong predictor of the burned area.

2. Box Plot: Area Burned per Month

- **Inference:** The box plot shows the distribution of the area burned across different months. If some months have significantly higher median burned areas or larger interquartile ranges, it suggests that certain months are more prone to severe fires. This could be due to seasonal factors, such as dry weather or wind conditions.

3. Correlation Heatmap

- **Inference:** The heatmap displays correlations between numerical variables, with strong correlations (either positive or negative) potentially highlighting key factors influencing the burned area. For example, a high correlation between temperature and wind could indicate that hotter days are often windier, possibly contributing to more severe fires.

4. Bar Plot: Fire Occurrences by Day

- **Inference:** The bar plot shows how often fires occur on each day of the week. If there is a noticeable difference in fire occurrences across days, it could be related to human activities (like weekend barbecues or work-related accidents).

5. Histogram: Distribution of Burned Area

- **Inference:** If the histogram shows a right-skewed distribution, it indicates that most fires affect small areas, while a few severe fires result in very large burned areas. This kind of distribution is typical for forest fire data, where smaller fires are more frequent than larger ones.

6. Line Plot: Average Temperature per Month

- **Inference:** This plot can help identify temperature trends over the year, such as the warmest months. A correlation with the area burned might be observed if months with higher average temperatures also show a greater extent of fires.

7. Density Plot: Temperature Distribution

- **Inference:** The density plot provides a smooth estimate of temperature distribution. Peaks in the plot could indicate the most common temperature ranges when fires occur, possibly suggesting conditions under which fires are more likely.

8. Scatter Plot: Wind Speed vs. Area Burned

- **Inference:** This plot can reveal if stronger winds are associated with larger fires. If a trend is visible, it suggests that wind might play a significant role in spreading fires.

GitHub Repository Link (Public):

https://github.com/deepsalunkhee/Learn-and-Practice/tree/master/SEM-7/BDA/R_Project

Output Screenshots:

```

#load the data

data <- read.csv("forestfires.csv", header = TRUE, sep = ",")
head(data)

  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 90.6 35.4 669.1 6.7 18.0 33 0.9 0.0 0
3 7 4 oct  sat 90.6 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

# Display the structure of the dataset
str(data)

'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: chr  "mar" "oct" "oct" "mar" ...
 $ day   : chr  "fri" "tue" "sat" "fri" ...
 $ FFMC  : num  86.2 90.6 90.6 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 ...

# Load necessary libraries
library(ggplot2)
library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

  filter, lag

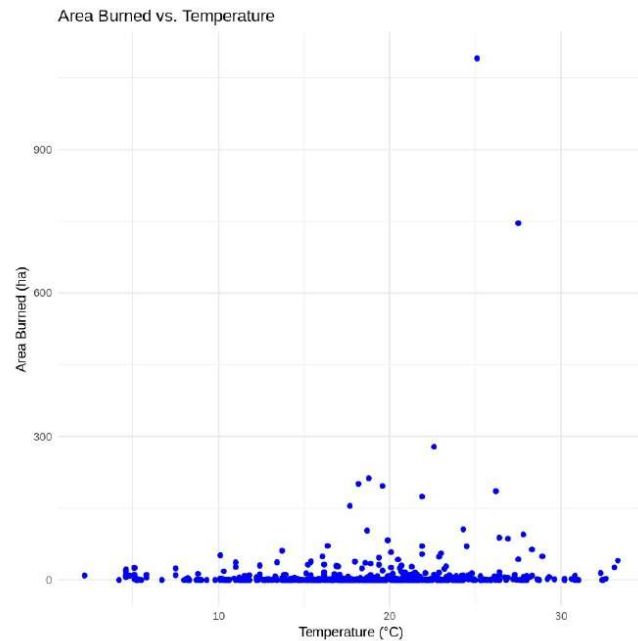
The following objects are masked from 'package:base':

  intersect, setdiff, setequal, union

# Scatter plot of area burned vs. temperature
ggplot(data, aes(x = temp, y = area)) +

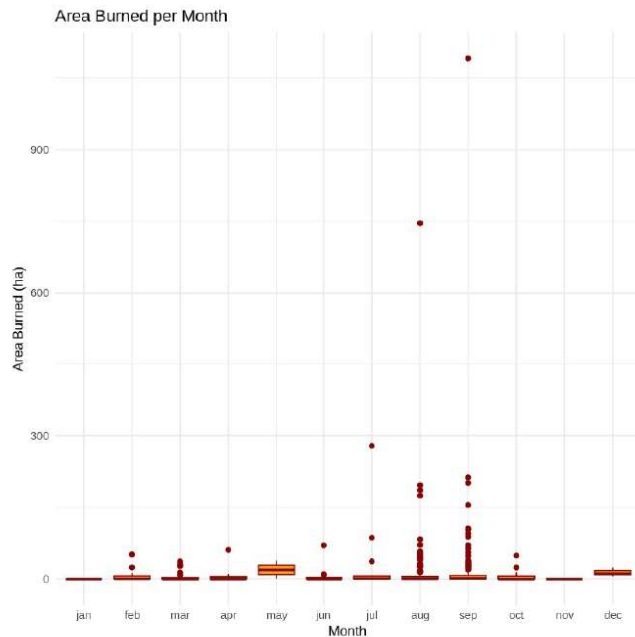
```

```
geom_point(color = "blue") +
labs(title = "Area Burned vs. Temperature", x = "Temperature (°C)",
y = "Area Burned (ha)") +
theme_minimal()
```



```
# Convert month to a factor for better visualization
data$month <- factor(data$month, levels = c("jan", "feb", "mar",
"apr", "may", "jun", "jul", "aug", "sep", "oct", "nov", "dec"))

# Box plot of area burned per month
ggplot(data, aes(x = month, y = area)) +
  geom_boxplot(fill = "orange", color = "darkred") +
  labs(title = "Area Burned per Month", x = "Month", y = "Area Burned
(ha)") +
  theme_minimal()
```



```
# Install the reshape2 package if not already installed
if (!require(reshape2)) {
  install.packages("reshape2")
}

# Load the library
library(reshape2)

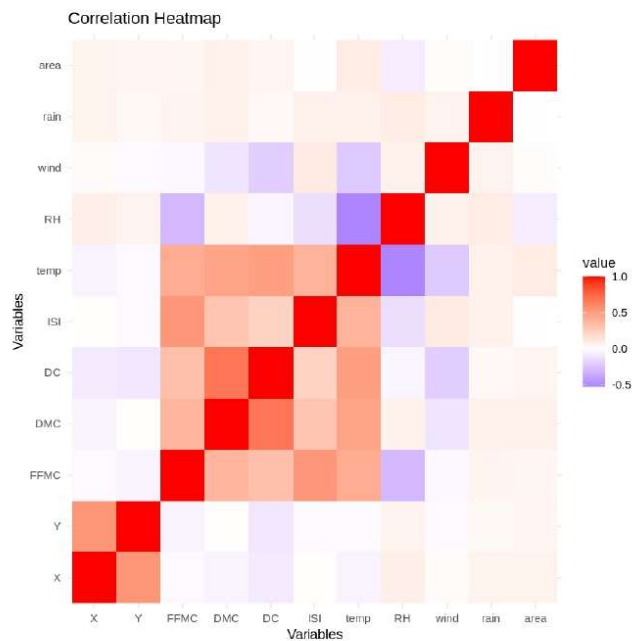
# Proceed with the correlation heatmap code
# Calculate the correlation matrix for numeric columns
numeric_data <- data %>% select_if(is.numeric)
correlation_matrix <- cor(numeric_data, use = "complete.obs")

# Heatmap of the correlation matrix
melted_correlation <- melt(correlation_matrix)

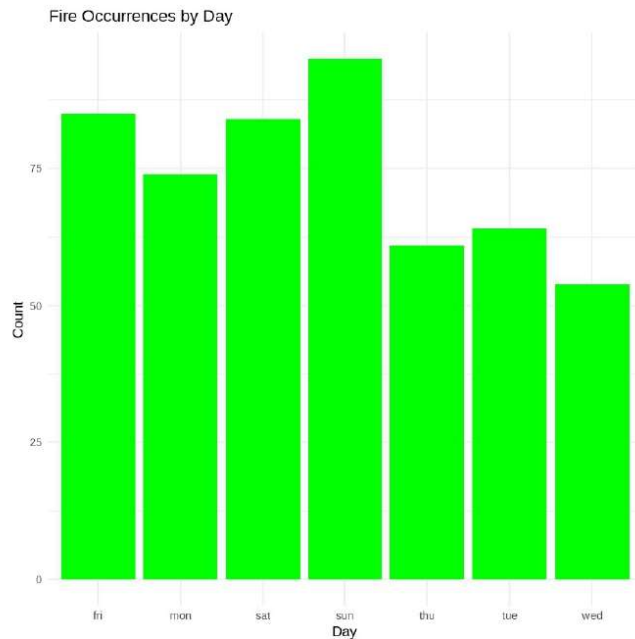
ggplot(melted_correlation, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
    midpoint = 0) +
  labs(title = "Correlation Heatmap", x = "Variables", y =
    "Variables") +
  theme_minimal()
```

```
Loading required package: reshape2
```

```
Warning message in library(package, lib.loc = lib.loc, character.only  
= TRUE, logical.return = TRUE, :  
"there is no package called 'reshape2'"  
Installing package into '/usr/local/lib/R/site-library'  
(as 'lib' is unspecified)  
also installing the dependency 'plyr'
```

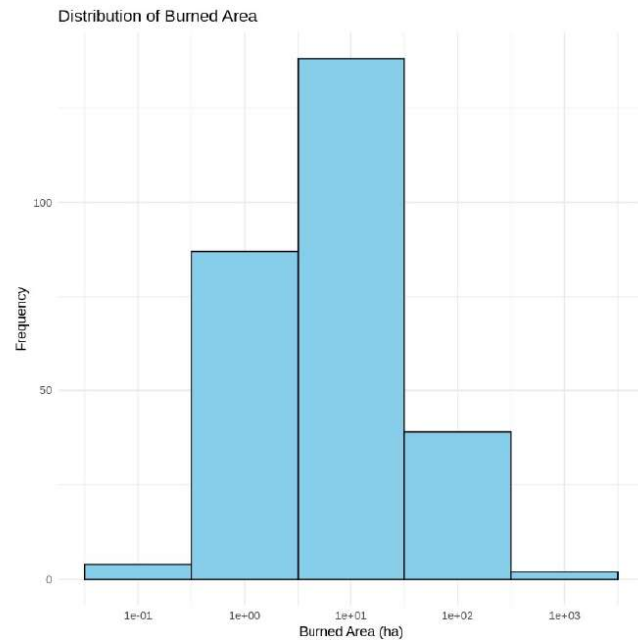


```
# Bar plot of fire occurrences by day of the week  
ggplot(data, aes(x = factor(day))) +  
  geom_bar(fill = "green") +  
  labs(title = "Fire Occurrences by Day", x = "Day", y = "Count") +  
  theme_minimal()
```

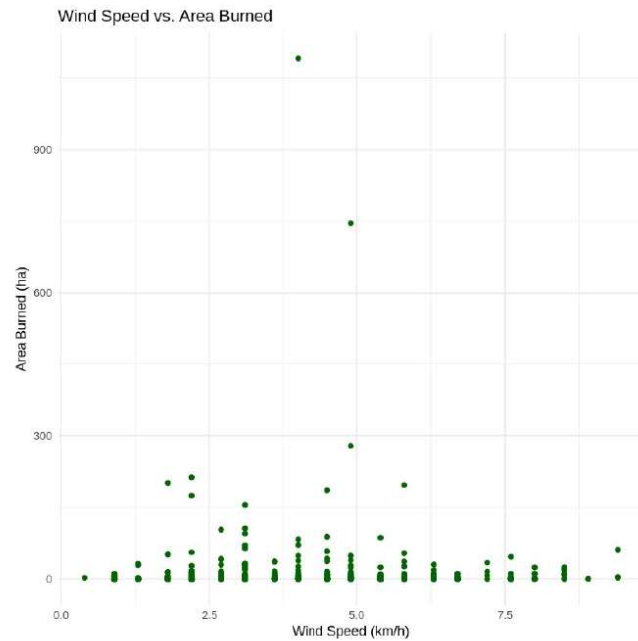



```
# Histogram of the burned area
ggplot(data, aes(x = area)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Burned Area", x = "Burned Area (ha)",
y = "Frequency") +
  theme_minimal() +
  scale_x_log10() # Use log scale to better visualize skewed data
```

Warning message in scale_x_log10():
 "log-10 transformation introduced infinite values."
 Warning message:
 "Removed 247 rows containing non-finite outside the scale range
 (`stat_bin()`)."

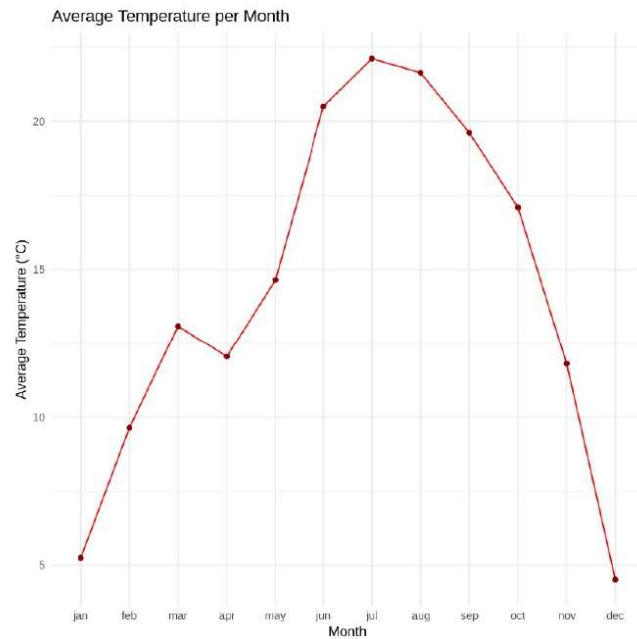


```
# Scatter plot of wind speed vs. area burned  
ggplot(data, aes(x = wind, y = area)) +  
  geom_point(color = "darkgreen") +  
  labs(title = "Wind Speed vs. Area Burned", x = "Wind Speed (km/h)",  
        y = "Area Burned (ha)") +  
  theme_minimal()
```

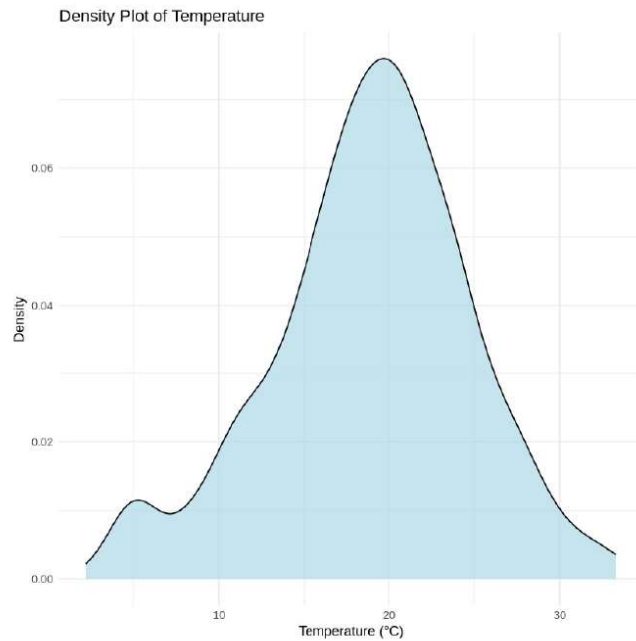


```
# Calculate the average temperature per month
avg_temp_per_month <- data %>%
  group_by(month) %>%
  summarize(avg_temp = mean(temp))

# Line plot of average temperature per month
ggplot(avg_temp_per_month, aes(x = month, y = avg_temp, group = 1)) +
  geom_line(color = "red") +
  geom_point(color = "darkred") +
  labs(title = "Average Temperature per Month", x = "Month", y =
"Average Temperature (°C)") +
  theme_minimal()
```



```
# Density plot of temperature distribution
ggplot(data, aes(x = temp)) +
  geom_density(fill = "lightblue", alpha = 0.7) +
  labs(title = "Density Plot of Temperature", x = "Temperature (°C)",
    y = "Density") +
  theme_minimal()
```



```
# Install and load GGally if not already installed
if (!require(GGally)) {
  install.packages("GGally")
}

library(GGally)

# Pair plot for numeric columns
numeric_data <- data %>% select_if(is.numeric)
ggpairs(numeric_data)

Loading required package: GGally

Warning message in library(package, lib.loc = lib.loc, character.only
= TRUE, logical.return = TRUE, :
"there is no package called 'GGally'"
Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)

also installing the dependencies 'patchwork', 'ggstats'

Registered S3 method overwritten by 'GGally':
  method from
```

`+ .gg` `ggplot2`

