William Shin

Professor Pham

Math 17 | Data Analysis and Statistical Programming using R

December 15, 2024

Investigating what drives the most Wildfires in California Using Fire Weather Index(FWI) Data

This project focuses on identifying the most significant factors contributing to wildfire occurrences and spread in California. Wildfires are a persistent challenge in the state, causing widespread environmental and economic damage. By examining the Fire Weather Index (FWI) data, which is a numeric rating that estimates the likelihood of a fire starting and spreading in forested areas, I aim to pinpoint the indices most strongly associated with wildfire severity. As the most widely used fire weather system in the world and developed in Canada, the FWI is based on three moisture codes and three fire behavior indices that account for the effects of wind and fuel moisture on fire spread and behavior. Additionally, I examine seasonal trends to identify periods of heightened activity, helping inform prevention and resource allocation strategies.

Methodology

The analysis began by exploring the seasonal patterns in wildfire occurrences and the area burned, using data from the Forest Fire dataset. To identify trends in wildfire activity over time, I utilized the group_by() function in R to summarize the total number of occurrences and

the total area burned for each month. This step helped in visualizing periods of heightened wildfire activity, specifically in August and September. To visualize these seasonal patterns, I created bar plots using the **ggplot2** package, showing the total occurrences of wildfires and the area burned in each month. The bar plots were customized with the **geom_col()** function, and the axes were adjusted using **coord_flip()** to make the data more accessible.

The core of the analysis focused on the percentage increase in key weather-related variables during the peak months of August and September. The variables of interest, which I will discuss them again later, included Drought Code (DC), Duff Moisture Code (DMC), Fine Fuel Moisture Code (FFMC), Initial Spread Index (ISI), temperature, relative humidity (RH), and wind speed. To calculate the percentage increase for each variable, I computed the mean of the variable for the months of August and September, then compared it with the overall mean for the entire dataset. The formula used for the calculation was:

$$\text{Percentage Increase} = \left(\frac{\text{Mean of August/September} - \text{Overall Mean}}{\text{Overall Mean}} \right) \times 100$$

And this was done separately for each variable, resulting in percentage increases for both August and September. These calculations were performed using R functions, and the results were visualized using bar charts to highlight which variables exhibited the most significant changes. To further refine the analysis, I computed the average percentage increase by combining the data from both August and September and compared it to the overall mean for each variable. This provided conclusive analysis into how the combination of these two months compared to the overall trends in the dataset. The **ggplot2** package was again used to visualize the final results, with bar charts presenting the percentage increase for each weather-related variable.

Dataset

The dataset, titled "Forest Fires," was compiled by Paulo Cortez and Aníbal Morais (University of Minho) in 2007. It is publicly available for research and includes detailed meteorological and fire-related data. This dataset was originally used to apply data mining techniques to predict forest fires and is detailed in the paper by Cortez and Morais titled *A Data Mining Approach to Predict Forest Fires using Meteorological Data*, presented at the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence. The dataset contains 517 records and 13 columns, with 12 input variables and one output variable.

The variables include spatial coordinates (X and Y) that represent locations within the Montesinho park map, temporal data such as the month of the year and the day of the week, and weather data including temperature in Celsius, relative humidity in percentage, wind speed in kilometers per hour, and rainfall in millimeters per square meter. The Fire Weather Index (FWI) codes, including Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), Drought Code (DC), and Initial Spread Index (ISI), quantify fire risk under various meteorological conditions. The output variable, burned area, measures the forest area burned in hectares. This variable exhibits a highly skewed distribution with many small fires and a few large ones. Most variables in the dataset are numerical, except for categorical variables like month and day. The dataset contains no missing values, simplifying the preprocessing stage. This dataset poses analytical challenges due to its skewed distribution and potential correlations among variables. Through statistical and computational techniques, my analysis seeks to uncover key drivers of wildfires, providing new mitigation and preparedness strategies.

1. Size and Scope:

- Number of Instances: 517 records
- Number of Attributes: 13 columns (12 input variables and 1 output variable)

2. Variables:

- Spatial coordinates (X, Y): Represent locations within the Montesinho park map.
- Temporal data (month, day): Indicate the month of the year and the day of the week.
- Weather data: Includes temp (temperature in Celsius), RH (relative humidity in %), wind (wind speed in km/h), and rain (rainfall in mm/m²).

Fire Weather Index codes:

- FFMC (Fine Fuel Moisture Code)
- DMC (Duff Moisture Code)
- DC (Drought Code)
- ISI (Initial Spread Index)
- RH (Relative Humidity)- These indices quantify fire risk under various meteorological conditions.
- Burned area (area): The output variable representing the forest area burned (in hectares). This variable exhibits a highly skewed distribution with many small fires and a few large ones.

4. Type of Data:

- Most variables are numerical, except for categorical variables (month, day).
- The dataset contains no missing values, which facilitates preprocessing.

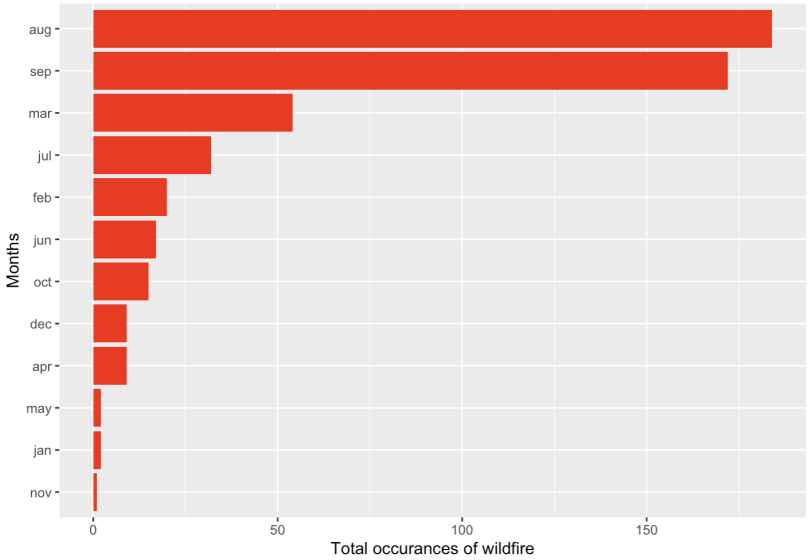
Result

1. The summary of each variables of the dataset:

```
       X               Y            month              day                FFMC       
 Min.   :1.000   Min.   :2.000   Length:268         Length:268         Min.   :63.50  
 1st Qu.:2.750   1st Qu.:4.000   Class :character   Class :character   1st Qu.:90.30  
 Median :5.000   Median :4.000   Mode  :character   Mode  :character   Median :91.65  
 Mean   :4.791   Mean   :4.358                                         Mean   :91.01  
 3rd Qu.:7.000   3rd Qu.:5.000                                         3rd Qu.:92.92  
 Max.   :9.000   Max.   :9.000                                         Max.   :96.20  
      DMC               DC              ISI              temp             RH       
 Min.   :  3.20   Min.   : 15.3   Min.   : 0.800   Min.   : 2.20   Min.   :15.00  
 1st Qu.: 82.62   1st Qu.:480.8   1st Qu.: 6.800   1st Qu.:16.10   1st Qu.:33.00  
 Median :111.70   Median :665.5   Median : 8.400   Median :20.10   Median :41.00  
 Mean   :114.28   Mean   :570.0   Mean   : 9.162   Mean   :19.26   Mean   :43.86  
 3rd Qu.:141.30   3rd Qu.:721.4   3rd Qu.:11.325   3rd Qu.:23.32   3rd Qu.:53.00  
 Max.   :291.30   Max.   :860.6   Max.   :22.700   Max.   :33.30   Max.   :96.00  
     wind            rain              area        
 Min.   :0.40   Min.   :0.0000   Min.   :  0.090  
 1st Qu.:2.70   1st Qu.:0.0000   1st Qu.:  2.138  
 Median :4.00   Median :0.0000   Median :  6.330  
 Mean   :4.11   Mean   :0.0291   Mean   : 17.929  
 3rd Qu.:4.90   3rd Qu.:0.0000   3rd Qu.: 14.845  
 Max.   :9.40   Max.   :6.4000   Max.   :278.530  
```

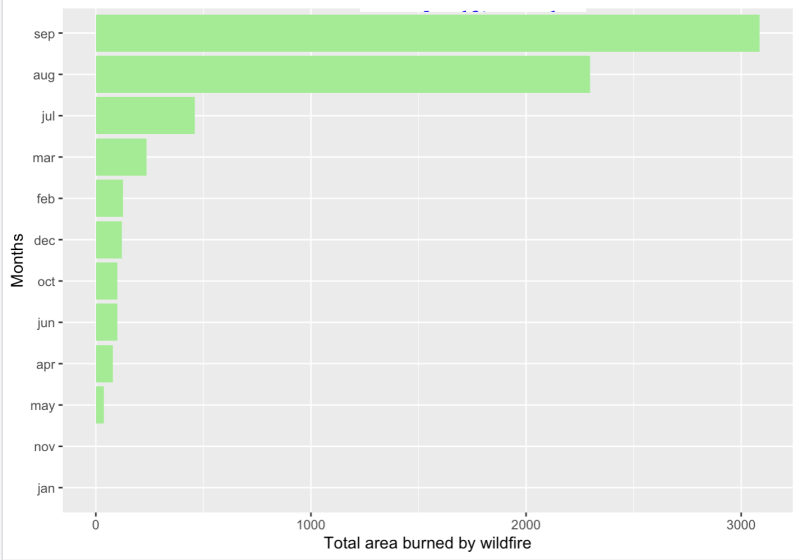
2. Using group_by() function to examine in which month(s) the wildfires have occurred the most and its graph:

```
# A tibble: 12 × 2
   month Count
   <chr> <int>
 1 aug     184
 2 sep     172
 3 mar      54
 4 jul      32
 5 feb      20
 6 jun      17
 7 oct      15
 8 apr       9
 9 dec       9
10 jan       2
11 may       2
12 nov       1
```

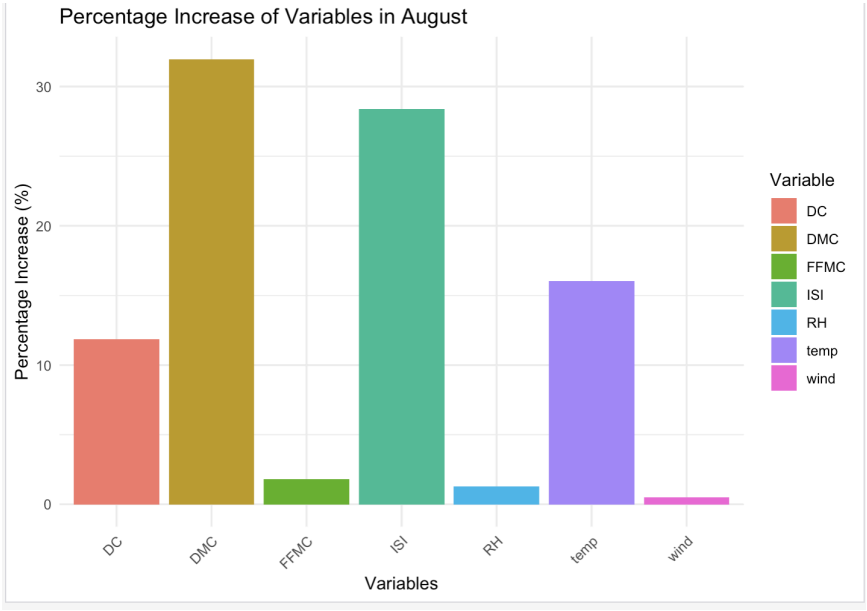


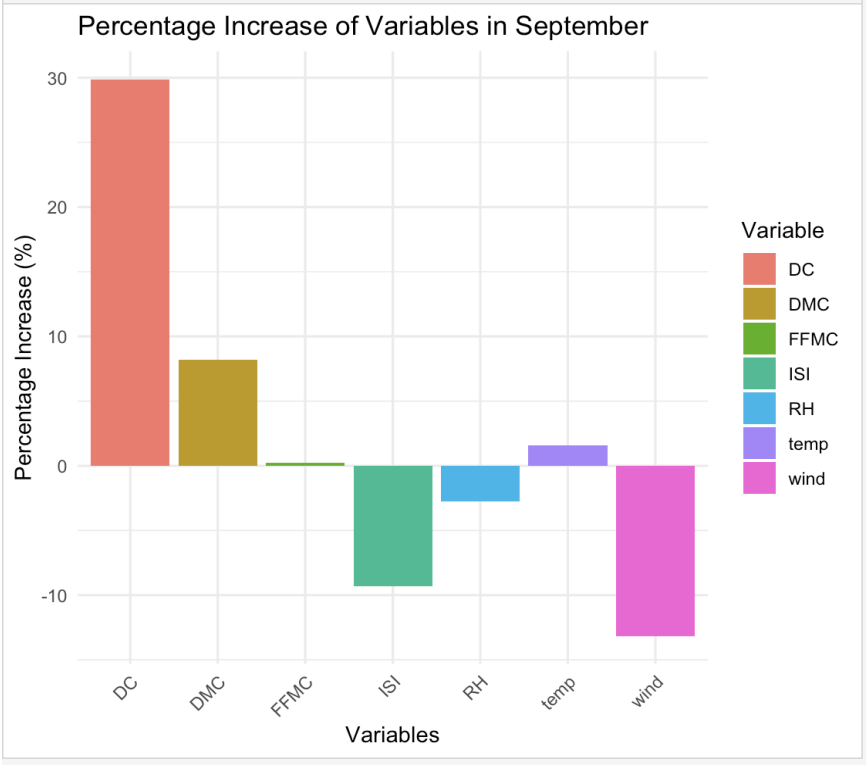
3. Using group_by() function to examine in which month(s) the wildfires have spreaded the most and its graph:

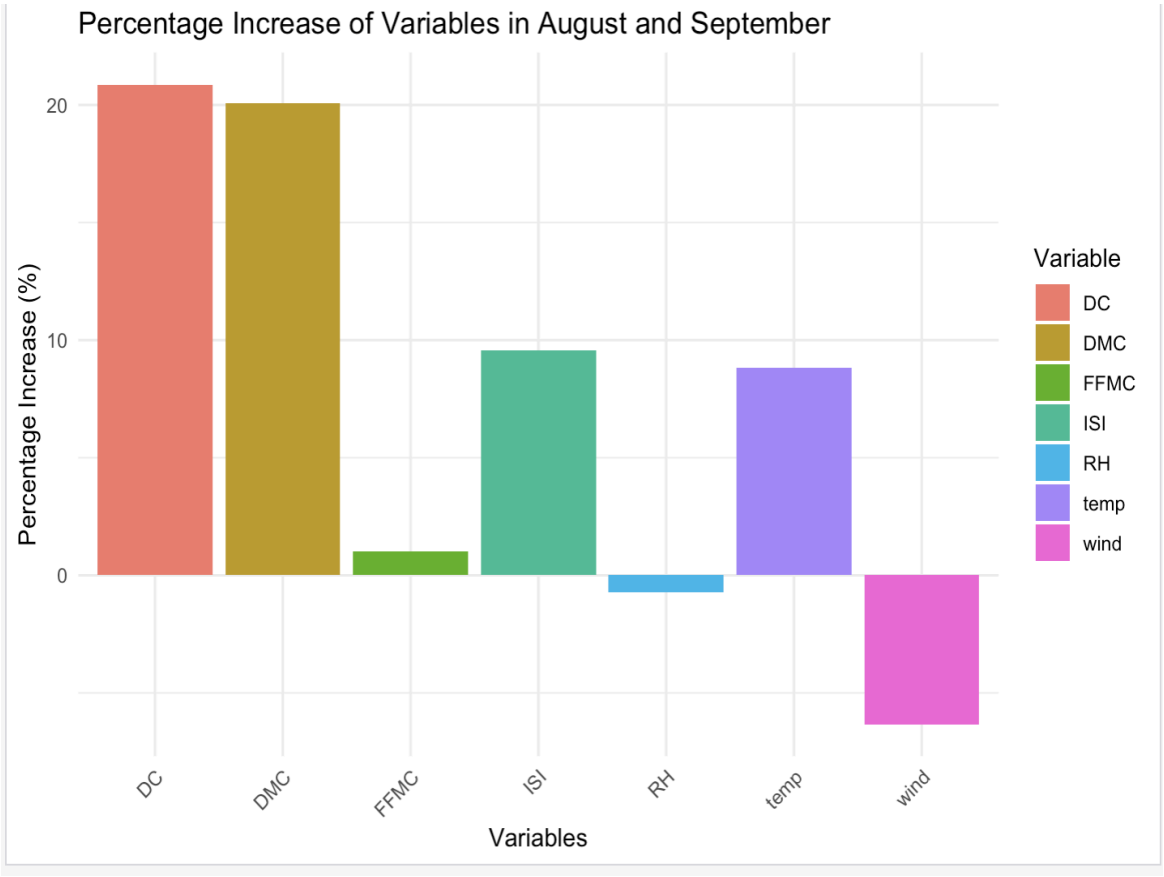
```
# A tibble: 12 × 2
   month  count
   <chr>  <dbl>
 1 sep   3086.
 2 aug   2298.
 3 jul    460.
 4 mar    235.
 5 feb    126.
 6 dec    120.
 7 oct     99.6
 8 jun     99.3
 9 apr     80.0
10 may     38.5
11 jan      0
12 nov      0
```



4. → Both graphs highlight the wildfire's occurrence and spread spikes in August and September.
5. Calculating the percentage increase of each FWI variable and examining which one(s) showed the highest spike during this period.







6. DC and DMC have the highest overall!

Conclusion

The analysis revealed that the Drought Code (DC) and Duff Moisture Code (DMC) experienced the highest percentage increases in August and September, indicating their strong potential influence on wildfire behavior during these months. These results suggest that drought conditions, as measured by (DC) and the moisture levels in the forest floor (DMC), were key factors contributing to the increased wildfire occurrence and area burned during this period. This finding highlights the importance of monitoring these variables, particularly in the months leading up to peak wildfire activity. DMC, or the Duff Moisture Code, measures the moisture content in deep organic layers of soil, while DC, the Drought Code, reflects moisture in deeper

layers of the soil, indicating prolonged drought conditions. DC, in particular, has intensified in recent years due to human activities such as land-use changes and climate change, leading to more severe droughts that significantly increase the likelihood and severity of wildfires. The significant increase in these variables could serve as an early warning system for wildfire risk, signifying the need for targeted prevention and mitigation strategies. In future research, further exploration into how these variables interact with other environmental and climate factors could provide more comprehensive insights into the increasing rate of wildfire occurrence. In conclusion, the results imply that, while predictive modeling was initially considered, a more focused analysis based on comparing mean value of specific months of heightened activity, combined the key weather-related variables such as FWI, offer more actionable and in-depth strategies for wildfire management and forecasting. The combination of August and September data proved to be a useful approach in understanding the spikes in wildfire occurrences, revealing that among FWI indexes, DMC and DC can significantly impact the spread and intensity of wildfires.

Copy of R Code

```r
f <- read.csv("forestfires.csv")
library(dplyr)
library(ggplot2)
f0 <- f |> group_by(month) |> summarize(Count=n()) |> arrange(-Count)
f0
ggplot(f0, aes(x = reorder(month, Count), y = Count)) +
  geom_col(fill = "red") +
  coord_flip() +
  labs(  x = "Months",
         y = "Total occurances of wildfire")
f1 <- f |> group_by(month) |> summarize(count=sum(area)) |> arrange(-count)
f1
ggplot(f1, aes(x = reorder(month, count), y = count)) +
  geom_col(fill = "lightgreen") +
  coord_flip() +
  labs(  x = "Months",
         y = "Total area burned by wildfire")

summary(f)
 f |> group_by(month) |> summarize(Mean=mean(FFMC)) |> arrange(-Mean)
f_august <- subset(f, month == "aug")
perc_increase_DC <- ((mean(f_august$DC) - mean(f$DC)) / mean(f$DC)) * 100
perc_increase_DMC <- ((mean(f_august$DMC) - mean(f$DMC)) / mean(f$DMC)) * 100
perc_increase_FFMC <- ((mean(f_august$FFMC) - mean(f$FFMC)) / mean(f$FFMC)) * 100
perc_increase_ISI <- ((mean(f_august$ISI) - mean(f$ISI)) / mean(f$ISI)) * 100
perc_increase_temp <- ((mean(f_august$temp) - mean(f$temp)) / mean(f$temp)) * 100
perc_increase_RH <- ((mean(f_august$RH) - mean(f$RH)) / mean(f$RH)) * 100
perc_increase_wind <- ((mean(f_august$wind) - mean(f$wind)) / mean(f$wind)) * 100

df_percentage_increase <- data.frame(
  Variable = c("DC", "DMC", "FFMC", "ISI", "temp", "RH", "wind"),
  PercentageIncrease = c(perc_increase_DC, perc_increase_DMC, perc_increase_FFMC, perc_increase_ISI,
                         perc_increase_temp, perc_increase_RH, perc_increase_wind)
)

ggplot(df_percentage_increase, aes(x = Variable, y = PercentageIncrease, fill = Variable)) +
  geom_bar(stat = "identity") +  # Use 'identity' to plot actual values
  labs(title = "Percentage Increase of Variables in August",
       x = "Variables",
       y = "Percentage Increase (%)") +
  theme_minimal() +  # Use a minimal theme
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```r
f_september <- subset(f, month == "sep")
f_september
# Calculate the percentage increase for each variable for September
perc_increase_DC1 <- ((mean(f_september$DC) - mean(f$DC)) / mean(f$DC)) * 100
perc_increase_DMC1 <- ((mean(f_september$DMC) - mean(f$DMC)) / mean(f$DMC)) * 100
perc_increase_FFMC1 <- ((mean(f_september$FFMC) - mean(f$FFMC)) / mean(f$FFMC)) * 100
perc_increase_ISI1 <- ((mean(f_september$ISI) - mean(f$ISI)) / mean(f$ISI)) * 100
perc_increase_temp1 <- ((mean(f_september$temp) - mean(f$temp)) / mean(f$temp)) * 100
perc_increase_RH1 <- ((mean(f_september$RH) - mean(f$RH)) / mean(f$RH)) * 100
perc_increase_wind1 <- ((mean(f_september$wind) - mean(f$wind)) / mean(f$wind)) * 100

# Create a new data frame with the results for September
df_percentage_increase_september <- data.frame(
  Variable = c("DC", "DMC", "FFMC", "ISI", "temp", "RH", "wind"),
  PercentageIncrease = c(perc_increase_DC1, perc_increase_DMC1, perc_increase_FFMC1, perc_increase_ISI1,
                         perc_increase_temp1, perc_increase_RH1, perc_increase_wind1)
)
df_percentage_increase_september
ggplot(df_percentage_increase_september, aes(x = Variable, y = PercentageIncrease, fill = Variable)) +
  geom_bar(stat = "identity") +  # Use 'identity' to plot actual values
  labs(title = "Percentage Increase of Variables in September",
       x = "Variables",
       y = "Percentage Increase (%)") +
  theme_minimal() +  # Use a minimal theme
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

perc_increase_DC2 <- (((mean(f_september$DC) + mean(f_august$DC)) / 2) - mean(f$DC)) / mean(f$DC) * 100
perc_increase_DMC2 <- (((mean(f_september$DMC) + mean(f_august$DMC)) / 2) - mean(f$DMC)) / mean(f$DMC) * 100
perc_increase_FFMC2 <- (((mean(f_september$FFMC) + mean(f_august$FFMC)) / 2) - mean(f$FFMC)) / mean(f$FFMC) * 100
perc_increase_ISI2 <- (((mean(f_september$ISI) + mean(f_august$ISI)) / 2) - mean(f$ISI)) / mean(f$ISI) * 100
perc_increase_temp2 <- (((mean(f_september$temp) + mean(f_august$temp)) / 2) - mean(f$temp)) / mean(f$temp) * 100
perc_increase_RH2 <- (((mean(f_september$RH) + mean(f_august$RH)) / 2) - mean(f$RH)) / mean(f$RH) * 100
perc_increase_wind2 <- (((mean(f_september$wind) + mean(f_august$wind)) / 2) - mean(f$wind)) / mean(f$wind) * 100

# Create a data frame for the percentage increases
df_percentage_increase_Augustandseptember <- data.frame(
  Variable = c("DC", "DMC", "FFMC", "ISI", "temp", "RH", "wind"),
  PercentageIncrease = c(
    perc_increase_DC2, perc_increase_DMC2, perc_increase_FFMC2, perc_increase_ISI2,
    perc_increase_temp2, perc_increase_RH2, perc_increase_wind2
  )
)

```

```r

df_percentage_increase_Augustandseptember
ggplot(df_percentage_increase_Augustandseptember, aes(x = Variable, y = PercentageIncrease, fill = Variable)) +
  geom_bar(stat = "identity") +  # Use 'identity' to plot actual values
  labs(title = "Percentage Increase of Variables in August and September",
       x = "Variables",
       y = "Percentage Increase (%)") +
  theme_minimal() +  # Use a minimal theme
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
ggplot(f, aes(x = FFMC, y = area, color = as.factor())) +
  geom_point() +
  facet_wrap(~month) +
  labs(
    title = "Scatter Plot of FFMC vs Area Burned by Month",
    x = "FFMC (Fine Fuel Moisture Code)",
    y = "Area Burned",
    color = "Month"
  ) +
  theme_minimal()
ggplot(f, aes(x = DC, y = area, color = as.factor(month))) +
  geom_point() +
  facet_wrap(~month) +
  labs(
    title = "Scatter Plot of DC vs Area Burned by Month",
    x = "DC (Fine Fuel Moisture Code)",
    y = "Area Burned",
    color = "Month"
  ) +
  theme_minimal()
ggplot(f, aes(x = DMC, y = area, color = as.factor(month))) +
  geom_point() +
  facet_wrap(~month) +
  labs(
    title = "Scatter Plot of DMC vs Area Burned by Month",
    x = "DMC (Fine Fuel Moisture Code)",
    y = "Area Burned",
    color = "Month"
  ) +
  theme_minimal()


```

Screenshot

```r
ggplot(f, aes(x = RH, y = area, color = as.factor(month))) +
  geom_point() +
  facet_wrap(~month) +
  labs(
    title = "Scatter Plot of RH vs Area Burned by Month",
    x = "RH (Fine Fuel Moisture Code)",
    y = "Area Burned",
    color = "Month"
  ) +
  theme_minimal()
ggplot(f, aes(x = ISI, y = area, color = as.factor(month))) +
  geom_point() +
  facet_wrap(~month) +
  labs(
    title = "Scatter Plot of ISI vs Area Burned by Month",
    x = "ISI (Fine Fuel Moisture Code)",
    y = "Area Burned",
    color = "Month"
  ) +
  theme_minimal()
hist(f$DMC, main="Distribution of DMC", xlab="DMC")
hist(f$area, main="Distribution of Area", xlab="Area")
```