W203 Lab 1: Forest Fire Data Exploration

Ben Arnoldy, Alexander Chung, Christina Papadimitriou, Ehsan Yousefzadeh

Package installation

Note that for the purpose of this report the installation of the following packages is necessary:

install.packages('gplots', dependencies = TRUE)

install.packages('dplyr')

install.packages('reshape')

install.packages('xcolor')

install.packages('kableExtra')

install.packages('PerformanceAnalytics')

install.packages('ggplot2')

Introduction

Research Question

This is an exploratory data analysis that aims to understand what factors lead to particularly damaging forest fires. The motivation behind the analysis is to eventually develop an early warning system for forest fires, which are a major environmental, economic, and safety concern in some locations.

Data Summary

The dataset explored in this study includes forest fire data from the Montesinho Natural Park in Portugal. It consists of 517 observations and 13 variables, including spatial location features within a 9x9 grid (X and Y axes), *month* and *day* that each fire occured, weather-related variables including temperature in degrees Celsius (*temp*), *wind* speed in km/h, percent of relative humidity (*RH*), and outside *rain* in mm/m2. Additionally, the dataset includes a set of variables from the Fire Weather Index (FWI), a Canadian system for rating forest fire danger. These include Fine Fuel Moisture Code (*FFMC*), Duff Moisture Code (*DMC*), Drought Code (*DC*), and Initial Spread Index (*ISI*). Lastly, the *area* variable indicates the forest region burned in hectares (ha)[1] as a result of each fire. The *area* variable is the outcome variable and will be used as a proxy to measure how damaging a fire is.

```r
library(knitr)
library(kableExtra)
variables = c("X", "Y", "month", "day", "FFMC", "DMC", "DC", "ISI","temp","RH","wind",
              "rain","area")
description = c("x-axis spatial coordinate","y-axis spatial coordinate",
                "month of the year", "day of the week", "index from the FWI system",
                "index from the FWI system","index from the FWI system",
                "index from the FWI system","temperature in Celsius degrees",
                "relative humidity in %","wind speed in km/h",
                "outside rain in mm/m2","the burned area of the forest (in hectares)")
```

[1] 1 hectare (ha) = 10,000 square meters = 2.47 acres

```
value_range = c("1 to 9", "2 to 9", ""jan" to "dec"", ""mon" to "sun"", "18.7 to 96.20",
                "1.1 to 291.3","7.9 to 860.6","0.0 to 56.10","2.2 to 33.30",
                "15.0 to 100","0.40 to 9.40","0.0 to 6.4","0.00 to 1090.84")
data_type = c("discrete","discrete", "categorical", "categorical", "continuous",
              "continuous","continuous","continuous","continuous","continuous",
              "continuous","continuous","continuous")

table1 = data.frame(variables, description, value_range, data_type)
kable(table1, format = "latex", booktabs = T,
      caption = "Variable Descriptions, Ranges and Data Types") %>%
  kable_styling(latex_options = c("striped","hold_position"))
```

Table 1: Variable Descriptions, Ranges and Data Types

variables	description	value_range	data_type
X	x-axis spatial coordinate	1 to 9	discrete
Y	y-axis spatial coordinate	2 to 9	discrete
month	month of the year	“jan” to “dec”	categorical
day	day of the week	“mon” to “sun”	categorical
FFMC	index from the FWI system	18.7 to 96.20	continuous
DMC	index from the FWI system	1.1 to 291.3	continuous
DC	index from the FWI system	7.9 to 860.6	continuous
ISI	index from the FWI system	0.0 to 56.10	continuous
temp	temperature in Celsius degrees	2.2 to 33.30	continuous
RH	relative humidity in %	15.0 to 100	continuous
wind	wind speed in km/h	0.40 to 9.40	continuous
rain	outside rain in mm/m2	0.0 to 6.4	continuous
area	the burned area of the forest (in hectares)	0.00 to 1090.84	continuous

Table 1 shows the variables in the dataset, their ranges and data types. The *X* and *Y* spatial coordinates are discete variables that take integer values from 1 to 9 and 2 to 9, respectively. The *month* and *day* are categorical variables. The rest of the variables (*FFMC*, *DMC*, *DC*, *ISI*, *temp*, *RH*, *wind*, *rain* and *area*) are continuous variables.

```
library(knitr)
library(kableExtra)
fires = read.csv("forestfires.csv")
kable(summary(fires), format = "latex", booktabs = T,
      caption = "Data Summary") %>%
  kable_styling(latex_options = c("striped", "hold_position", "scale_down"))
```

Table 2: Data Summary

X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
Min. :1.000	Min. :2.0	aug :184	fri:85	Min. :18.70	Min. : 1.1	Min. : 7.9	Min. : 0.000	Min. : 2.20	Min. : 15.00	Min. :0.400	Min. :0.00000	Min. : 0.00
1st Qu.:3.000	1st Qu.:4.0	sep :172	mon:74	1st Qu.:90.20	1st Qu.: 68.6	1st Qu.:437.7	1st Qu.: 6.500	1st Qu.:15.50	1st Qu.: 33.00	1st Qu.:2.700	1st Qu.:0.00000	1st Qu.: 0.00
Median :4.000	Median :4.0	mar : 54	sat:84	Median :91.60	Median :108.3	Median :664.2	Median : 8.400	Median :19.30	Median : 42.00	Median :4.000	Median :0.00000	Median : 0.52
Mean :4.669	Mean :4.3	jul : 32	sun:95	Mean :90.64	Mean :110.9	Mean :547.9	Mean : 9.022	Mean :18.89	Mean : 44.29	Mean :4.018	Mean :0.02166	Mean : 12.85
3rd Qu.:7.000	3rd Qu.:5.0	feb : 20	thu:61	3rd Qu.:92.90	3rd Qu.:142.4	3rd Qu.:713.9	3rd Qu.:10.800	3rd Qu.:22.80	3rd Qu.: 53.00	3rd Qu.:4.900	3rd Qu.:0.00000	3rd Qu.: 6.57
Max. :9.000	Max. :9.0	jun : 17	tue:64	Max. :96.20	Max. :291.3	Max. :860.6	Max. :56.100	Max. :33.30	Max. :100.00	Max. :9.400	Max. :6.40000	Max. :1090.84
NA	NA	(Other): 38	wed:54	NA	NA	NA	NA	NA	NA	NA	NA	NA

Table 2 displays the summary statistics of all the variables within the dataset. When evaluating the data quality there were not any major issues to address. The dataset was complete, with no missing values across all variables. One important observation is that the outcome variable (*area*) has zero values for 47.8% of the observations (247 out of 517). Since the purpose of this report is to identify factors that lead to particularly damaging forest fires, where damage is proxied using the burned *area*, we choose to perform part of our analysis by grouping the observations into three categories (Zero Damage, Small Damage, and Large Damage). This categorization will be explained further in the next section of the report.

```
nrow(fires) # number of total observations
```

```
## [1] 517
```

```
nrow(fires[fires$area == 0,]) # number of observations with zero area value
```

```
## [1] 247
```

In terms of data preparation, one important step was to re-order the levels of the categorical (factor) variables, *month* and *day*. This ensures that the months and days are ordered correctly when displayed in graphs and tables.

```
fires$month = factor(fires$month, levels=c("jan", "feb", "mar", "apr", "may", "jun",
                                            "jul", "aug", "sep", "oct", "nov", "dec"))
fires$day = factor(fires$day, levels=c("mon", "tue", "wed", "thu", "fri", "sat", "sun"))
```

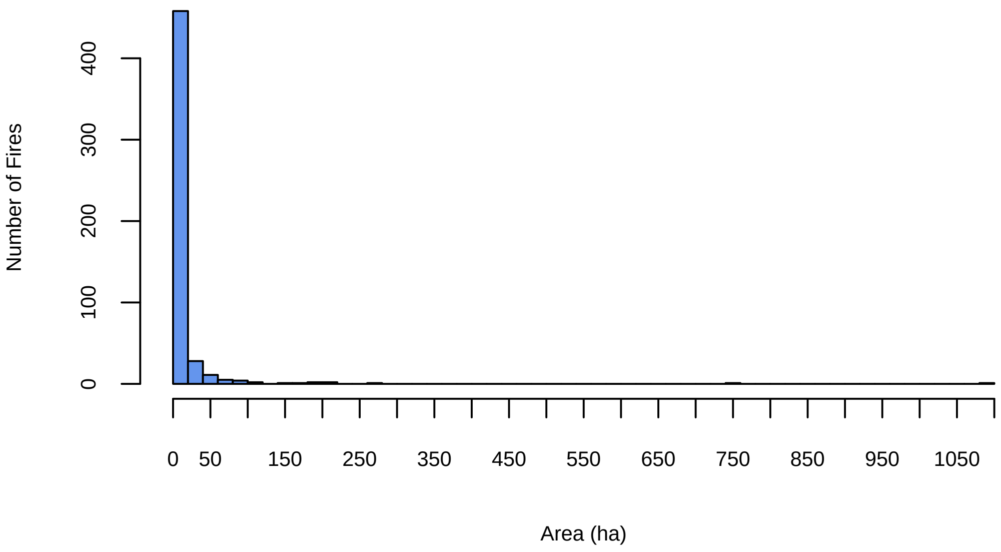
Univariate Analysis of Key Variables

We will now perform a univariate analysis of the key variables. The main goal of the univariate analysis is to gain a better understanding of the distribution of each variable across the 517 observations of fire instances.

Output Variable

```
hist(fires$area, breaks = 50, col="cornflowerblue",
     main = "Figure 1. Histogram of Burned Area (in hectares)", xlab = 'Area (ha)',
     ylab = 'Number of Fires', cex.main=0.8, cex.lab=0.7, xaxt = "n", yaxt = "n")
axis(1, at = seq(0, 1100, by = 50), cex.axis = 0.7)
axis(2, cex.axis = 0.7)
```

Figure 1. Histogram of Burned Area (in hectares)



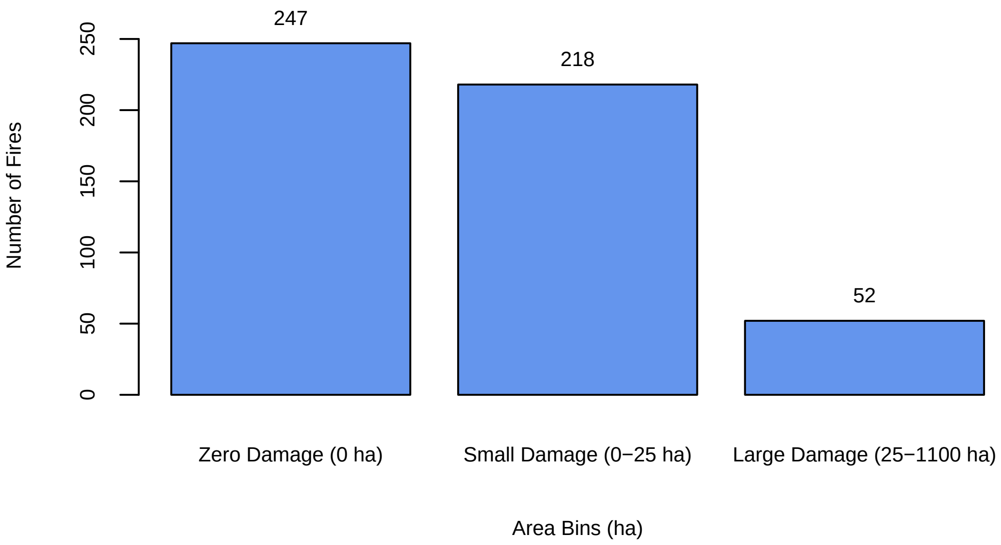
We start our univariate analysis by investigating the distribution of the burned *area* variable which measures the damage caused by each fire. The histogram of the *area* variable (Figure 1) has a strong positive skew, showing that most observations are concentrated to the left on the x-axis. From Table 2, we note that the mean (12.85 ha) is much larger than the median (0.52 ha), which is typical for very positively skewed variables. As discussed in the introduction section, this positive skewness is partially attributed to the fact that almost half of the observations (247 out of 517) have zero burned area.

```r
# creating area_bins variable of fires data.frame
fires$area_bins = cut(fires$area, breaks = c(0,0.09,25, Inf),
                      right = FALSE, labels = c("Zero Damage (0 ha)",
                      "Small Damage (0-25 ha)", "Large Damage (25-1100 ha)"))

# creating subset data.frames based on the area bins
fires.zero = subset(fires, area == 0)
fires.small = subset(fires, area > 0 & area <25)
fires.mega = subset(fires, area >= 25)

bp = barplot(table(fires$area_bins),
             main="Figure 2. Distribution of Fires across Area Bins",
             ylim=c(0, 280), xlab = 'Area Bins (ha)', ylab = "Number of Fires",
             cex.main=0.8, cex.lab=0.7, yaxt = "n", cex.names=0.7, col="cornflowerblue")
text(x = bp, y = table(fires$area_bins), label = round(table(fires$area_bins),2),
     pos = 3, cex = 0.7)
axis(2, cex.axis = 0.7)
```

Figure 2. Distribution of Fires across Area Bins



In order to see a clearer picture of the *area* distribution we split the variable into three buckets: fires with zero burned area (Zero Damage: area = 0), fires with burned area greater than zero and smaller than 25 hectares (Small Damage: $0 <$ area < 25 ha), and fires with burned area greater than equal to 25 hectares (Large Damage: area $>=$ 25) (see Figure 2). The reasoning behind the three groupings is based on the following logic. First, fires with zero burned area are grouped together since they are fires with no damage (according to the proxy used for this project). The threshold choice to split the following two groups (Small Damage and Large Damage) was based on a tradeoff between leaving enough observations in the largest bucket and choosing a number for burned area that was big enough to classify 'Large Damage' fires. Additionally, in the scientific community, there is a wide variability of area burned in hectares cut-offs to define large fires ranging from 20 to 40,000 ha. We made a conservative choice of 25 ha as the cut-off between small and large damage fires. Therefore, our choice of 25 ha, leaves 218 observations in the 'Small Damage' group and 52 observations in the 'Large Damage' group.

A futher note here about our choice of bin break points: We explored other break points between the small and large area bins including large area >= 40, 50, 75, and 100. The correlations between area and the other variables (explained later in the report) did not grow stronger with these alternate choices, except at the 100 breakpoint. However, we would have been left with only 14 observations in an area >=100 bin. We were not comfortable violating the general rule of thumb that 30+ observations are needed to have an adequate sample size for basic statistical analysis. In fact, if we were to stick to that rule of thumb, we could not set the break point higher than area >=40.

Month and Day Variables

Next, we examine the distributions of the *month* and *day* variables (i.e. number of observations that fall under each month and each day). Figure 3, shows the number of fires that occured by month and we notice that the majority of fires took place in August and September. Figure 4, shows how fires are distributed across days of the week. Interestigly, we observe that there are more fires during the weekend (Friday-Sunday) than during the week (Monday-Thursday).

```
bp = barplot(table(fires$month), main="Figure 3. Distribution of Fires across Months",
             cex.main=0.8, cex.lab=0.7, cex.names=0.7, ylim=c(0, 210), yaxt = "n",
             xlab=" ", ylab = "Number of Fires",  col="seagreen3")
text(x = bp, y = table(fires$month), label = round(table(fires$month),2),
     pos = 3, cex = 0.7)
axis(2, cex.axis = 0.7)

bp2 = barplot(table(fires$day),
              main="Figure 4. Distribution of Fires across Days of the Week",
              cex.main=0.8, cex.lab=0.7, cex.names=0.7, ylim = c(0, 110), yaxt = "n",
              xlab=" ", ylab = "Number of Fires",  col="seagreen3")
text(x = bp2, y = table(fires$day), label = round(table(fires$day),2), pos = 3, cex = 0.7)
axis(2, cex.axis = 0.7)
```

Figure 3. Distribution of Fires across Months

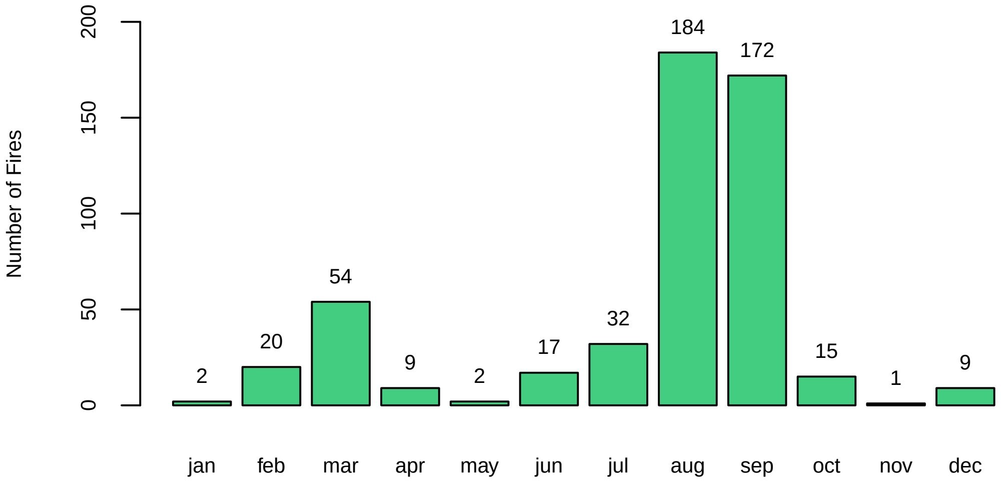
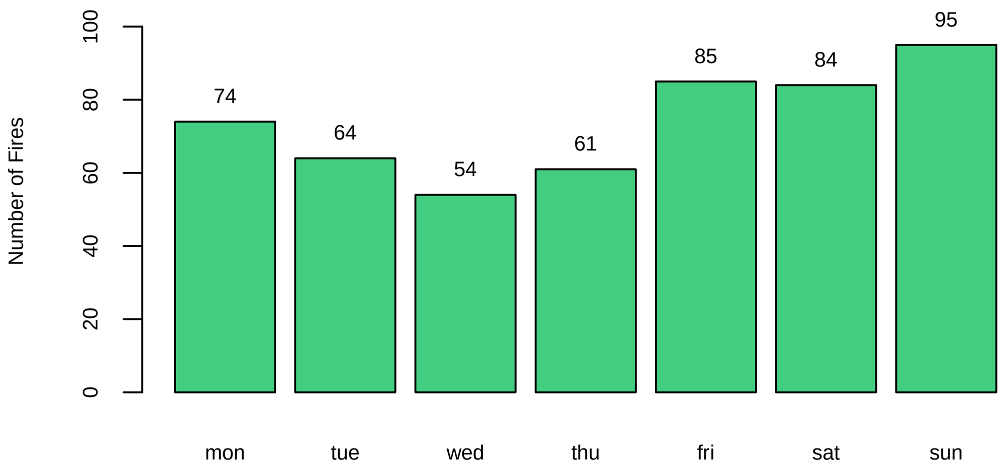


Figure 4. Distribution of Fires across Days of the Week



Weather Variables

Next, we look at the histograms of the weather variables (Figures 5-8).

After plotting the *temperature* variable (Figure 6) we can see that the curve is not quite normally distributed given a small cluster of data points around 5 degrees celsius. The mean of the temperature is 18.89 degrees celsius.

Wind speed (Figure 5) and *RH* (Figure 7) have a slightly positive skew. Most observations of the wind variable falling under 2-6 km/h with a mean of 4.018 km/h and a median of 4.000 km/h. The RH variable is positively skewed with a mean of 44.29 % and a median of 42.00 %.

Finally, with regards to the *rain* variable, all observations except for 8 register 0 mm/m2. Given that there are only 8 observations of rain instance greater than 0, we will not continue any further analysis of this variable.

```r
par(mfrow=c(2,2))
hist(fires$wind, breaks = 10, main = "Figure 5. Histogram of Wind Speed (km/h)",
     col = "lightslateblue", cex.main=0.8, cex.lab=0.7,
     yaxt = "n", xaxt = "n", xlab = "Wind Speed (km/h)")
axis(2, cex.axis = 0.7)
axis(1, cex.axis = 0.7)
hist(fires$temp, breaks = 15, main = "Figure 6. Histogram of Temperature (deg C)",
     col = "lightslateblue", cex.main=0.8, cex.lab=0.7,
     yaxt = "n", xaxt = "n", xlab = "Temperature (degrees C)")
axis(2, cex.axis = 0.7)
axis(1, cex.axis = 0.7)
hist(fires$RH, breaks = seq(0,100,by=10), main = "Figure 7. Histogram of Relative Humidity",
     yaxt = "n", xaxt = "n", cex.main=0.8, cex.lab=0.7,
     col = "lightslateblue", xlab = "Relative Humidity (%)")
axis(2, cex.axis = 0.7)
axis(1, cex.axis = 0.7)
hist(fires$rain, breaks = 20, main = "Figure 8. Histogram of Rain (mm/m2)",
     col = "lightslateblue", cex.main=0.8, cex.lab=0.7,
     yaxt = "n", xaxt = "n", xlab = "Rain (mm/m^2)")
axis(2, cex.axis = 0.7)
axis(1, cex.axis = 0.7)
```

Figure 5. Histogram of Wind Speed (km/h)

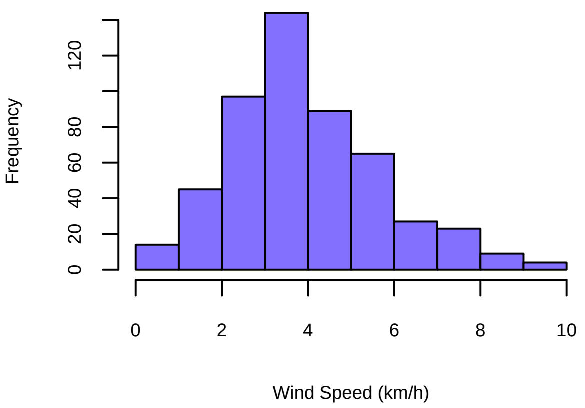


Figure 6. Histogram of Temperature (deg C)

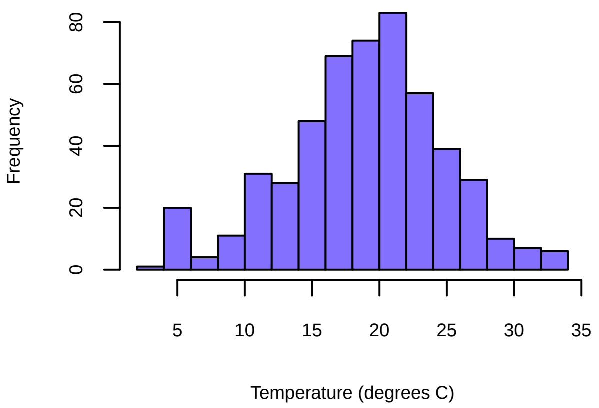


Figure 7. Histogram of Relative Humidity

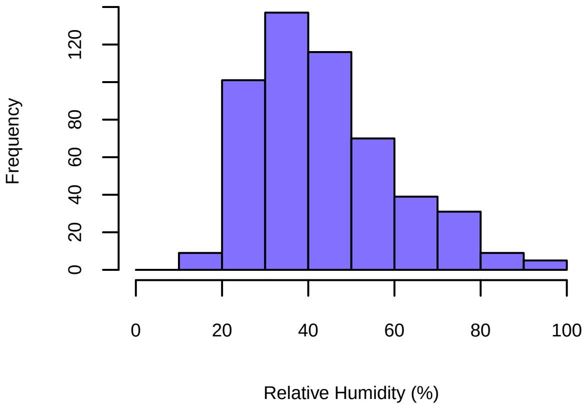
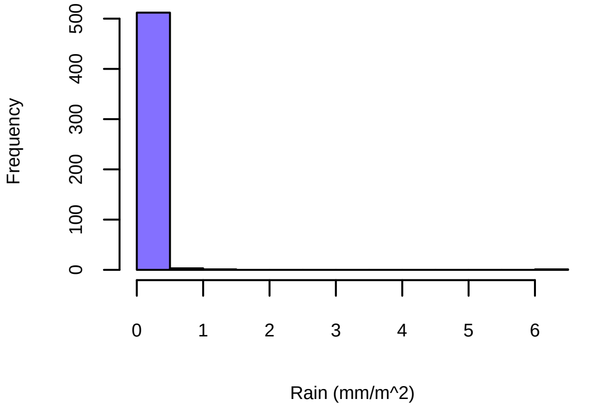


Figure 8. Histogram of Rain (mm/m2)



```
nrow(fires[fires$rain>0,c("rain","area_bins")])
```

```
## [1] 8
```

Indices

The graphs below (Figures 9-12) show the distribution of the 4 indices mentioned in our report: FFMC, DMC, DC, and ISI. For our report we will not dive into the criteria of how these indices were created as that would be beyond the scope of this report. However, the histrograms of the variables deserve a closer look as none of them appear to be normally distributed.

For the FFMC index we see that the chart is negatively skewed with the majority of values clustered towards the mean of 90.64 and a few outlier values that are well below the mean.

When looking at the histrogram of the DMC index, we see that the distribution is rather bimodal. A bulk of the values are clustered around the 100-150 values while we also have a cluster of values around the 0-50 range.

Looking at the chart for DC, we see that the curve is negatively skewed and somewhat bimodal with the mean being 547.9. We see a large group of values a bit above the mean and then we see a somewhat smaller cluster of values in the 0-100 bucket.

Finally, we get to the ISI index which shows relatively normal distribution with a slight positive skew and a mean of 9.022.

```r
par(mfrow=c(2,2))
hist(fires$FFMC, col = "cornflowerblue",
     main = "Figure 9. Histogram of FFMC Index", cex.main=0.8, cex.lab=0.7,
     yaxt = "n", xaxt = "n", xlab = "FFMC")
axis(2, cex.axis = 0.7)
axis(1, cex.axis = 0.7)
hist(fires$DMC, col = "cornflowerblue",
     main = "Figure 10. Histogram of DMC Index", cex.main=0.8, cex.lab=0.7,
     yaxt = "n", xaxt = "n", xlab = "DMC")
axis(2, cex.axis = 0.7)
axis(1, cex.axis = 0.7)
hist(fires$DC, col = "cornflowerblue",
     main = "Figure 11. Histogram of DC Index", cex.main=0.8, cex.lab=0.7,
     yaxt = "n", xaxt = "n", xlab = "DC")
axis(2, cex.axis = 0.7)
axis(1, cex.axis = 0.7)
hist(fires$ISI, breaks = 10, col = "cornflowerblue", cex.main=0.8, cex.lab=0.7,
     main = "Figure 12. Histogram of ISI Index",
     yaxt = "n", xaxt = "n", xlab = "ISI")
axis(2, cex.axis = 0.7)
axis(1, cex.axis = 0.7)
```

Figure 9. Histogram of FFMC Index

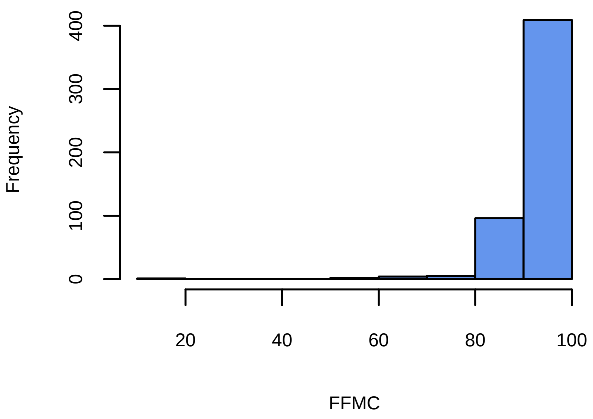


Figure 10. Histogram of DMC Index

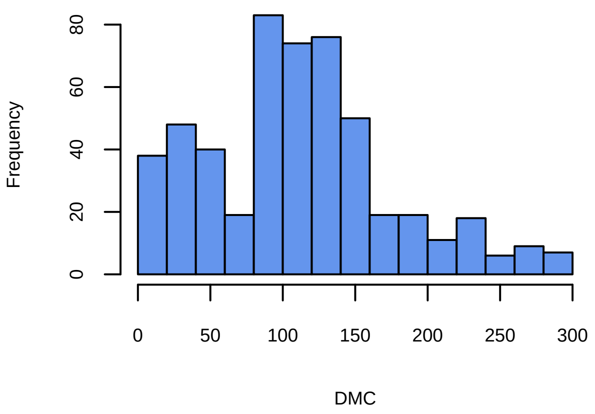


Figure 11. Histogram of DC Index

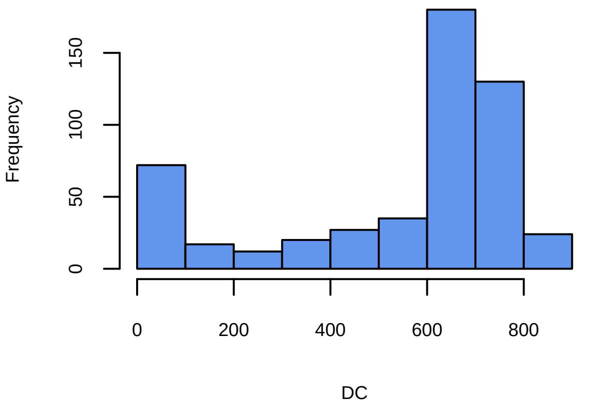
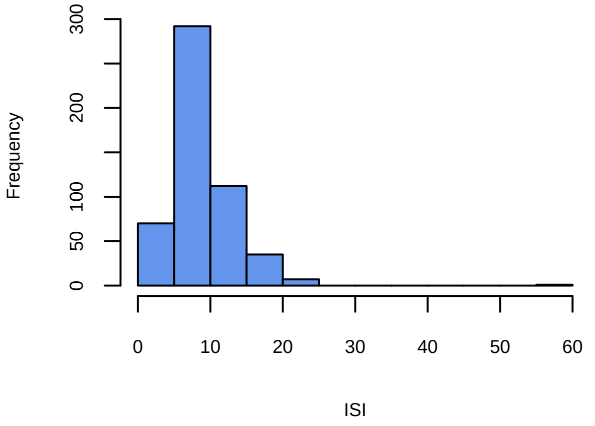


Figure 12. Histogram of ISI Index



Analysis of Key Relationships

Spatial Variables

Figure 13, shows the number of fires that occured in each location of the park based on the X and Y coordinates (X,Y). The figure shows that the fires were not evenly distributed accross the park, but concentrated in specific locations. The positions with the highest fire occurencies are (8,6) with 52 fires, (6,5) with 49 fires, and (7,4) with 45 fires.

Figure 14, shows the total *area* burned in each location of the park. Here we notice that the three locations mentioned above, with the largest number of fire occurencies, also have the highest total burned area, 1265.3 ha, 1384.05 ha, and 474.37 ha for (8,6), (6,5), and (7,4) respectively. Additionally, we notice that a few locations with very small number of fires (1-2) have a relatively large area burned. For example, location (8,8) consists of only one fire instance but it has a burned area of 185.76 ha, indicating that there was a very large fire.

```r
library(reshape)
library(gplots)
```

```
t1 = cast(fires, Y ~ X, value = "area")
t2 = cast(fires, Y ~ X, value = "area", fun.aggregate=sum)

col = colorRampPalette(c("slategray1","steelblue1","royalblue1"))(25)
par(cex.main=0.7)
heatmap.2(as.matrix(t1), cellnote = round(as.matrix(t1),2),
          main = "Figure 13. Number of Fires by Location",
          notecol="black", notecex=0.8,  density.info="none", trace="none",
          xlab = "X Position", ylab = "Y Position", srtCol = 0,
          cexRow = 1, cexCol = 1, margins =c(4,4),
          col = col, key.title = NA, symm=TRUE, Rowv=NA, Colv=NA )
par(cex.main=0.7)
heatmap.2(as.matrix(t2), cellnote = round(as.matrix(t2),1),
          main = "Figure 14. Sum of Burned Area (ha) by Location",
          notecol="black", notecex=0.8,  density.info="none", trace="none",
          xlab = "X Position", ylab = "Y Position", srtCol = 0,
          cexRow = 1, cexCol = 1.0, margins =c(4,4),
          col = col, key.title = NA, symm=TRUE, Rowv=NA, Colv=NA )
```

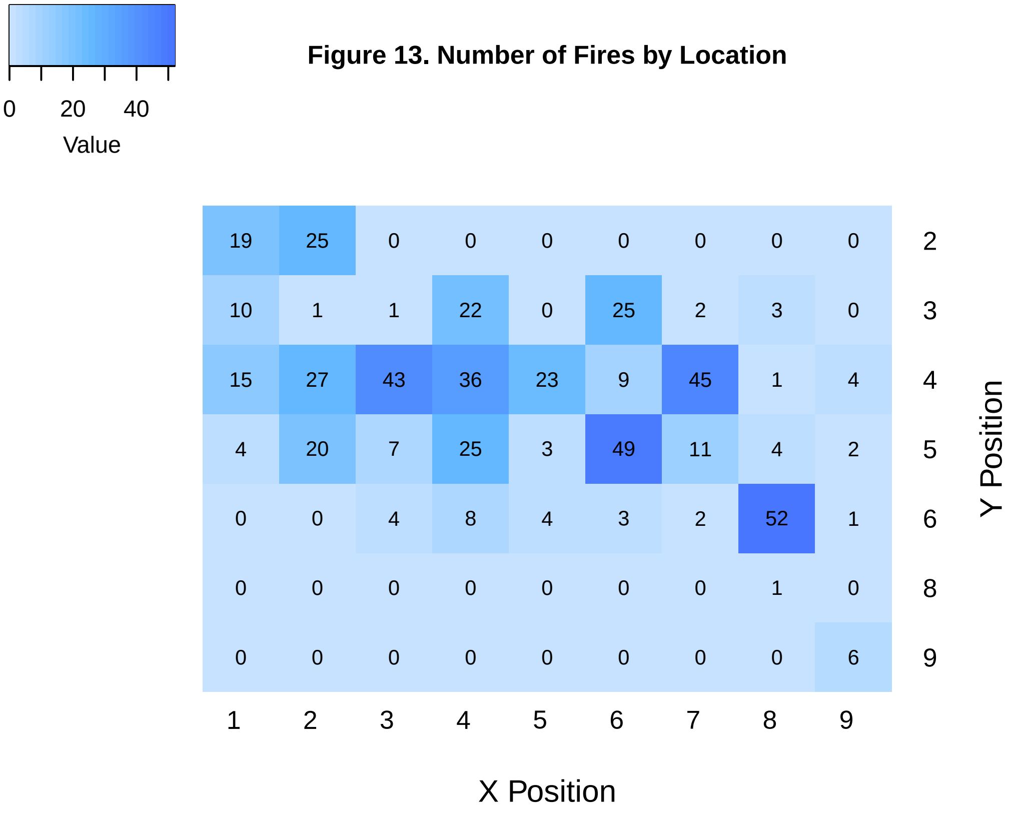


Figure 13. Number of Fires by Location

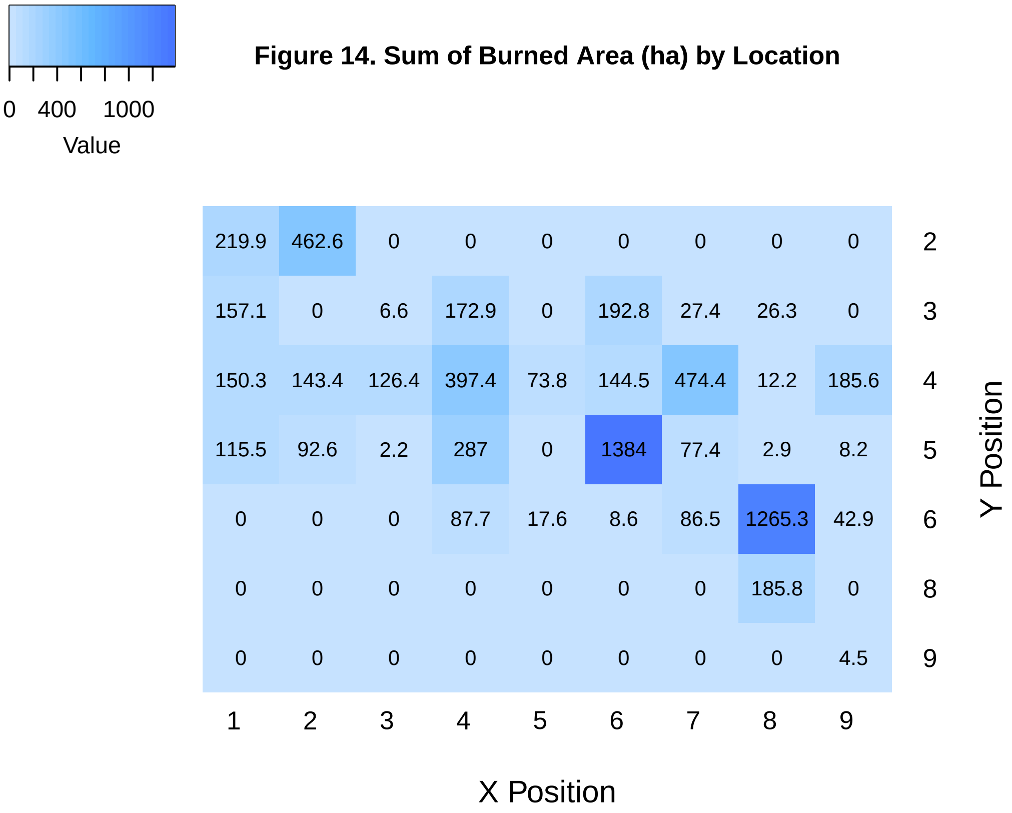


Figure 14. Sum of Burned Area (ha) by Location

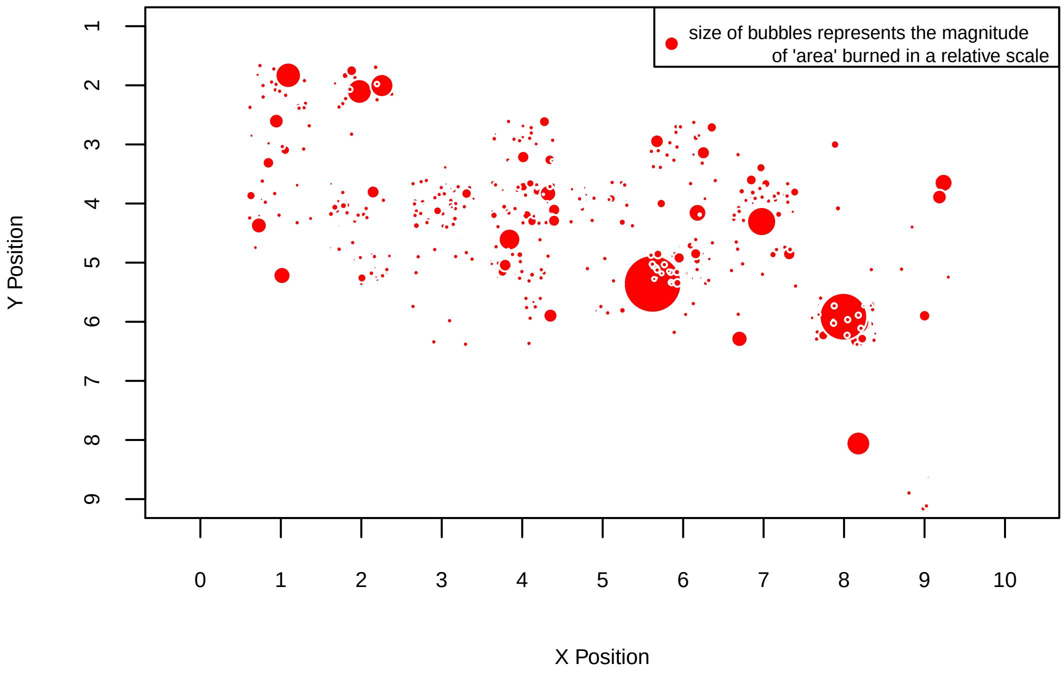
Y Position \ X Position	1	2	3	4	5	6	7	8	9
2	219.9	462.6	0	0	0	0	0	0	0
3	157.1	0	6.6	172.9	0	192.8	27.4	26.3	0
4	150.3	143.4	126.4	397.4	73.8	144.5	474.4	12.2	185.6
5	115.5	92.6	2.2	287	0	1384	77.4	2.9	8.2
6	0	0	0	87.7	17.6	8.6	86.5	1265.3	42.9
8	0	0	0	0	0	0	0	185.8	0
9	0	0	0	0	0	0	0	0	4.5

In order to get a better understanding of the distribution of fires across the area of the park based on the size of the fires, we look at Figure 15. This figure represents each fire by location (X and Y positions) and by magnitude of burned area (i.e. damage). The larger the damage the larger the "bubble" on the graph. This figure, helps us visualize the locations of the most damaging fires. We observe a cluster of large fires in the (8,6) position which is the second largest location in terms of total area burned. We also observe a cluster of large fires in the (6,5) position which is the first largest location in terms of total area burned. Both locations, have multiple fire occurencies but what differentiates them from other clusters, is the fact that they both have one very large fire instance in terms of area burned.

```r
radius = sqrt( fires$area/pi)

symbols(jitter(fires$X, factor=2), jitter(fires$Y, factor=2),
        main = "Figure 15. Fire Positions by Size of Burned Area (ha)",
        cex.main=0.8, cex.lab=0.7,
        yaxt = "n", xaxt = "n", ylim=c(9,1),
        circles=radius, inches=0.15, fg="white", bg="red", xlab = "X Position",
        ylab = "Y Position")
axis(2, cex.axis = 0.7, at = seq(9, 1, by = -1))
axis(1, cex.axis = 0.7, at = seq(0, 10, by = 1))
legend("topright",
       legend=c("size of bubbles represents the magnitude
                of 'area' burned in a relative scale"), pch = 21,
       cex=0.6, col = "white", pt.bg = "red", pt.cex = c(1))
```

Figure 15. Fire Positions by Size of Burned Area (ha)



Correlations

We notice several important features in the correlation matrix (Figure 16):

1. There are no correlations equal to or greater than 0.1 between our primary variable of interest (area) and any of the other variables. We know from the earlier histogram of the area variable that there is an extreme positive skew of the area observations, and this could be preventing achieving higher correlation values.

2. There are larger correlations between some of the other variables. Most of the largest correlations occur between the fire indices. We flag this as an indicator that the indices may be constructed from some of the same weather data variables. We will explore this further in the secondary effect section. Two particularly large correlations among the indices were DMC and DC (correlation of 0.68), and ISI and FFMC (correlation of 0.53).

3. In contrast to the fire indices, the weather variables showed lower correlations between each other - with the notable exception of temperature and RH (correlation of -0.53). The next largest correlation was between temperature and wind (correlation of -0.23). The remaining weather variable pairings showed 0.1 correlation or smaller.

4. Correlations between a fire index and a weather variable were strongest when the weather variable was temperature, which correlated to DC at 0.50, with DMC at 0.47, with FFMC at 0.43, and ISI at 0.39.

```
library(gplots)
cor_matrix = cor(fires[ ,c("FFMC", "DMC", "DC", "ISI","temp", "RH", "wind",
                           "rain", "area")], use = "complete.obs")
col = colorRampPalette(c("red","white","blue"))(20)
par(cex.main=0.7)
heatmap.2(cor_matrix, cellnote = round(cor_matrix,2),
          main = "Figure 16. Correlation Matrix",
          notecol="black", notecex=0.8, density.info="none", trace="none",
          cexRow = 1, cexCol = 1, margins =c(5,5),
          col = col, key.title = NA, symm=TRUE,
          Rowv=NA, Colv=NA )
```

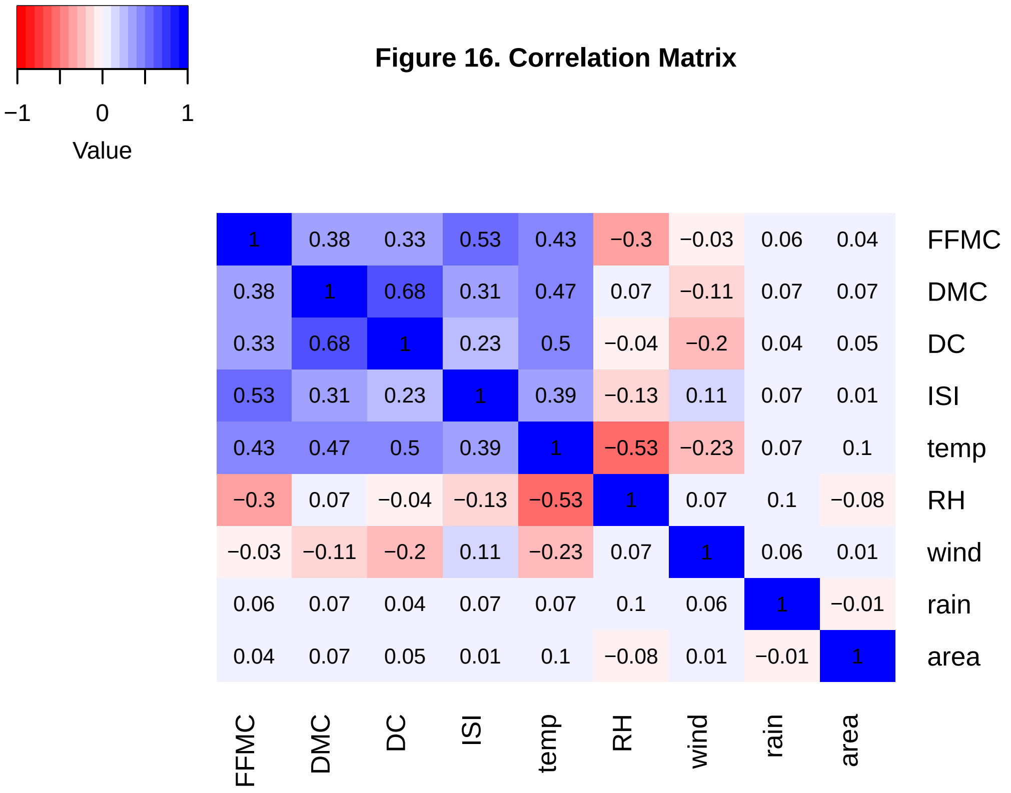


Figure 16. Correlation Matrix

	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
FFMC	1	0.38	0.33	0.53	0.43	−0.3	−0.03	0.06	0.04
DMC	0.38	1	0.68	0.31	0.47	0.07	−0.11	0.07	0.07
DC	0.33	0.68	1	0.23	0.5	−0.04	−0.2	0.04	0.05
ISI	0.53	0.31	0.23	1	0.39	−0.13	0.11	0.07	0.01
temp	0.43	0.47	0.5	0.39	1	−0.53	−0.23	0.07	0.1
RH	−0.3	0.07	−0.04	−0.13	−0.53	1	0.07	0.1	−0.08
wind	−0.03	−0.11	−0.2	0.11	−0.23	0.07	1	0.06	0.01
rain	0.06	0.07	0.04	0.07	0.07	0.1	0.06	1	−0.01
area	0.04	0.07	0.05	0.01	0.1	−0.08	0.01	−0.01	1

As discussed previously, the extreme positive skew of the area variable led us to bin our outcome variable into zero damage, small damage, and large damage groupings. We now look at the correlations between these groupings and the other variables in our sample (Table 3).

We note a few features in this tables that can help guide our analysis:

1. We have omitted Zero Damage bin data from the table. There are no correlations between area and the other variables in the Zero Damage bin because the observations all have the same zero value for area. Similarly, we have ommited rain from the Large Damage bin correlations because all rain values are zero.
2. Binning the data allows us to see slightly stronger correlations between the area variable and some of the other variables than we couldn't see in the unbinned data. The largest correlations are found in the Large Damage bin, between temperature and area (0.25), RH and area (-0.21), and DMC and area (0.2).
3. The Small Damage bin also showed stronger correlations with the area variable than seen in the unbinned data, but to a lesser degree than the Large Damage bin. While most of the correlations with the area variable in the Large Damage bin were positive, the correlations with the area variable in the Small Damage bin were negative. This suggests fire may behave differently at smaller sizes than larger sizes, providing support to our decision to bin the fires by size.

```
library(stargazer)
library(reshape)

smallbin_corrs = c(with(fires.small, cor(temp, area)), with(fires.small, cor(RH, area)), with(fires.smal
largebin_corrs = c(with(fires.mega, cor(temp, area)), with(fires.mega, cor(RH, area)), with(fires.mega,

bins_corrs <- data.frame(smallbin_corrs, largebin_corrs)

names(bins_corrs) <- c("Small Damage fires", "Large Damage fires")
row.names(bins_corrs) <- c("temp", "RH", "wind", "rain", "FFMC", "DMC", "DC", "ISI")

stargazer(bins_corrs, summary=FALSE, title = "Correlations with fire area variable", type = "latex", di
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Apr 01, 2018 - 19:27:18

Table 3: Correlations with fire area variable

	Small Damage fires	Large Damage fires
temp	-0.19	0.25
RH	-0.05	-0.21
wind	0.15	-0.01
rain	0.05	
FFMC	-0.15	0.11
DMC	-0.07	0.20
DC	-0.05	0.14
ISI	-0.12	0.03

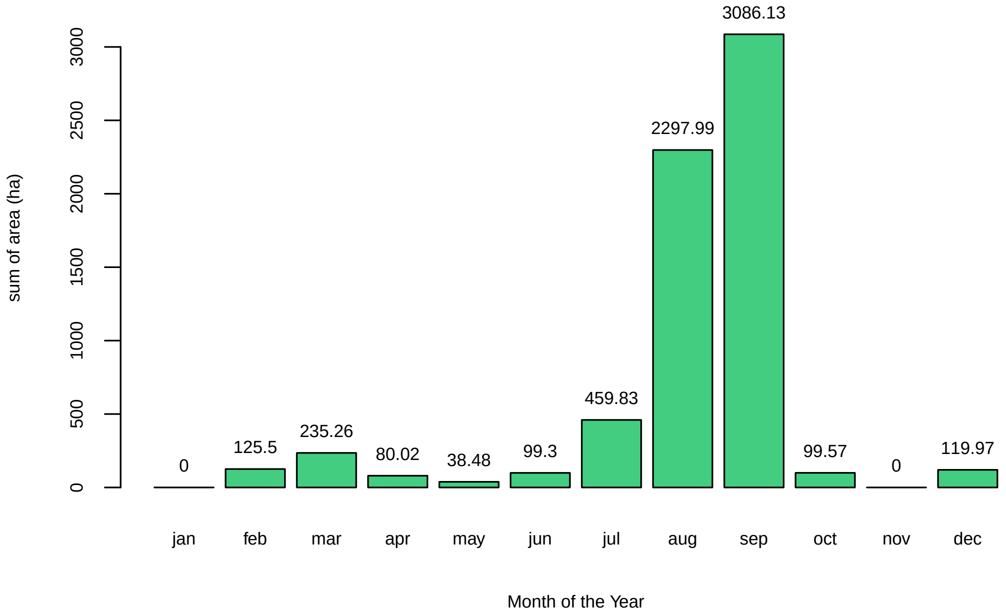
Area burned by Month and Day

This section examines the relationship between the area burned and the “time” variables (i.e. *month* and *day*). Figure 17, shows the sum of area burned in hactares by month. September is the month with the most area burned (3086 ha), followed by August (2297.99 ha). In Figure 3 of the univariate analysis section, we saw that most fires occur in August (184 fires) and September (172 fires). Even though August is the month with the most fire instances, September is the month with the most damage caused by fires.

Figure 18, shows the sum of area burned in hactares by day of the week. Here we observe that Saturday has the biggest amount of total burned area (2144.86 ha), more than twice the size of the other days. This is an interesting finding, considering that in terms of number of fire instances (see Figure 4), Saturday comes 3rd with 84 fires after Sunday (95 fires) and Friday (85 fires).

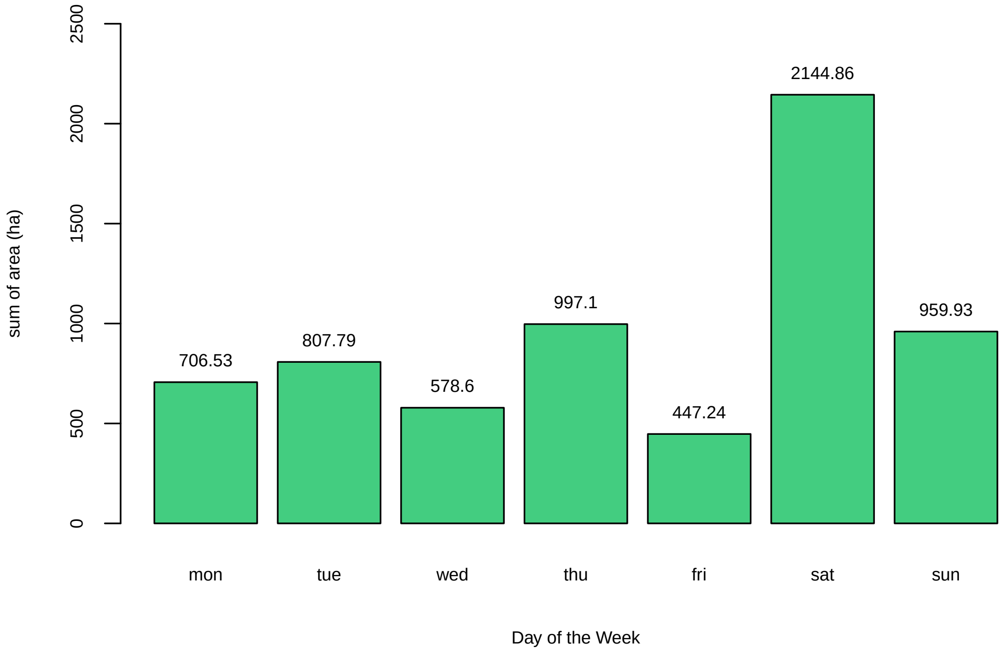
```
area_month_sum = by(fires$area, fires$month,  sum)
bp = barplot(area_month_sum, main="Figure 17. Total Burned Area (ha) by Month",
             ylim = c(0,3400), cex.main=0.8, cex.lab=0.7, cex.names=0.7, yaxt = "n",
             xlab="Month of the Year", ylab = "sum of area (ha)",  col="seagreen3")
text(x = bp, y = area_month_sum, label = round(area_month_sum,2), pos = 3, cex = 0.7)
axis(2, cex.axis = 0.7)
```

Figure 17. Total Burned Area (ha) by Month



```r
area_day_sum = by(fires$area, fires$day,  sum)
bp2 = barplot(area_day_sum, main="Figure 18. Total Burned Area (ha) by Day of the Week",
              ylim = c(0,2500), cex.main=0.8, cex.lab=0.7, cex.names=0.7, yaxt = "n",
              xlab="Day of the Week", ylab = "sum of area (ha)",  col="seagreen3")
text(x = bp2, y = area_day_sum, label = round(area_day_sum,2), pos = 3, cex = 0.7)
axis(2, cex.axis = 0.7)
```

Figure 18. Total Burned Area (ha) by Day of the Week



Given the dramatic clustering in the time variables with far more fires on weekends and in late summer months, we decided to look more closely at the variables across time and within key time periods.

First, we compared our bucketed data across time to see whether the proportions of zero, small, and large damage fires change by month or day. We see below (Figures 19 & 20), however, that the bucketed data are distributed roughly proportionally across the time variables. We note a slight uptick in large damage fires in September from August, and a slight uptick in small and zero damage fires on Sunday versus Saturday.

```
library(ggplot2)

# see how the fires of each type stack up each month
cp = ggplot(fires, aes(x = month, y = area_bins, fill = area_bins)) + geom_bar(stat = "identity") + them
cp + labs(title = "Figure 19. Fires by month, broken up by size")

cp2 = ggplot(fires, aes(x = day, y = area_bins, fill = area_bins)) + geom_bar(stat = "identity") + theme
cp2 + labs(title = "Figure 20. Fires by day of week, broken up by size")
```

Figure 19. Fires by month, broken up by size

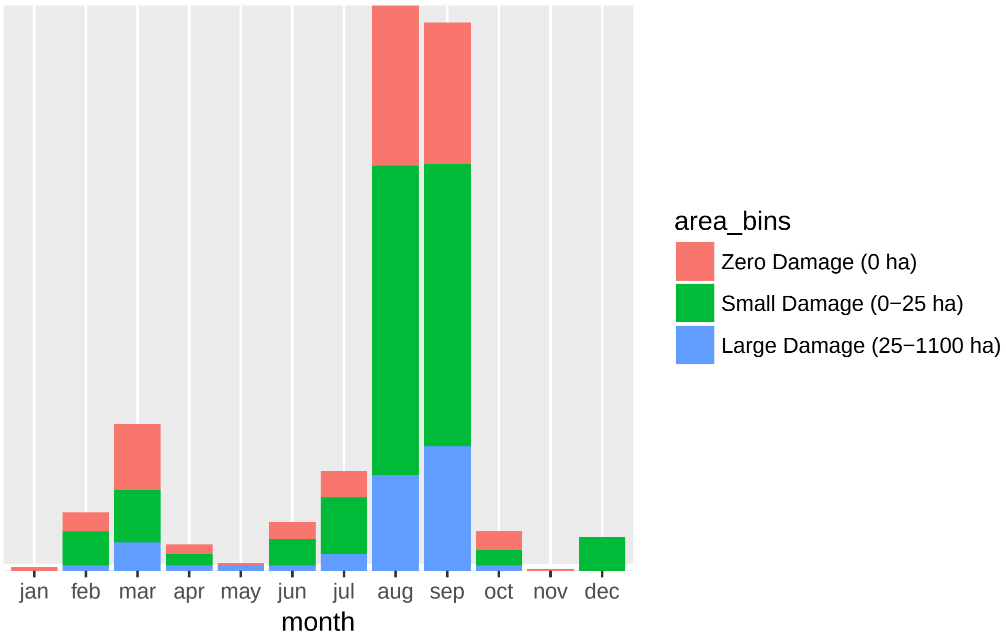
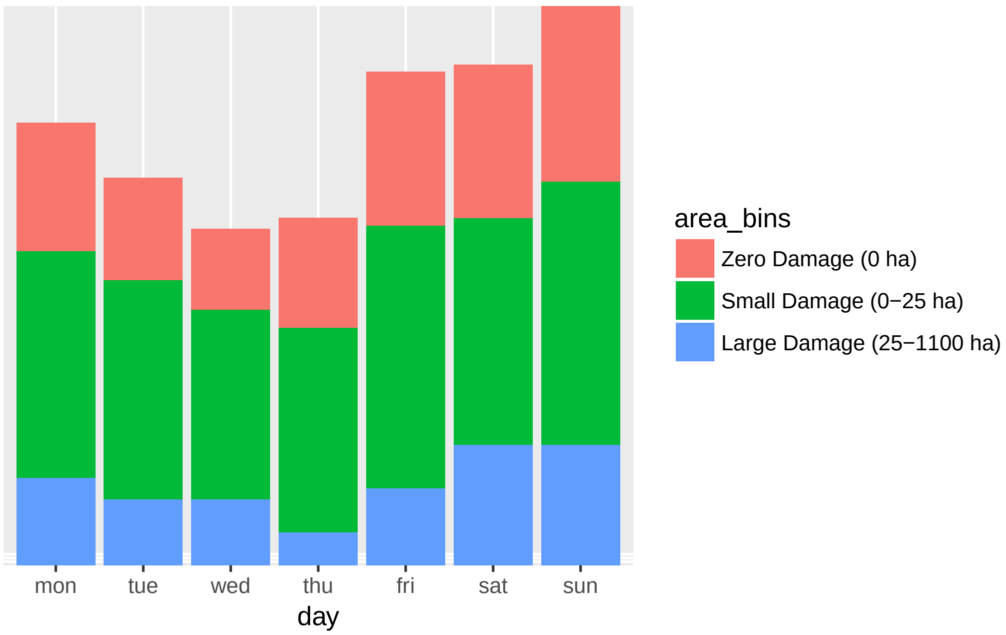


Figure 20. Fires by day of week, broken up by size



Our next step was to run boxplots for each month and each day for all of the variables. The day of week boxplots did not reveal any interesting variations, however the monthly plots show noticeable seasonal variations. We note that temperature peaks in the summer months of July and August and begins to decline in September (Figure 21); this does not map exactly to the summer spike in fires, which happens a little later in August and September. The monthly plots of the Drought Code (Figure 23) shed some light on why the summer spike in fires may occur later than the temperature spike - drought conditions reach their highest levels in September. Additionally, we can see the relative humidity falling in September from August (Figure 22). Meanwhile, the monthly plots of relative humidity, wind, and ISI index (Figures 22, 24 & 25) suggest these may be critical factors for the small spike in fires in March.

```
boxplot(fires$temp ~ fires$month, col="orange", cex.main=0.8, cex.lab=0.7, cex.names=0.7, yaxt = "n",
        main="Figure 21. Monthly Distribution of Temperature", xlab=" ", ylab="Temp")
axis(2, cex.axis = 0.7)
boxplot(fires$RH ~ fires$month, col="orange", cex.main=0.8, cex.lab=0.7, cex.names=0.7, yaxt = "n",
        main="Figure 22. Monthly Distribution of Relative Humidity (RH)", xlab=" ", ylab="RH")
axis(2, cex.axis = 0.7)
boxplot(fires$DC ~ fires$month, col="orange", cex.main=0.8, cex.lab=0.7, cex.names=0.7, yaxt = "n",
        main="Figure 23. Monthly Distribution of Drought Conditions (DC)", xlab=" ", ylab="DC")
axis(2, cex.axis = 0.7)
boxplot(fires$wind ~ fires$month, col="orange", cex.main=0.8, cex.lab=0.7, cex.names=0.7, yaxt = "n",
        main="Figure 24. Monthly Distribution of Wind", ylab="Wind", xlab=" ")
axis(2, cex.axis = 0.7)
boxplot(fires$ISI ~ fires$month, col="orange", cex.main=0.8, cex.lab=0.7, cex.names=0.7, yaxt = "n",
        main="Figure 25. Monthly Distribution of Initial Spread Index (ISI)", xlab=" ", ylab="ISI")
axis(2, cex.axis = 0.7)
```

Figure 21. Monthly Distribution of Temperature

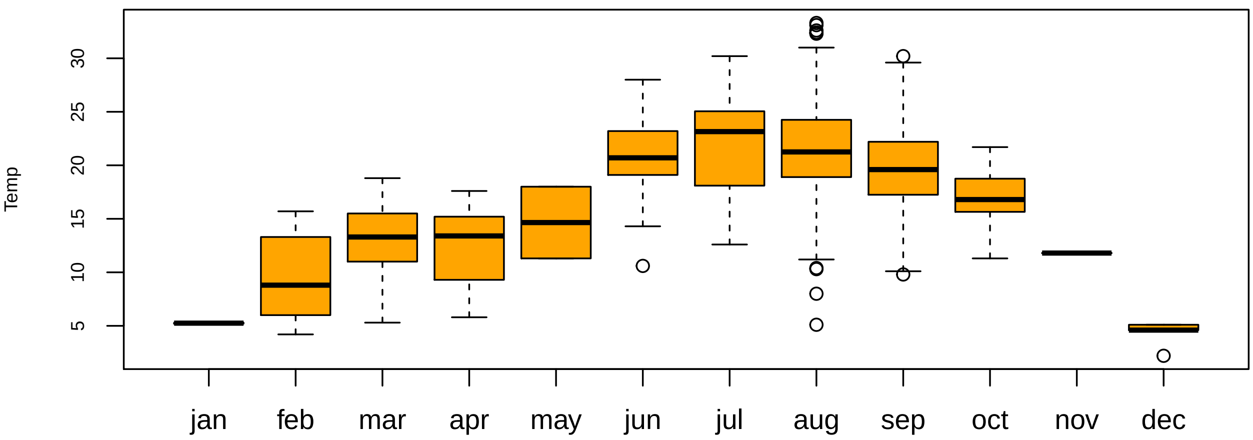


Figure 22. Monthly Distribution of Relative Humidity (RH)

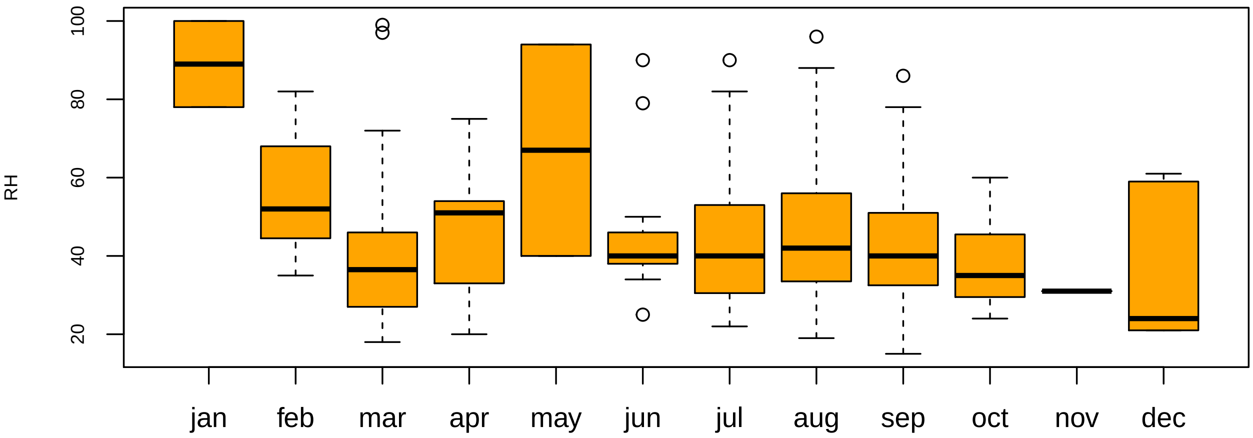


Figure 23. Monthly Distribution of Drought Conditions (DC)

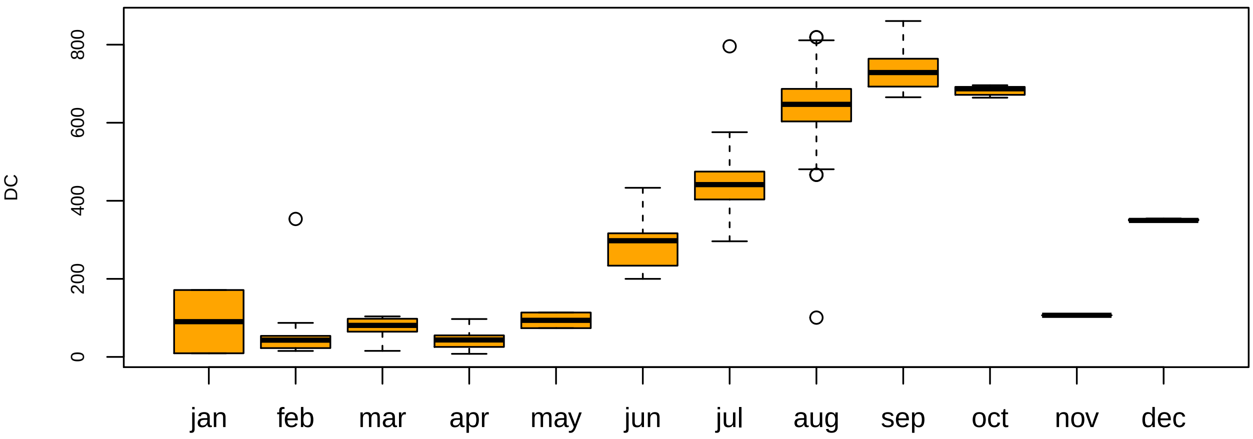


Figure 24. Monthly Distribution of Wind

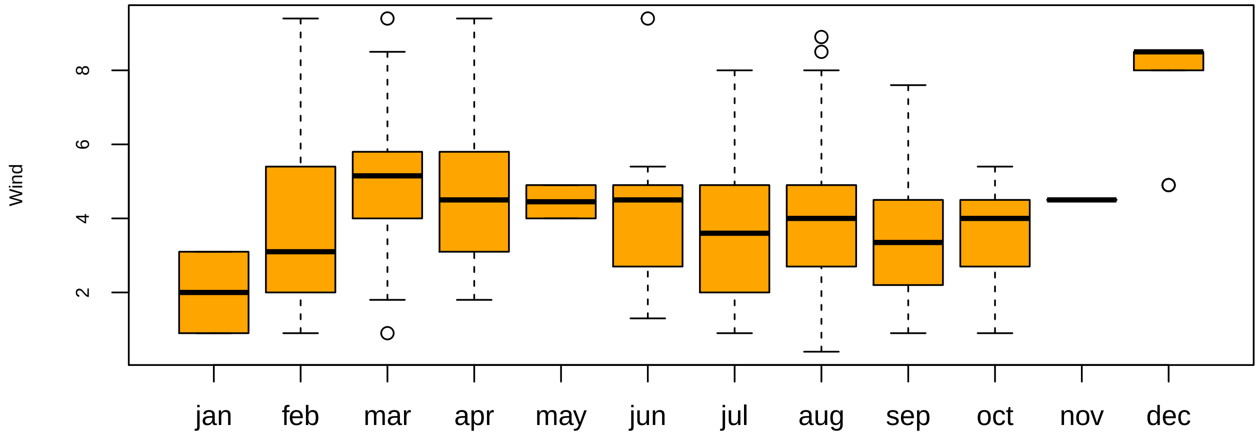
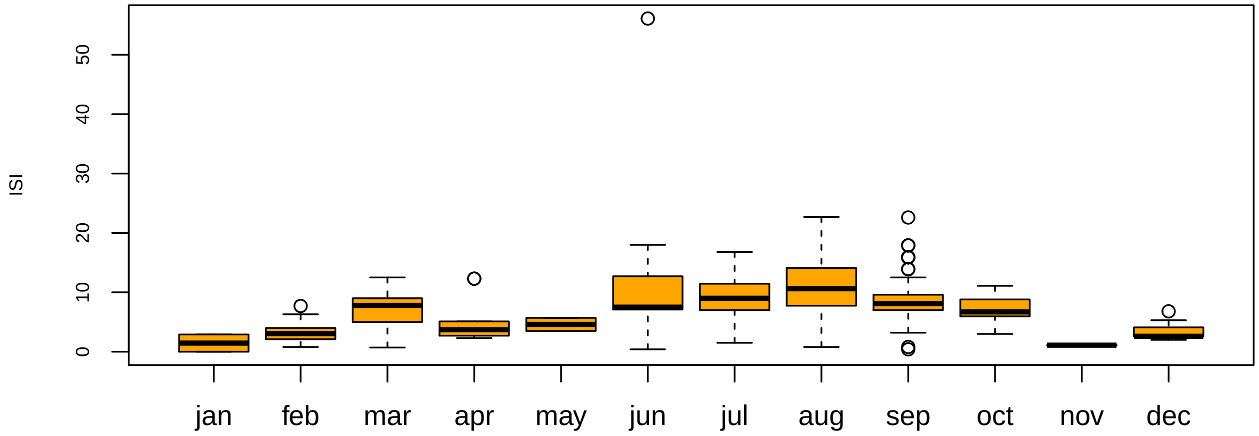


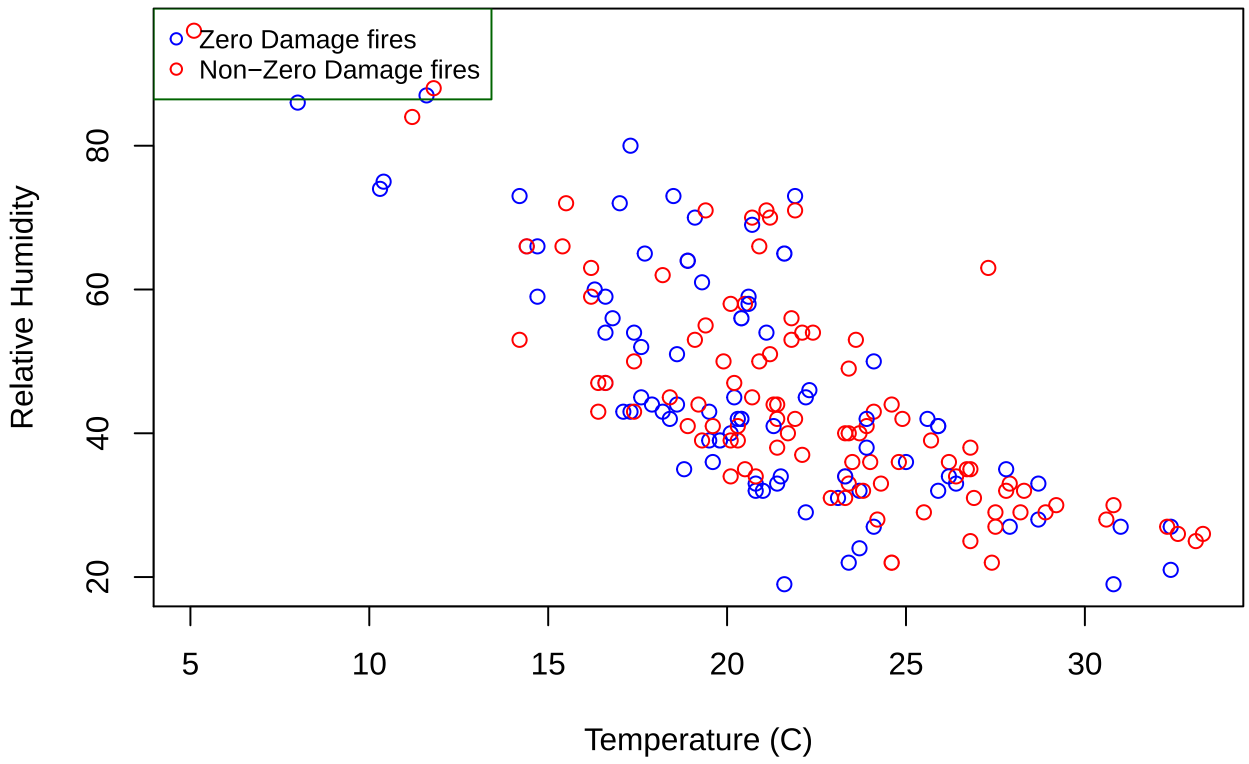
Figure 25. Monthly Distribution of Initial Spread Index (ISI)



Next, we looked at the data within individual months with spikes in fire activity: March, August, and September. This did not result in more substantial correlations with the fire area variable. Correlations in the +/- 0.2 to 0.3 range only reemerged when we further narrowed the data to large damage fires. Adding the dimension of the fire size bins in the August data showed how higher temperatures and lower relative humidity were linked to larger fires. In the chart below (Figure 26), the blue dots represent the zero damage fires, and the red dots are the non-zero damage fires (Small & Large Damage). We can see that for one of the hottest months (August), the smallest fires happened more frequently on cooler, damper days and the larger fires happened more frequently on hotter, drier days. This finding and surrounding exploration suggests that the role of temperature and humidity in fires is clearer when looking at larger-sized fires in hotter months.

```
fires.aug = subset(fires, month == "aug")
plot(fires.aug$temp, fires.aug$RH, xlab = "Temperature (C)", ylab = "Relative Humidity",
     main="Figure 26. Temperature and Relative Humidity in August",
     col=ifelse(fires.aug$area_bins=="Zero Damage (0 ha)", "blue", "red"))
legend("topleft", pch=c(1,1), col=c("blue", "red"), c("Zero Damage fires", "Non-Zero Damage fires"), bt
```

Figure 26. Temperature and Relative Humidity in August



Analysis of Secondary Effects

After viewing the correlation matrix on figure 16 as previously shown, we notice some confounding effects on the variables we have identified.

Referring back to Figure 16, we can view the correlation matrix to get a better understanding of correlations amongst the variables themselves. Looking solely at the weather variables (temperature, wind, rain, and RH), we notice that temperature has a correlation of -0.53 to RH and a correlation of -.23 to wind. Figure 27 below demonstrates one of these relationships, between temperature and relative humidity, and the correlation is visible from the scatter plot. Thus, it is difficult for us to solely look at the relationship of temperature with the outcome variable (area), because RH and wind can have confounding effects on temperature.

Upon solely viewing the correlation matrix for the indices (Figure 16), we notice a relatively high positive correlation within the indices themselves. The positive correlation values for the indices (FFMC, DMC, DC, and ISI) range from a positive correlation of 0.23 to 0.68. Figure 28 displays one of the high correlations of 0.68 between DMC and the DC indices.

Looking at the relationship between the indices and the weather variables on the correlation matrix, we notice that temperature, RH, and wind each have correlations with the indices with temperature being the most profound relationship with the indices. Temperature has a range of correlations of 0.39 to 0.5 across the indices, RH has a range of correlations of -0.3 to 0.07, and wind has a range of correlations of -0.2 to 0.11 across the indices. Figure 29 displays one of the high correlation (0.47) of Temperature to the indices (DMC).

In addition, many confounding variables could exist that aren't contained within this sample dataset. For example, the speed of response of local fire fighters could influence if small fires grow to large ones. In another example, we have access to wind speed but not the exposure of terrain within the park that could shield certain areas from its effects. This may explain why the number of fires by location are concentrated in certain regions as show in Figure 13.

```r
par(mfrow=c(3,1))
plot(fires$temp, fires$RH, main="Figure 27. Temperature (degrees C) vs Relative Humidity", xlab = "Tempe
plot(fires$DMC, fires$DC, main="Figure 28. DMC Index vs DC Index", xlab = "DMC Index", ylab = "DC Index
plot(fires$temp, fires$DMC, main="Figure 29. Temperature (degrees C) vs DMC Index", xlab = "Temperature
```

Figure 27. Temperature (degrees C) vs Relative Humidity

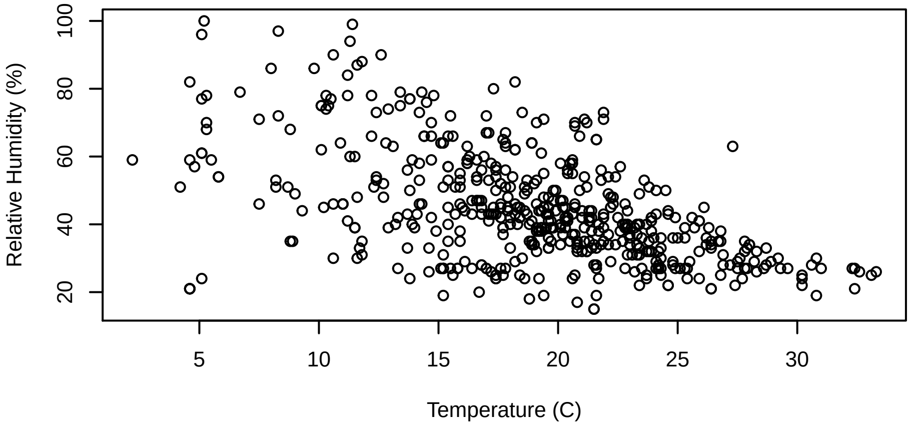


Figure 28. DMC Index vs DC Index

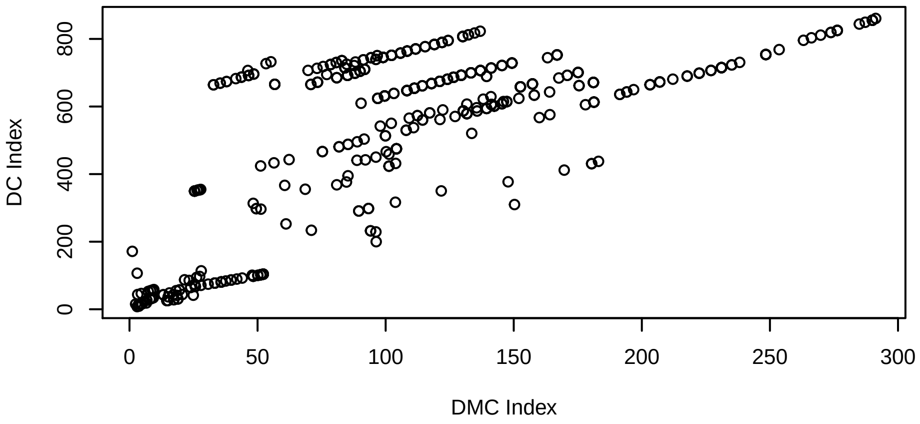
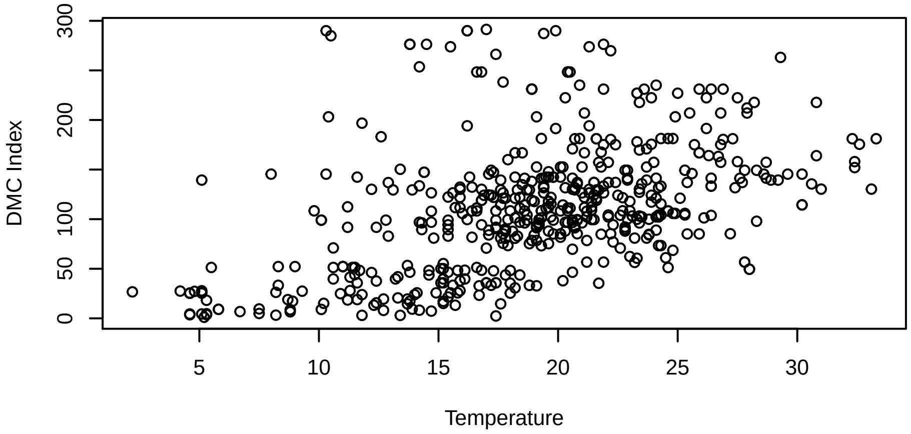


Figure 29. Temperature (degrees C) vs DMC Index



Conclusion

This study explores the variables that have an impact on damaging fires as defined by the total area burned. Through an exploratory data analysis, we uncovered some notable observations, but no single variable alone had a strong correlation with the output variable (i.e. area). From this dataset, we can conclude that burned area is not highly correlated with a single factor but, instead, is influenced by an interplay of many variables. Below we will discuss our key takeaways from this exploratory analysis.

Looking at the distribution of fire instances and burned area across the 9x9 spatial grid, we found that the most damaging fires were concentrated in certain locations within the park. Next, when we looked at the correlations between the outcome variable (area burned) and the rest of the variables we saw that there was no correlation greater than 0.1. After binning the fire observations into three groups based on damage size (zero, small and large), we found that there were stronger correlations between the burned area and the other variables for the non-zero damage fires. The largest correlations are found in the Large Damage group, between area and temperature (0.25), RH (-0.21) and DMC (0.2). Another notable conclusion is that the correlations between area and the other variables have opposite signs in the Large Damage and Small Damage groups, which suggests that fires may behave differently depending on their size.

When we examined the relationship between fire damage (i.e. burned area) and the month and day variables, we discovered that there were more fire instances and more burned area during the late summer months (August and September), and during the weekend. To explore the time component further, we looked at the distribution of the indices and weather variables across the months of the year. One notable finding was that the Drought Code (DC) reached its highest level in September, which is the month with the highest total area burned across the year. Additionally, during one of the hottest months (August) with the most fire occurrences, we found that the most damaging fires happened more frequently on days with high temperature and low relative humidity.

Our analysis on confounding effects showed that it is tricky to examine the bivariate relationship of a weather or index variable alone with the area burned, since there might be some confounding effects that influence this relationship. Specifically, temperature and relative humidity were highly correlated with most of the index variables, indicating that the indices may be constructed from some of the same weather variables.

Our analysis has yielded several insights into forest fires across several important variables (spatial, time, weather and fire indices). Our findings have created a starting roadmap to do additional statistical testing and analysis.