Introduction to Data Mining and Programming in R

Instructions

Download `forestfires.tsv data` and documentation, then use `dplyr` verbs and `ggplot2`/`ggpairs` visualizations to conduct an exploratory data analysis on the forest fires data set. Start by answering the following questions, and then pursue what interests you on your own. Submit a document with your answers, brief explanations, and any relevant figures.

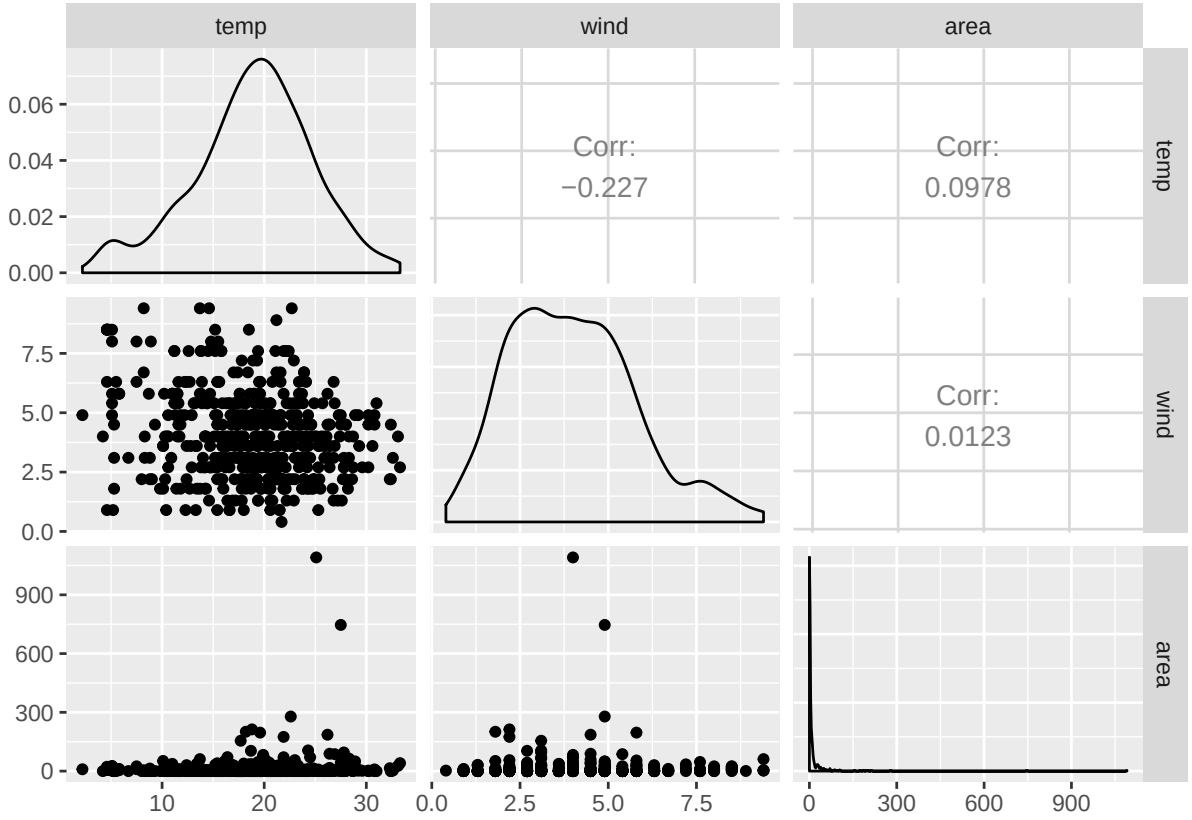
```
library(dplyr)
library(ggplot2)
library(GGally)

### Load the forest fires datasets on your machine by setting the working directory
### or specify the directory of your data

### This code is commented out and will NOT run
### ff = read.delim("forestfires.tsv", sep = '\t', header = TRUE)
```

1. We hypothesize that `temp` and `wind` are correlated with burn area (just called `area` in the data). Test that hypothesis with ggpairs.

```
ggpairs(select(ff, temp, wind, area))
```



A. What is the correlation between temp and burn area?

The correlation between burn area and temperature is about 0.1, which is a fairly small positive correlation. So increasing temperatures are slightly associated with increasing burn area in these data.

B. What is the correlation between wind and burn area?

The correlation between burn area and wind is about 0.01, which is a very small correlation.

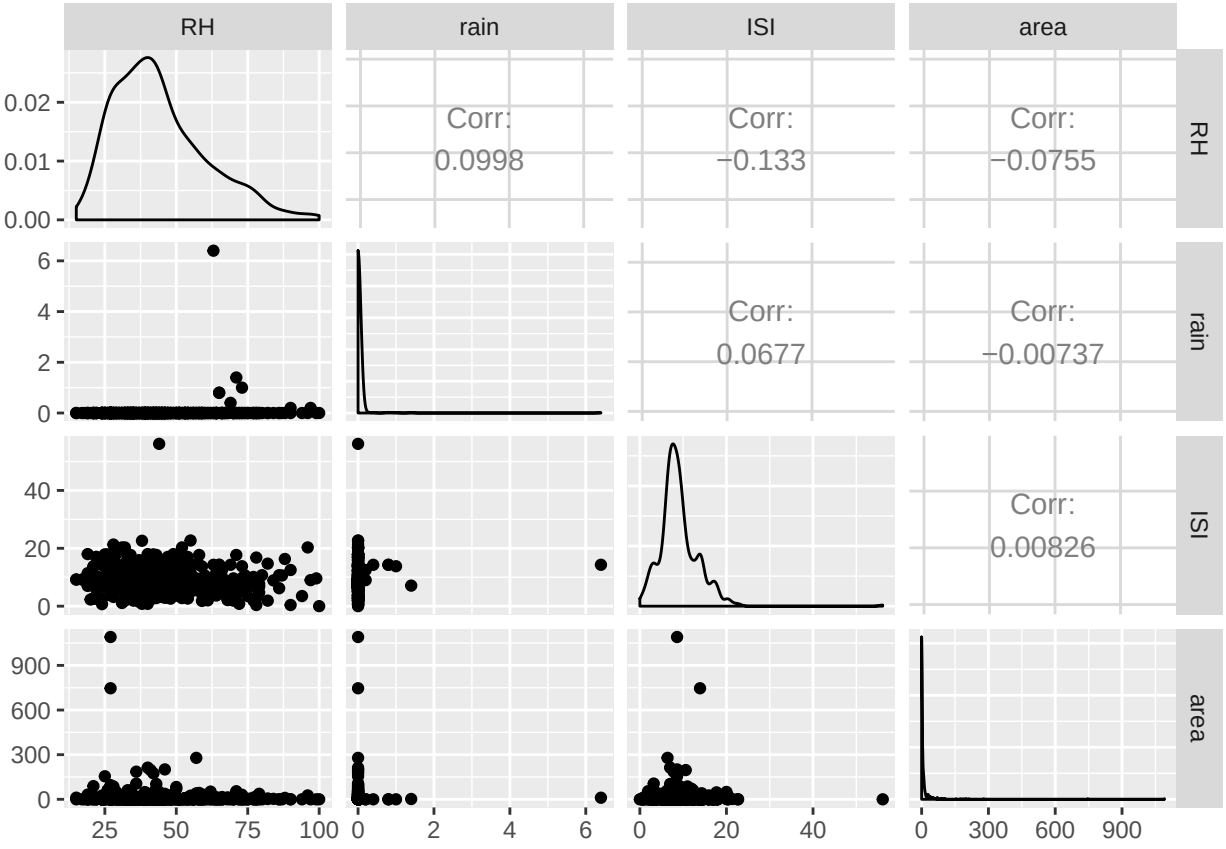
C. Is there evidence to support the hypothesis that temp and wind are correlated with burn area?

Based on the data, the evidence for a relationship between temp or wind and burn area is very limited.

2. Make some hypotheses about three additional variables that could be positively and negatively correlated with `burn_area`. Test your hypotheses using ggpairs.

Two variables we consider here that could be negatively associated with burn area are relative humidity (`RH`) and `rain`. We do find a negative correlation but it is very small, slightly highr for relative humidity. We also looked ta the `ISI`, which was minimally correlated with burn area.

```
ggpairs(select(ff, RH, rain, ISI, area))
```



3. Does burn area seems to be related to the season of the year?

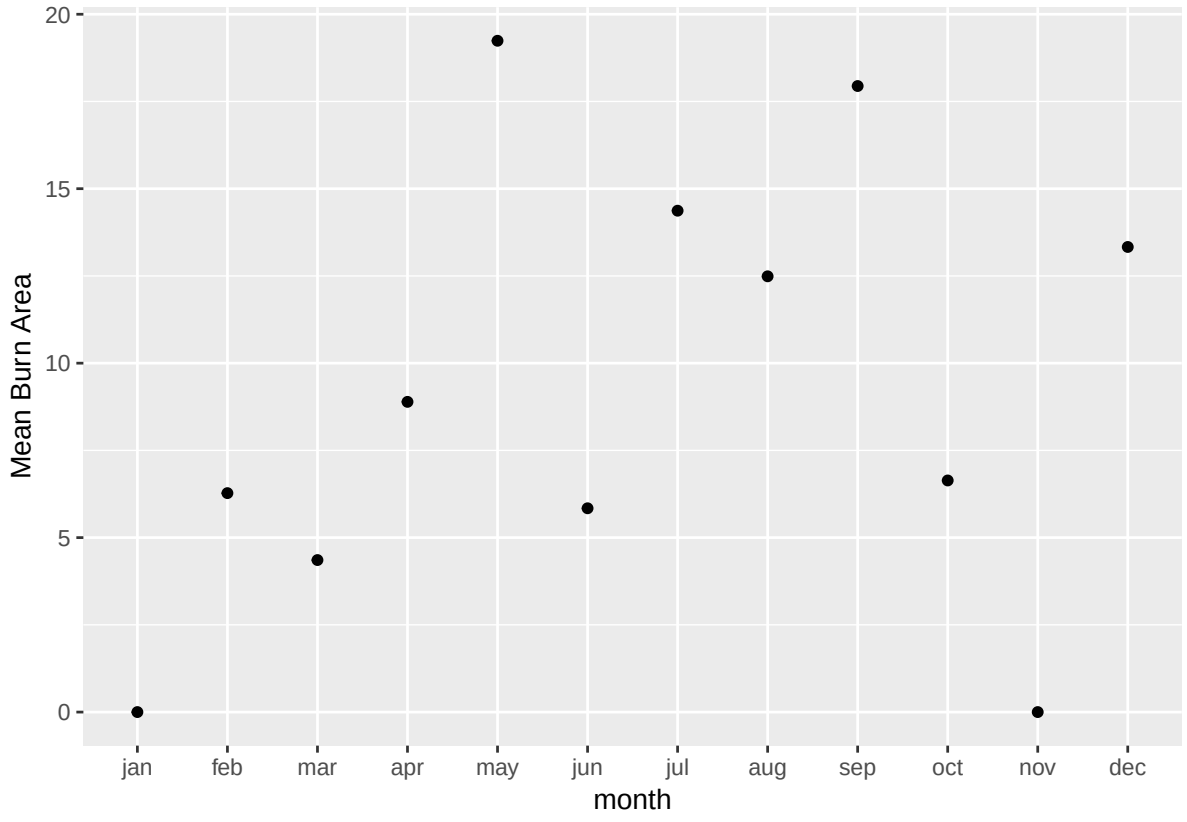
   A. Use `dplyr` verbs to aggregate by month and compute the mean burn area. Do you see evidence of a seasonal relationship?
   B. Which month has the largest average burn area?

   The highest mean burn area was in May. Burn area is generally elevated in the late spring and summer months.

```
seasonal_burn_area = ff %>%
  group_by(month) %>%
  summarize(
    mean_area = mean(area)
    )

# put the months in the right order
seasonal_burn_area$month = factor(seasonal_burn_area$month,
                                  levels = c("jan", "feb", "mar", "apr", "may", "jun",
                                             "jul", "aug", "sep", "oct", "nov", "dec"))

ggplot(data = seasonal_burn_area, aes(x = month, y = mean_area)) +
  geom_point() +
  scale_y_continuous("Mean Burn Area")
```



4. Are there limitations in your data that could invalidate your findings?

The months are not evenly sampled so it may be difficult to extract actual seasonal relationships from the data. The outcome is skewed towards zero, so most forest fires were small in magnitude. We look at the log of the burn area, to reduce the skew, compared to some predictors below:

```
ff$log_area = log(ff$area + 1)

ggpairs(select(ff, temp, wind, RH, rain, ISI, log_area))
```