Creating a machine learning analysis plan

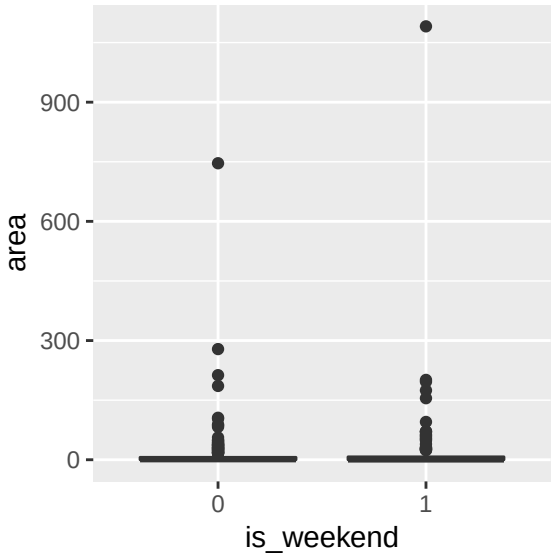
Instructions

If you haven't already, download `forestfires.tsv data` and documentation. Use the `caret` package to prepare the forest fire data for a machine learning analysis. When you finish preparing the data, save your script and upload it to Canvas.

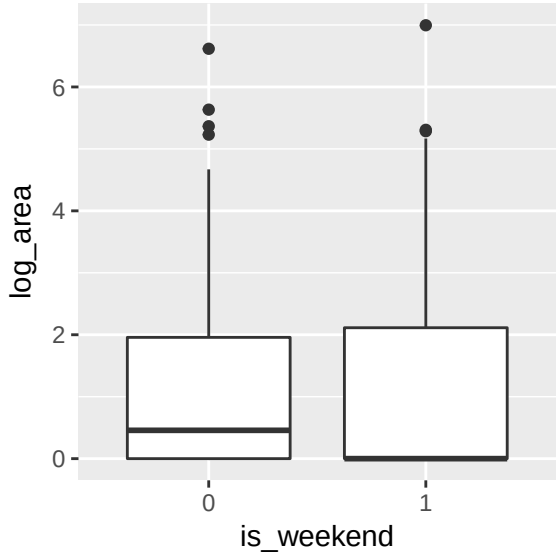
```
library(caret)

### 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. Make at least one new feature and plot it against the burn area. From visual inspection does there appear to be a relationship between the new feature and the burn area?

```
ff$is_weekend = ifelse(ff$day %in% c("sat", "sun"), 1, 0)
ff$is_weekend = factor(ff$is_weekend)

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

ggplot(ff, aes(x = is_weekend, y = area)) +
  geom_boxplot()
```




```
ggplot(ff, aes(x = is_weekend, y = log_area)) +
  geom_boxplot()
```



Since the area variable is so skewed, it is easier to see relationships on the log scale (second plot). The relationship between ***is_weekend*** *and burn area is not overwhelming but the median is lower on weekend. There is also a ggpairs plot of the* ***is_weekend*** *variable in the tutorial.*

2. Use `createDataPartition` to split 80% of the forest fire data into a training set.

```
in_train = createDataPartition(y = ff$log_area, p = 0.8, list = FALSE)
ff_train = ff[in_train, ]
ff_test = ff[-in_train, ]
```

3. Use `preProcess` to prepare your data for analysis. What, if any, variables were removed for near zero variance?

*One variable (**`rain`**) is removed. Following the tutorial, using the **`model.matrix`** method helps find other near zero variance categories in the factor variables.*

```
preprocess_steps = preProcess(select(ff_train, FFMC, DMC, DC, ISI, temp, RH, wind, rain),
                              method = c("center", "scale", "nzv"))

preprocess_steps
```

```
## Created from 416 samples and 8 variables
##
## Pre-processing:
##   - centered (7)
##   - ignored (0)
##   - removed (1)
##   - scaled (7)
```

```
nearZeroVar(ff_train, saveMetrics = TRUE)
```

```
##             freqRatio percentUnique zeroVar   nzv
## X            1.085714     2.1634615   FALSE FALSE
## Y            1.687500     1.4423077   FALSE FALSE
## month        1.145985     2.8846154   FALSE FALSE
## day          1.212121     1.6826923   FALSE FALSE
## FFMC         1.090909    23.5576923   FALSE FALSE
## DMC          1.142857    45.1923077   FALSE FALSE
## DC           1.142857    45.9134615   FALSE FALSE
## ISI          1.055556    26.9230769   FALSE FALSE
## temp         1.333333    43.2692308   FALSE FALSE
## RH           1.350000    17.5480769   FALSE FALSE
## wind         1.045455     5.0480769   FALSE FALSE
## rain       204.000000     1.6826923   FALSE  TRUE
## area        66.000000    48.7980769   FALSE FALSE
## is_weekend   1.849315     0.4807692   FALSE FALSE
## log_area    66.000000    48.7980769   FALSE FALSE
```