Homework 2

Tucker Morgan - tlm2152

3/8/2022

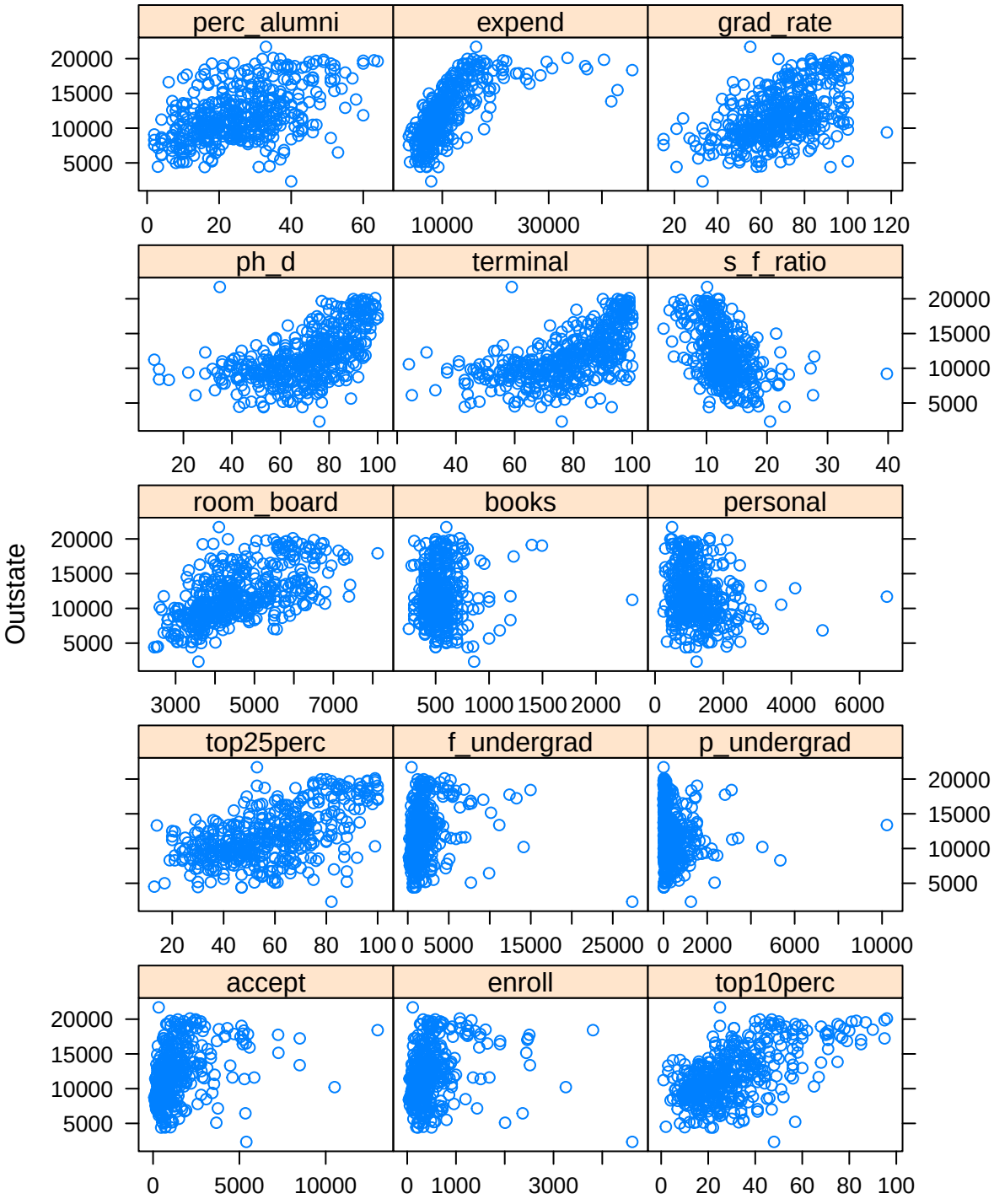
```
library(tidyverse)
library(caret)
```

First, let's import and partition our data set.

```
college_full <- read_csv("./Homework 2/College.csv") %>%
  janitor::clean_names()
# creating data partitioning index
part_index <- createDataPartition(y = college_full$outstate,
                                  p = 0.8,
                                  list = FALSE)

college_trn <- college_full[part_index, ] %>% # training data
  select(-college) # removing college name because unique identifier
college_tst <- college_full[-part_index, ] %>% # test data
  select(-college) # removing college name because unique identifier
```

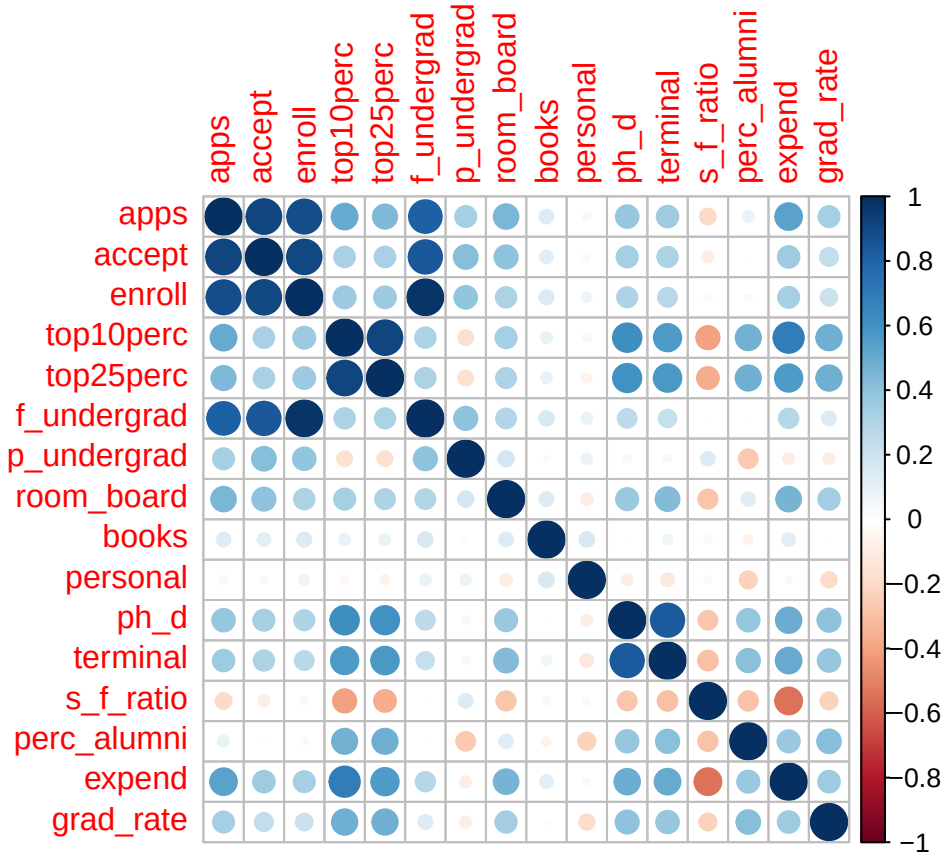
(a) And now we'll look at some exploratory analysis using the training data set.

```
# making a model matrix for caret
college_x <- model.matrix(outstate ~ ., college_trn)[,-1]
college_y <- college_trn$outstate

# setting plot parameters
theme1 <- trellis.par.get()
theme1$plot.symbol$col <- rgb(.1, .1, .6, .4)
theme1$plot.symbol$pch <- 16
theme1$plot.line$col <- rgb(.8, .1, .1, .8)
theme1$plot.line$lwd <- 2
theme1$strip.background$col <- rgb(.0, .2, .6, .2)

featurePlot(college_x[,-1], college_y,
            plot = "scatter",
            labels = c("", "Outstate"),
            type = c("p"),
            layout = c(3, 5))
```



Looking at this plot, there are varying relationships between the response, `outstate`, and the array of predictors. It seems like there could be linear relationships between out of state tuition and room and board costs (`room_board`), percentage of new students from the top 10% of their class (`top10perc`), and percentage of new students from the top 25% of their class (`top25perc`). Let's look at a a correlation plot of these predictors.

```
corrplot::corrplot(cor(college_x),
                   method = "circle",
                   type = "full")
```



It's clear from this plot that the number of applications, number of accepted students, and number of enrolled students have a high correlation with each other along with the number of full time undergraduate students, `f_undergrad`. There is also an understandably strong relationship between `top10perc` and `top25perc`. Another strong correlation is between percent of faculty with PhD's and percent of faculty with terminal degrees. This is likely because terminal degrees are often PhD's.