Assignment 2

Zheyan Liu

Contents

Intro and data preparation

we build nonlinear models using the “College” data. The dataset contains statistics for 565 US Colleges from a previous issue of US News and World Report. The response variable is the out-of-state tuition (Outstate).

Read data

Drop college column

```
df =
  read_csv('data/College.csv', show_col_types = FALSE) %>%
  janitor::clean_names() %>%
  select(-college)
```

Split the dataset into training and testing

Partition the dataset into two parts: training data (80%) and test data (20%).

```
trainRows <- createDataPartition(y = df$outstate, p = 0.8, list = FALSE)
training_df = df[trainRows, ]
testing_df = df[-trainRows, ]

x_train <- model.matrix(outstate~.,training_df)[,-1]
y_train <- training_df$outstate

x_test <- model.matrix(outstate~.,testing_df)[,-1]
y_test <- testing_df$outstate
```

(a) Perform exploratory data analysis using the training data

There are 17 variables in the data and 453 observations.

summary statistics

All variables are continuous

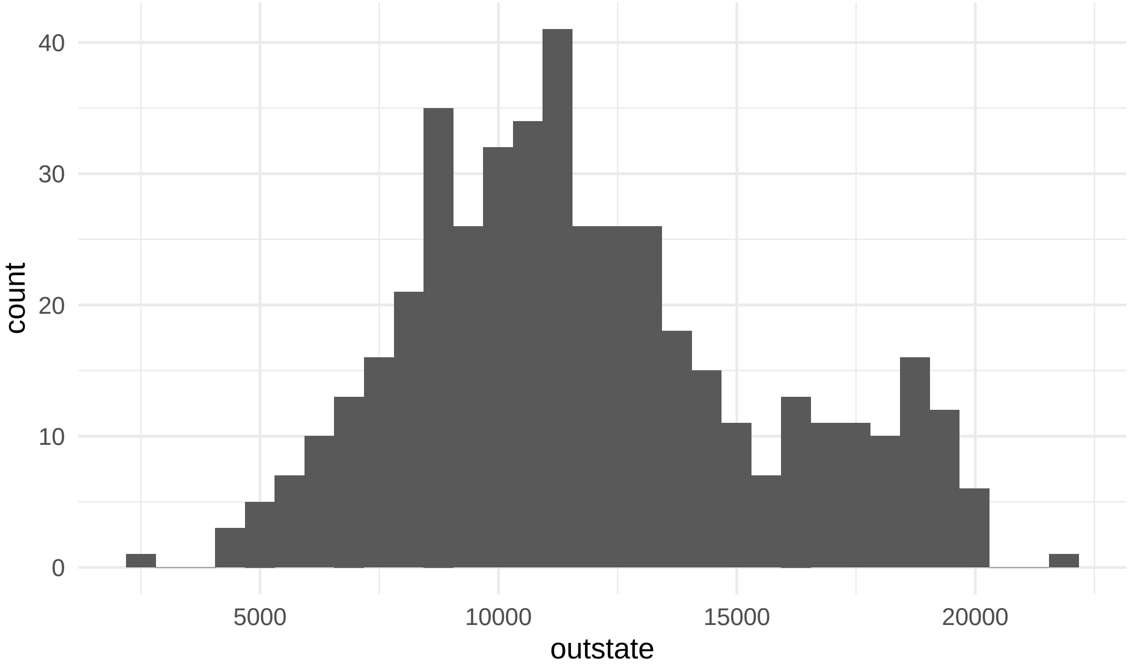
```
summary(training_df)
```

```
##       apps           accept          enroll         top10perc
##  Min.   :   81   Min.   :   72   Min.   :  35.0   Min.   : 1.00
##  1st Qu.:  619   1st Qu.:  503   1st Qu.: 212.0   1st Qu.:17.00
##  Median : 1109   Median :  858   Median : 328.0   Median :25.00
##  Mean   : 2013   Mean   : 1329   Mean   : 464.5   Mean   :29.82
##  3rd Qu.: 2212   3rd Qu.: 1580   3rd Qu.: 523.0   3rd Qu.:37.00
##  Max.   :20192   Max.   :13007   Max.   :4615.0   Max.   :96.00
##    top25perc       f_undergrad     p_undergrad         outstate
##  Min.   :  9.00   Min.   :  139   Min.   :    1.0   Min.   : 2340
##  1st Qu.: 43.00   1st Qu.:  879   1st Qu.:   61.0   1st Qu.: 9100
##  Median : 56.00   Median : 1280   Median :  209.0   Median :11200
##  Mean   : 57.65   Mean   : 1906   Mean   :  461.9   Mean   :11850
##  3rd Qu.: 71.00   3rd Qu.: 2041   3rd Qu.:  580.0   3rd Qu.:13960
##  Max.   :100.00   Max.   :27378   Max.   :10221.0   Max.   :21700
##    room_board        books          personal          ph_d           terminal
##  Min.   :2460   Min.   : 250   Min.   : 300   Min.   :  8.00   Min.   : 24.00
##  1st Qu.:3730   1st Qu.: 450   1st Qu.: 800   1st Qu.: 61.00   1st Qu.: 68.00
##  Median :4390   Median : 500   Median :1100   Median : 74.00   Median : 81.00
##  Mean   :4609   Mean   : 553   Mean   :1217   Mean   : 71.81   Mean   : 79.11
##  3rd Qu.:5420   3rd Qu.: 600   3rd Qu.:1500   3rd Qu.: 85.00   3rd Qu.: 92.00
##  Max.   :8124   Max.   :2340   Max.   :6800   Max.   :100.00   Max.   :100.00
##    s_f_ratio       perc_alumni         expend          grad_rate
##  Min.   : 2.50   Min.   : 2.00   Min.   : 3365   Min.   : 15.00
##  1st Qu.:11.10   1st Qu.:16.00   1st Qu.: 7440   1st Qu.: 58.00
##  Median :12.80   Median :25.00   Median : 9060   Median : 69.00
##  Mean   :12.96   Mean   :25.95   Mean   :10547   Mean   : 69.01
##  3rd Qu.:14.60   3rd Qu.:34.00   3rd Qu.:11711   3rd Qu.: 82.00
##  Max.   :39.80   Max.   :64.00   Max.   :56233   Max.   :118.00
```

histogram of response variable

Distribution of outstate is close to normal distribution, much outstate is around 10000 except a second peak around 17500

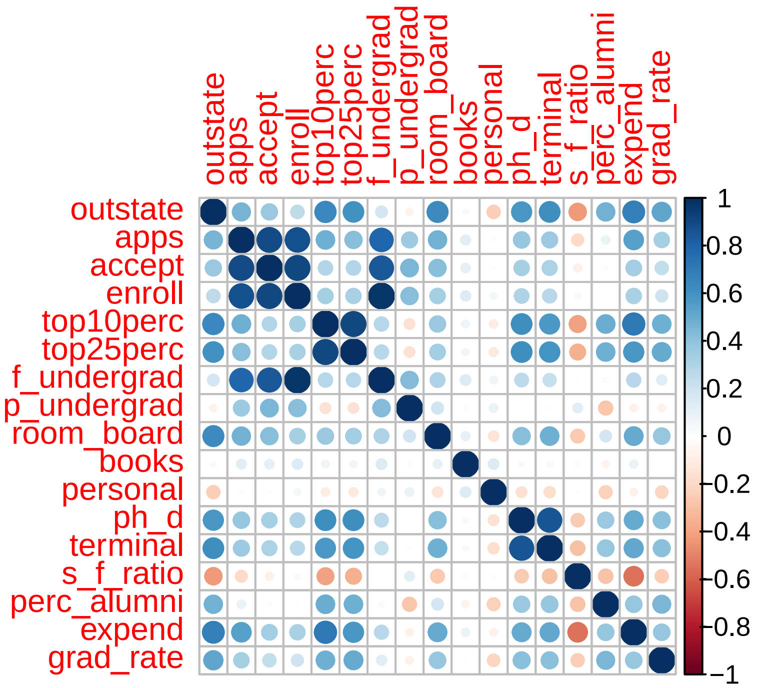
```
ggplot(training_df, aes(x=outstate)) +
  geom_histogram(bins = 32)
```



correlation of response vs. predictors

Correlation plot shows that some variables are highly correlated with outstate and there is multicollinearity.

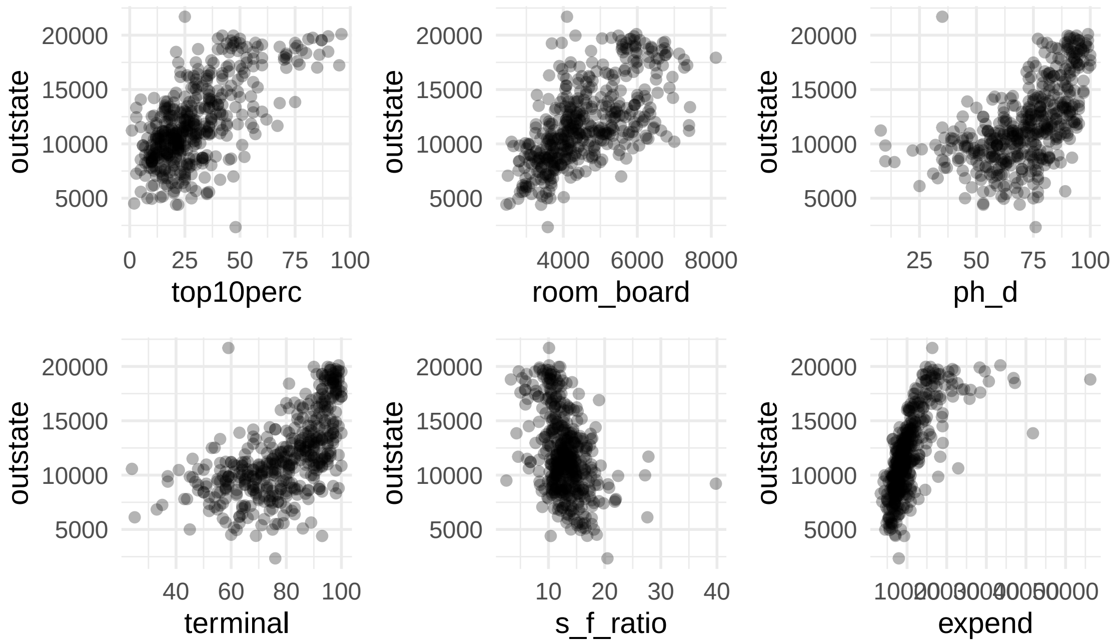
```
corrplot::corrplot(cor(training_df %>% select(where(is.numeric)) %>% relocate(outstate)), method = "cir
```



Plot scatterplot of response variables with selection of highly correlated predictors including top10perc, room_board, ph_d, terminal, s_f_ratio and expend. Only s_f_ratio is negatively correlated.

```
library(patchwork)
p1 = ggplot(training_df, aes(x=top10perc, y=outstate)) + geom_point(alpha=0.3)
p2 = ggplot(training_df, aes(x=room_board, y=outstate)) + geom_point(alpha=0.3)
p3 = ggplot(training_df, aes(x=ph_d, y=outstate)) + geom_point(alpha=0.3)
p4 = ggplot(training_df, aes(x=terminal, y=outstate)) + geom_point(alpha=0.3)
p5 = ggplot(training_df, aes(x=s_f_ratio, y=outstate)) + geom_point(alpha=0.3)
p6 = ggplot(training_df, aes(x=expend, y=outstate)) + geom_point(alpha=0.3)

(p1 + p2 + p3)/(p4 + p5 + p6)
```



(b) Fit smoothing spline models using Terminal as the only predictor of Outstate

(c) Fit a generalized additive model (GAM) using all the predictors.

(d) Train a multivariate adaptive regression spline (MARS) model using all the predictors

(e) Model selection