HW2

Min Kim

2025-02-03

```
library(here)
```

```
## here() starts at /Users/kim/University of Michigan Dropbox/Min Kim/2025 Winter Term/DATASCI 415
```

```
library(ISLR)
```

1.

a.

```
college <- read.csv(here("Data Files", "College.csv"))
```

b.

```
rownames(college) = college[,1]
fix(college)
college = college[-1]
fix(college)
```

c.

i.

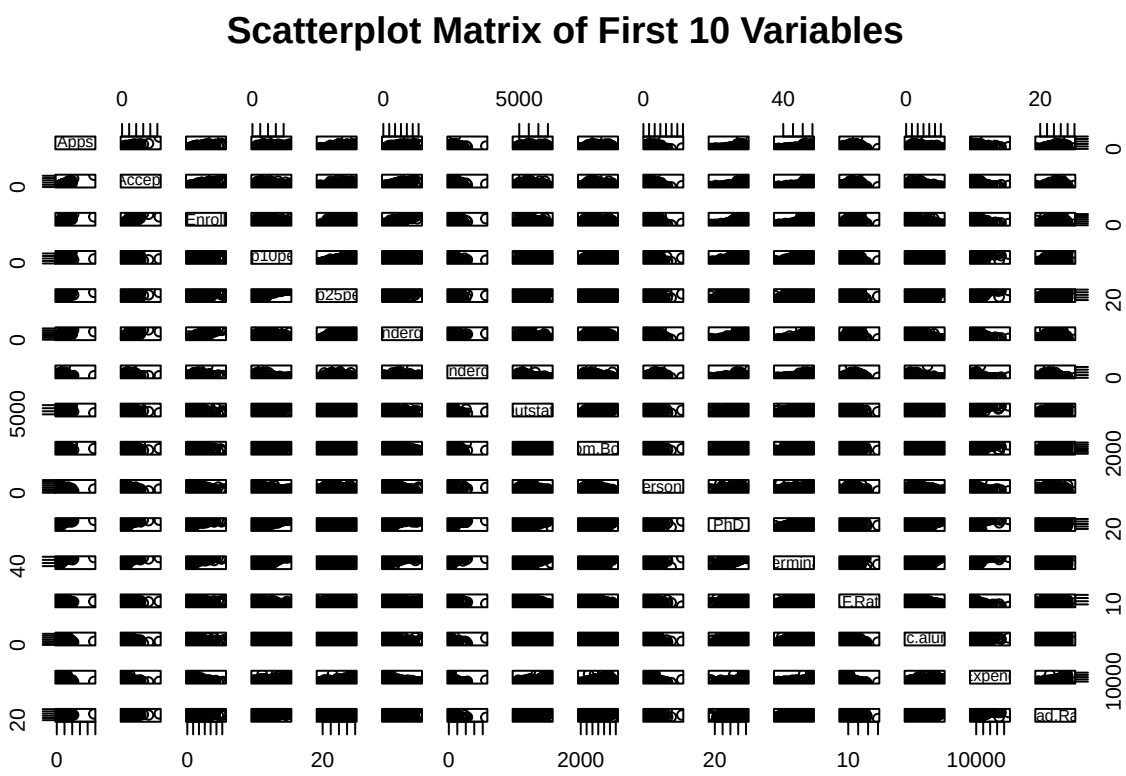
```
summary(college)
```

```
##    Private                 Apps           Accept          Enroll
##  Length:777         Min.   :   81   Min.   :   72   Min.   :  35
##  Class :character   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242
##  Mode  :character   Median : 1558   Median : 1110   Median : 434
##                     Mean   : 3002   Mean   : 2019   Mean   : 780
##                     3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902
##                     Max.   :48094   Max.   :26330   Max.   :6392
##    Top10perc       Top25perc      F.Undergrad     P.Undergrad
##  Min.   : 1.00   Min.   :  9.0   Min.   :  139   Min.   :    1.0
##  1st Qu.:15.00   1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0
##  Median :23.00   Median : 54.0   Median : 1707   Median :  353.0
##  Mean   :27.56   Mean   : 55.8   Mean   : 3700   Mean   :  855.3
##  3rd Qu.:35.00   3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0
##  Max.   :96.00   Max.   :100.0   Max.   :31643   Max.   :21836.0
##     Outstate       Room.Board       Books           Personal
##  Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   : 250
##  1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850
```

```
##  Median : 9990   Median :4200   Median : 500.0   Median :1200
##  Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   :1341
##  3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700
##  Max.   :21700   Max.   :8124   Max.   :2340.0   Max.   :6800
##       PhD            Terminal       S.F.Ratio      perc.alumni
##  Min.   :  8.00   Min.   : 24.0   Min.   : 2.50   Min.   : 0.00
##  1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00
##  Median : 75.00   Median : 82.0   Median :13.60   Median :21.00
##  Mean   : 72.66   Mean   : 79.7   Mean   :14.09   Mean   :22.74
##  3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00
##  Max.   :103.00   Max.   :100.0   Max.   :39.80   Max.   :64.00
##      Expend        Grad.Rate
##  Min.   : 3186   Min.   : 10.00
##  1st Qu.: 6751   1st Qu.: 53.00
##  Median : 8377   Median : 65.00
##  Mean   : 9660   Mean   : 65.46
##  3rd Qu.:10830   3rd Qu.: 78.00
##  Max.   :56233   Max.   :118.00
```

ii.

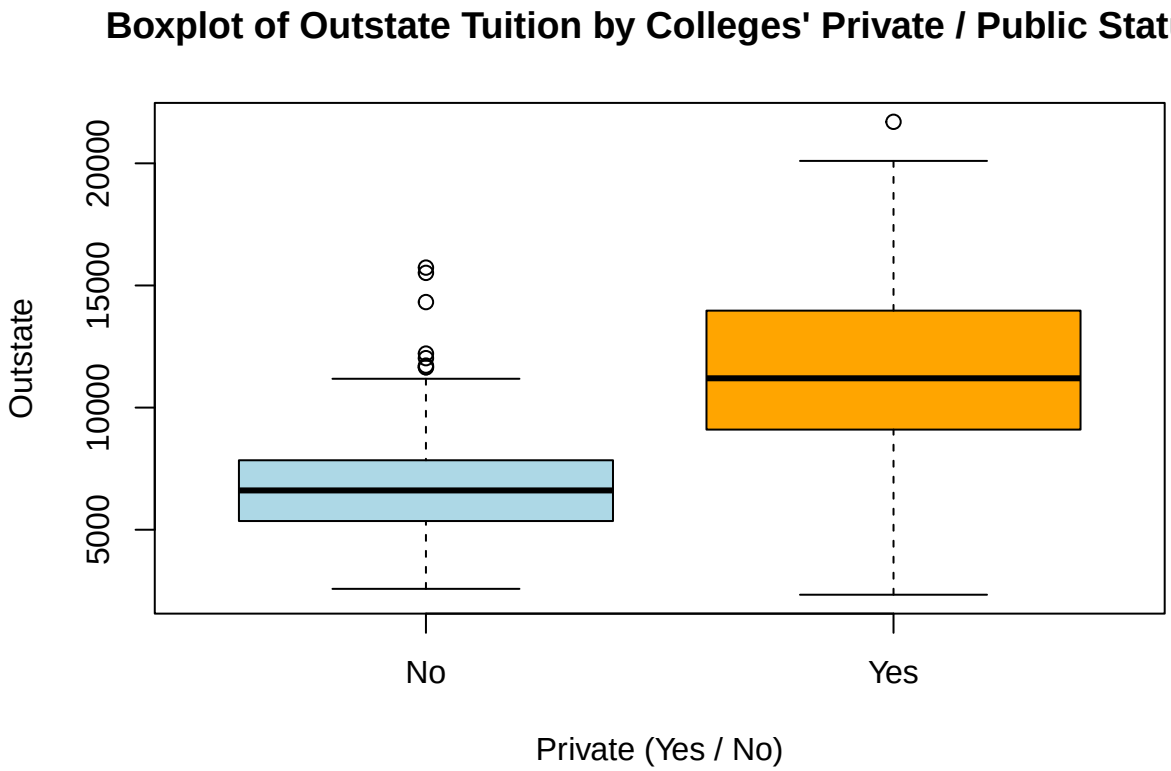
```r
college_numeric <- college[, sapply(college[, 1:10], is.numeric)]
pairs(college_numeric, main = "Scatterplot Matrix of First 10 Variables")
```



iii.

```r
college$Private <- as.factor(college$Private)
plot(Outstate ~ Private, data = college,
     main = "Boxplot of Outstate Tuition by Colleges' Private / Public Status",
```

```
    xlab = "Private (Yes / No)",
    ylabs = "Outstate Tuition",
    col = c("lightblue", "orange")
    )
```



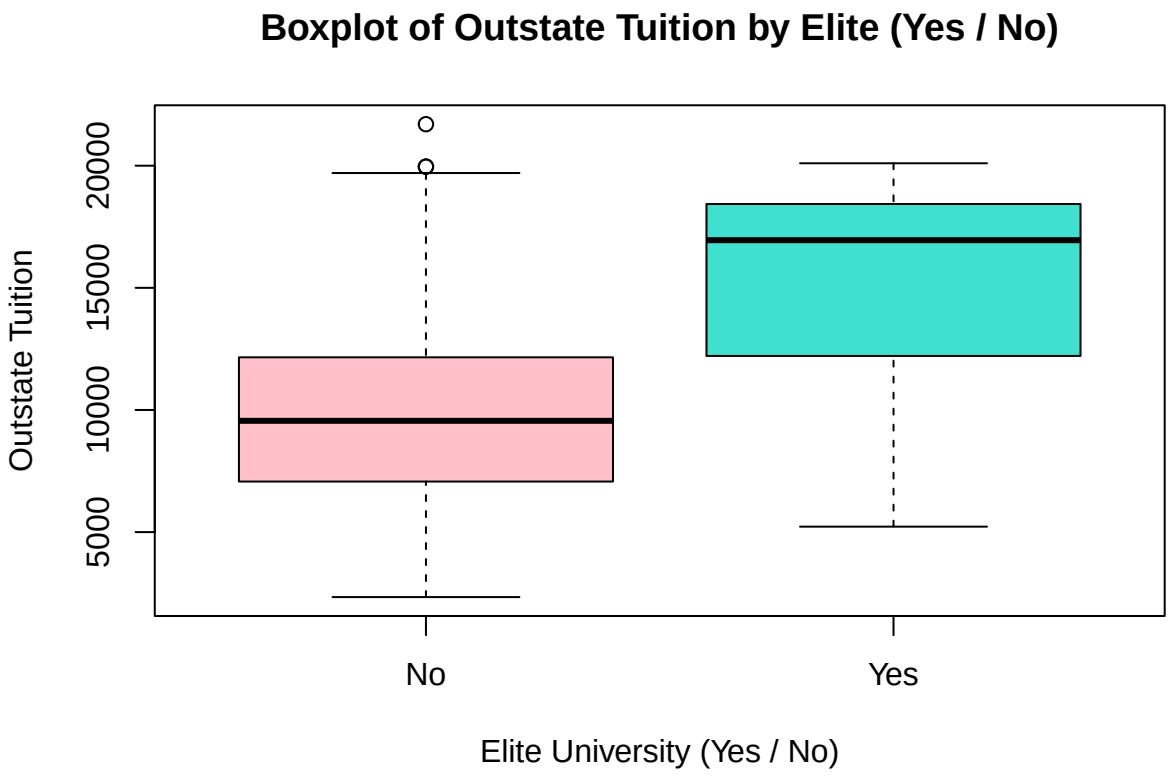
iv.

```
Elite = rep("No", nrow(college))
Elite[college$Top10perc > 50] = "Yes"
Elite = as.factor(Elite)
college = data.frame(college, Elite)

summary(college)
```

```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00
##            Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56
##            3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00
##            Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00
##    Top25perc      F.Undergrad     P.Undergrad         Outstate
##  Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340
##  1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320
##  Median : 54.0   Median : 1707   Median :  353.0   Median : 9990
##  Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441
##  3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925
##  Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700
##    Room.Board       Books           Personal         PhD
```

```
##  Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00
##  1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00
##  Median :4200   Median : 500.0   Median :1200   Median : 75.00
##  Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66
##  3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00
##  Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00
##     Terminal       S.F.Ratio      perc.alumni        Expend
##  Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186
##  1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751
##  Median : 82.0   Median :13.60   Median :21.00   Median : 8377
##  Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660
##  3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830
##  Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233
##    Grad.Rate      Elite
##  Min.   : 10.00   No :699
##  1st Qu.: 53.00   Yes: 78
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00
```

```r
plot(Outstate ~ Elite, data = college,
     main = "Boxplot of Outstate Tuition by Elite (Yes / No)",
     xlab = "Elite University (Yes / No)",
     ylab = "Outstate Tuition",
     col = c("pink", "turquoise")
     )
```

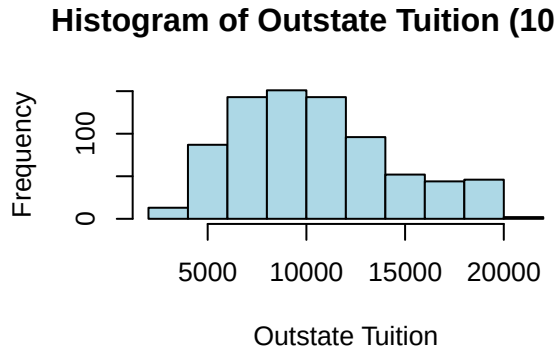


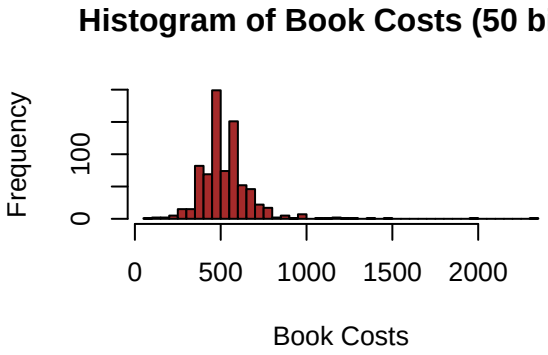
v.

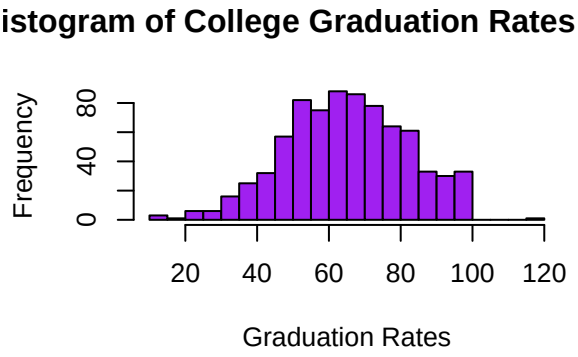
```
par(mfrow = c(2,2))
# using the str() function, identify which quantitative variables to use
str(college)
```

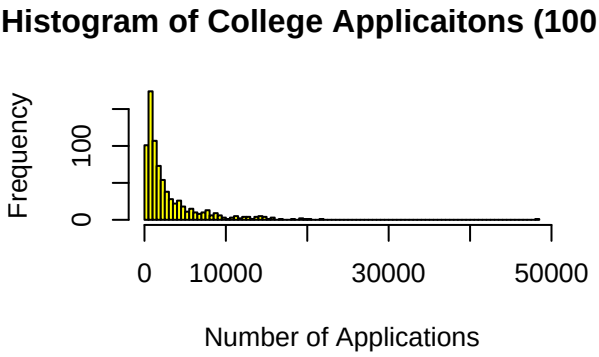
```
## 'data.frame':    777 obs. of  19 variables:
##  $ Private    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Apps       : num  1660 2186 1428 417 193 ...
##  $ Accept     : num  1232 1924 1097 349 146 ...
##  $ Enroll     : num  721 512 336 137 55 158 103 489 227 172 ...
##  $ Top10perc  : num  23 16 22 60 16 38 17 37 30 21 ...
##  $ Top25perc  : num  52 29 50 89 44 62 45 68 63 44 ...
##  $ F.Undergrad: num  2885 2683 1036 510 249 ...
##  $ P.Undergrad: num  537 1227 99 63 869 ...
##  $ Outstate   : num  7440 12280 11250 12960 7560 ...
##  $ Room.Board : num  3300 6450 3750 5450 4120 ...
##  $ Books      : num  450 750 400 450 800 500 500 450 300 660 ...
##  $ Personal   : num  2200 1500 1165 875 1500 ...
##  $ PhD        : num  70 29 53 92 76 67 90 89 79 40 ...
##  $ Terminal   : num  78 30 66 97 72 73 93 100 84 41 ...
##  $ S.F.Ratio  : num  18.1 12.2 12.9 7.7 11.9 9.4 11.5 13.7 11.3 11.5 ...
##  $ perc.alumni: num  12 16 30 37 2 11 26 37 23 15 ...
##  $ Expend     : num  7041 10527 8735 19016 10922 ...
##  $ Grad.Rate  : num  60 56 54 59 15 55 63 73 80 52 ...
##  $ Elite      : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 1 1 ...
```

```
# plot histograms
hist(college$Outstate,
     breaks = 10,
     main = "Histogram of Outstate Tuition (10 bins)",
     xlab = "Outstate Tuition",
     col = "lightblue"
     )

hist(college$Books,
     breaks = 50,
     main = "Histogram of Book Costs (50 bins)",
     xlab = "Book Costs",
     col = "brown"
     )

hist(college$Grad.Rate,
     breaks = 20,
     main = "Histogram of College Graduation Rates (20 bins)",
     xlab = "Graduation Rates",
     col = "purple"
     )

hist(college$Apps,
     breaks = 100,
     main = "Histogram of College Applicaitons (100 bins)",
     xlab = "Number of Applications",
     col = "yellow"
     )
```



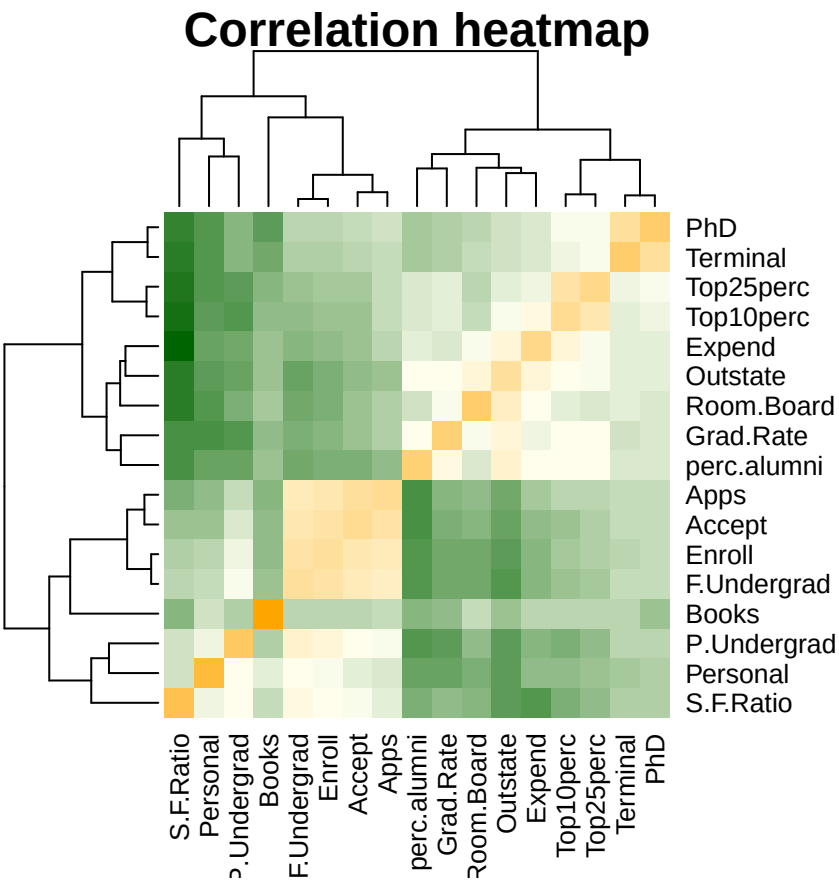






vi.

```
par(mfrow = c(1,2))
# Correlation visualization using correlation Matrix and heatmap
numeric_vars <- college[, sapply(college, is.numeric)]
cor_matrix <- cor(numeric_vars, use = "complete.obs")
heatmap(cor_matrix, main = "Correlation heatmap",
        col = colorRampPalette(c("darkgreen", "ivory", "orange"))(50)
        )
```



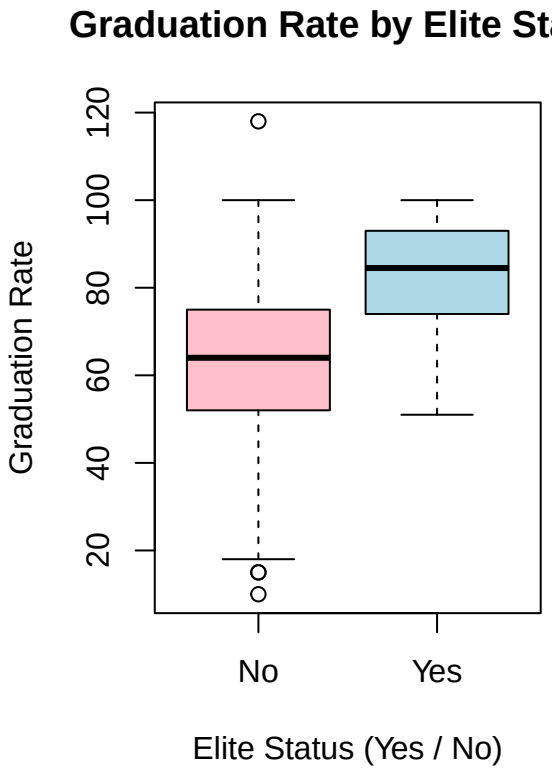

```
# Group comparisons using mean comparison and boxplots
tapply(college$Outstate, college$Private, mean)
```

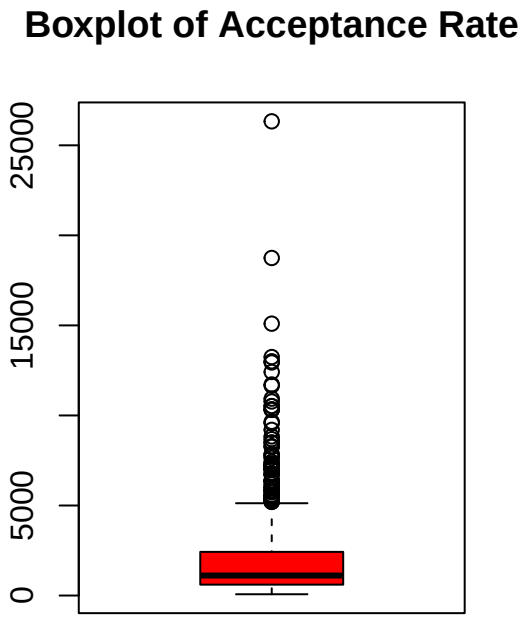
```
##       No      Yes
##  6813.41 11801.69
```

```
tapply(college$Grad.Rate, college$Elite, mean)
```

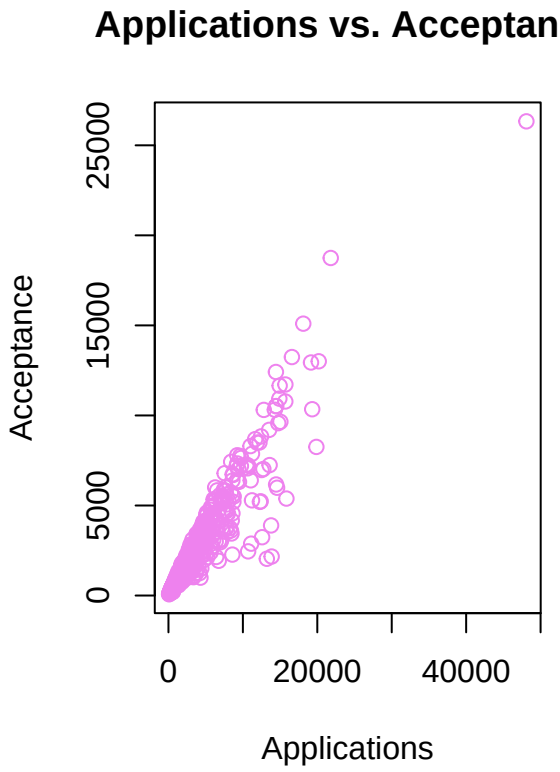
```
##       No      Yes
## 63.46352 83.38462
```

```
boxplot(Grad.Rate ~ Elite, data = college,
        main = "Graduation Rate by Elite Status",
        xlab = "Elite Status (Yes / No)",
        ylab = "Graduation Rate",
        col = c("pink", "lightblue")
        )

# Graphical and numerical analysis to identify potential outliers
boxplot(college$Accept,
        main = "Boxplot of Acceptance Rate",
        col = "red"
        )
```






```r
# Scatterplots between variables
plot(college$Apps, college$Accept,
     main = "Applications vs. Acceptance",
     xlab = "Applications",
     ylab = "Acceptance",
     col = "violet"
     )
```



2.

```
str(Carseats)
```

```
## 'data.frame':    400 obs. of  11 variables:
##  $ Sales       : num  9.5 11.22 10.06 7.4 4.15 ...
##  $ CompPrice   : num  138 111 113 117 141 124 115 136 132 132 ...
##  $ Income      : num  73 48 35 100 64 113 105 81 110 113 ...
##  $ Advertising: num  11 16 10 4 3 13 0 15 0 0 ...
##  $ Population : num  276 260 269 466 340 501 45 425 108 131 ...
##  $ Price       : num  120 83 80 97 128 72 108 120 124 124 ...
##  $ ShelveLoc   : Factor w/ 3 levels "Bad","Good","Medium": 1 2 3 3 1 1 3 2 3 3 ...
##  $ Age         : num  42 65 59 55 38 78 71 67 76 76 ...
##  $ Education   : num  17 10 12 14 13 16 15 10 10 17 ...
##  $ Urban       : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 2 2 1 1 ...
##  $ US          : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 1 2 1 2 ...
```

a.

```
mlr <- lm(Sales ~ Price + Urban + US, data = Carseats)
summary(mlr)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

b.

Note that Sales and Price are quantitative variables, whereas Urban and US are qualitative variables that take 1 as "no" and 2 as "yes". The intercept coefficient indicates the predicted Sales value when all predictor variables - Price, Urban, and US - are at reference level, hence when Price == 0, Urban == 1(no), and US == 1(no). The Price coefficient indicates that for every unit increase of the Price variable, the price company charges for car seats at each site, Sales decrease by 0.054459, holding other variables constant. The Urban coefficient indicates the difference in the Sales variable for stores located in urban areas compared to non-urban areas when other variables are held constant.

c.

```
coeff <- coef(mlr)
cat("Sales =", coeff[1], "+", coeff[2], "* Price +", coeff[3], "* UrbanYes +", coeff[4], "* USYes\n" )
```

```
## Sales = 13.04347 + -0.05445885 * Price + -0.02191615 * UrbanYes + 1.200573 * USYes
```

We can also write this as

$$Y = 13.0435 - 0.05446\beta_1 - 0.02193\beta_2 + 1.2006\beta_3$$

, where 13.0435 is

$$\beta_0$$

, the intercept,

$$\beta_1$$

as Price,

$$\beta_2$$

as Urban(Yes), and

$$\beta_3$$

as US(Yes).

d.

Based on the p-values of the predictors, we can reject the null hypothesis

$$H_0$$

for predictors Price and US(Yes).

```r
summary(mlr)$coeff
```

```
##                Estimate  Std. Error      t value     Pr(>|t|)
## (Intercept) 13.04346894 0.651012245  20.03567373 3.626602e-62
## Price       -0.05445885 0.005241855 -10.38923205 1.609917e-22
## UrbanYes    -0.02191615 0.271650277  -0.08067781 9.357389e-01
## USYes        1.20057270 0.259041508   4.63467306 4.860245e-06
```

e.

```r
new_mlr <- lm(Sales ~ Price + US, data = Carseats)
summary(new_mlr)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

f.

To compare how well the models in (a) and (e) fit the data, let us use R-squared, Adjusted R-squared, and AIC. Based on our results, the both models explain approximately 23.9% of the variance in Sales, with negligible difference. The AIC suggests that our new model in (e) fits the data more parsimonious in fitting the data compared to the model in (a).

```r
# r-squared
summary(mlr)$r.squared
```

```
## [1] 0.2392754
```

```r
summary(new_mlr)$r.squared
```

```
## [1] 0.2392629
```

```r
summary(mlr)$adj.r.squared
```

```
## [1] 0.2335123
```

```r
summary(new_mlr)$adj.r.squared
```

```
## [1] 0.2354305
```

```r
AIC(mlr)
```

```
## [1] 1865.312
```

```r
AIC(new_mlr)
```

```
## [1] 1863.319
```