HW1

STAT 5211

Emma Johnson

January 22, 2025

8: College Data

a)

```
college <- read.csv("C:/Users/emma3/Downloads/College.csv")
```

b)

```
rownames(college) <- college[, 1]
college <- college[, -1]
# View(college)
```

c)

```
# ensuring the binary variable complies as a factor:
college$Private <- as.factor(college$Private)
```

i)

```
summary(college)
```

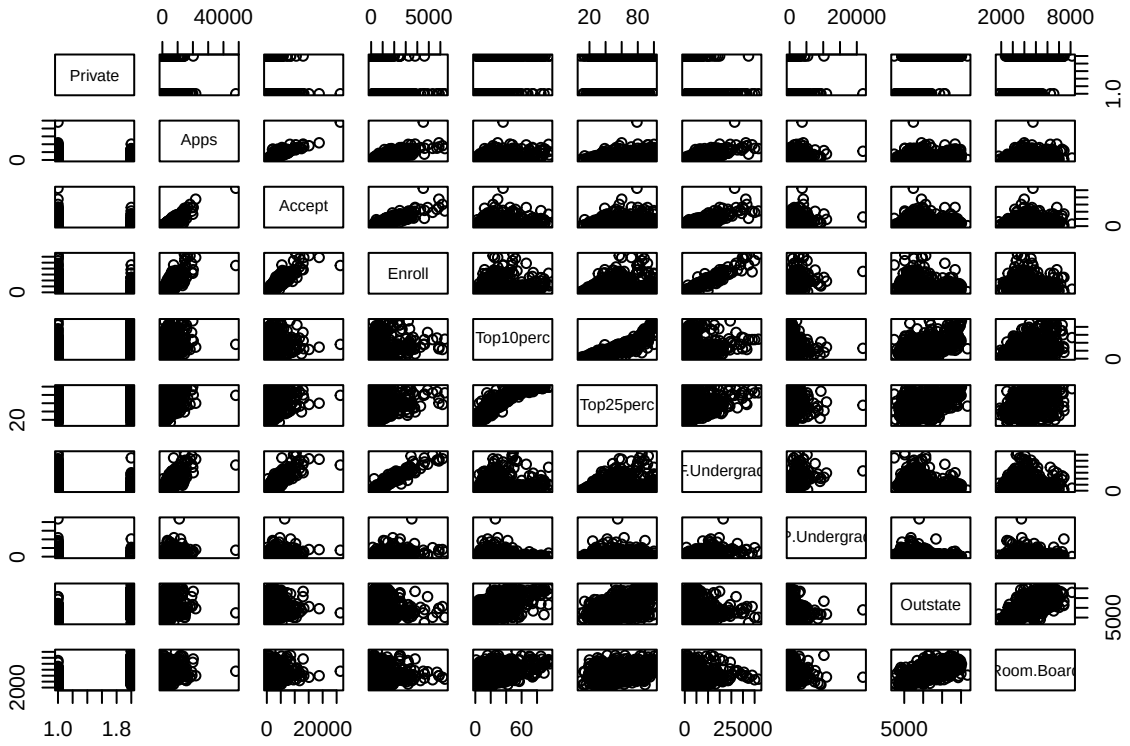
```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00
##            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
##  Min.   : 10.00
##  1st Qu.: 53.00
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00
```

ii)

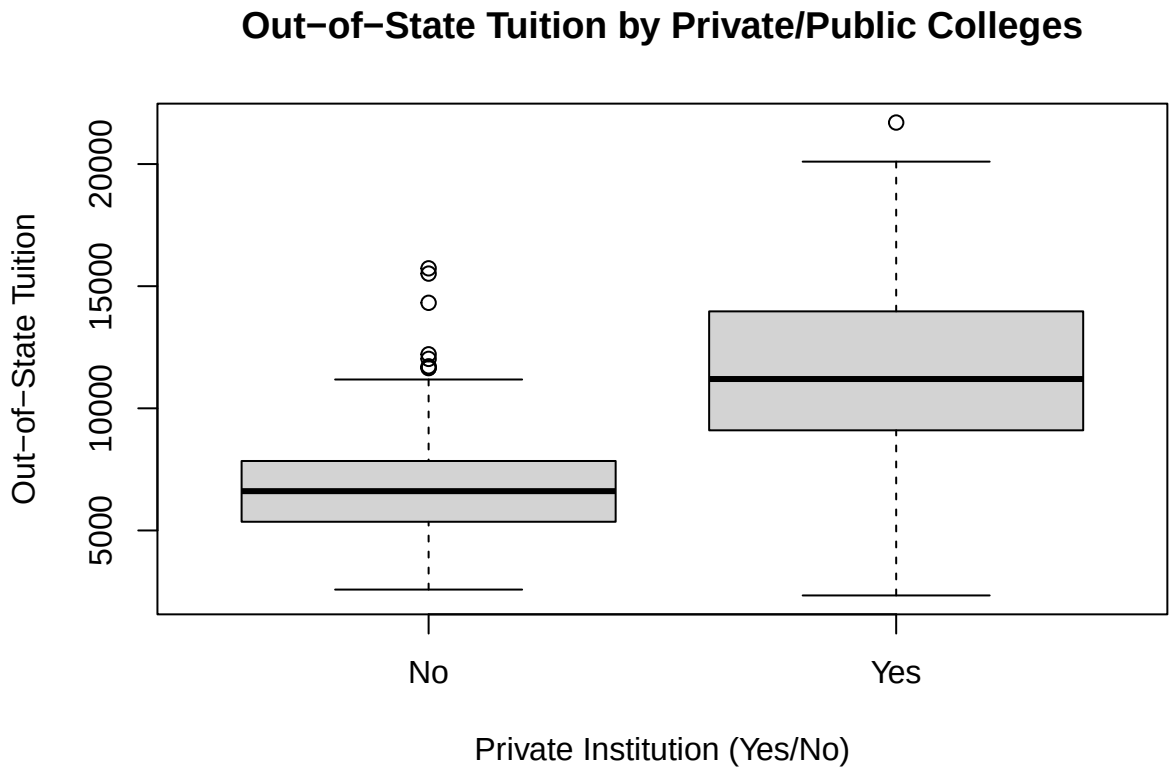
```
college$Private <- as.factor(college$Private)

pairs(college[,1:10])
```



iii)

```
plot(college$Outstate ~ college$Private,
     main = "Out-of-State Tuition by Private/Public Colleges",
     xlab = "Private Institution (Yes/No)",
     ylab = "Out-of-State Tuition")
```



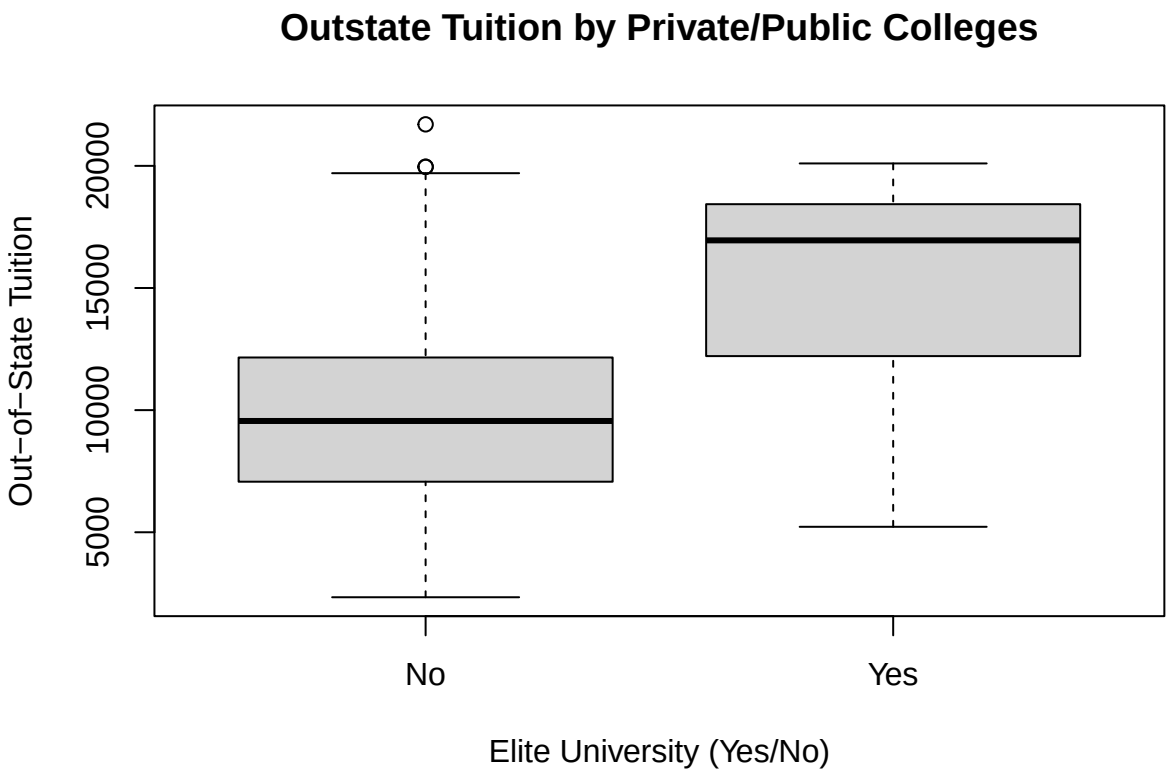
iv)

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

summary(college$Elite)
```

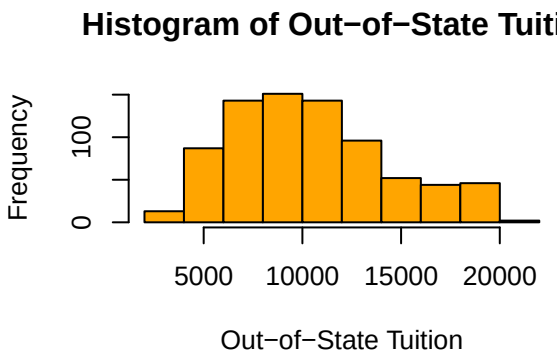
```
##  No Yes
## 699  78
```

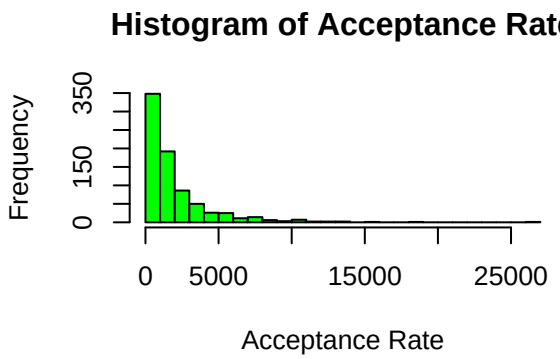
```
plot(college$Outstate ~ college$Elite,
     main = "Outstate Tuition by Private/Public Colleges",
     xlab = "Elite University (Yes/No)",
     ylab = "Out-of-State Tuition")
```

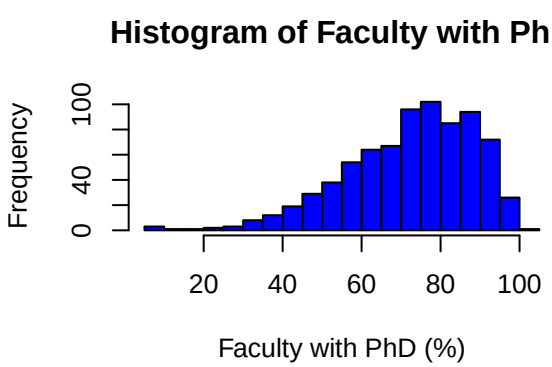


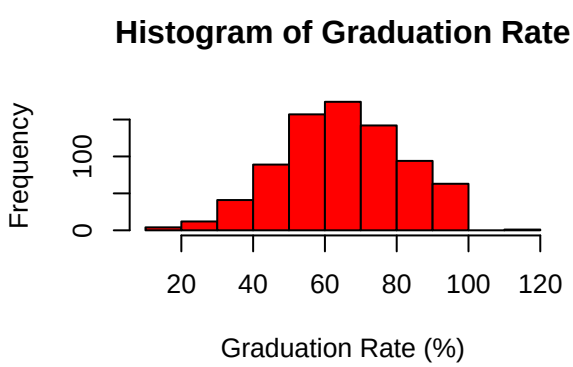
v)

```
par(mfrow = c(2, 2))
hist(college$Outstate, breaks = 8,
     main = "Histogram of Out-of-State Tuition",
     xlab = "Out-of-State Tuition",
     col = "orange")
hist(college$Accept, breaks = 30,
     main = "Histogram of Acceptance Rate",
     xlab = "Acceptance Rate",
     col = "green")
hist(college$PhD, breaks = 15,
     main = "Histogram of Faculty with PhD",
     xlab = "Faculty with PhD (%)",
     col = "blue")
hist(college$Grad.Rate, breaks = 10,
     main = "Histogram of Graduation Rate",
     xlab = "Graduation Rate (%)",
     col = "red")
```










```
library(dplyr)
```

vi)

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# creating a function to display the percentages of various responses w/in data frame:
percents <- function(data, colname){
  percentage <- data %>%
    count({{colname}}) %>%
    mutate(percentage = n / sum(n) * 100)
  return(percentage)
}

# Percentage of private colleges from all data:
percents(college, Private)
```

```
##   Private   n percentage
```

```
## 1      No 212   27.28443
## 2     Yes 565   72.71557
```

```
# Percentage of elite colleges from all data:
percents(college, Elite)
```

```
##   Elite   n percentage
## 1    No 699   89.96139
## 2   Yes  78   10.03861
```

```
# Creating a variable for colleges who are BOTH private & elite:
college$PrivElite <- case_when(college$Private == "Yes" & college$Elite == "Yes" ~ 1,
                               T ~ 0) # 1 means "Yes", 0 means "No"

percents(college, PrivElite) # only ~8.4% are both private & elite
```

```
##   PrivElite   n percentage
## 1         0 712  91.634492
## 2         1  65   8.365508
```

```
# figuring out if Out-of-State Tuition is higher among Private/Elite colleges:
summary(college$Outstate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2340    7320    9990   10441   12925   21700
```

```
med.Outstate <- summary(college$Outstate)[3] # median tuition for all colleges in data set
p75.Outstate <- summary(college$Outstate)[5] # 75th percentile for tuition
p90.Outstate <- quantile(college$Outstate, 0.9) # 90th percentile for tuition

# how many Priv/Elite colleges have a tuition over the ___ percentile?
  # True = the tuition is over the ___ percentile
# 50th:
college %>%
  filter(PrivElite == 1) %>%
  count(Outstate > med.Outstate)
```

```
##   Outstate > med.Outstate  n
## 1                   FALSE  3
## 2                    TRUE 62
```

```
# 75th:
college %>%
  filter(PrivElite == 1) %>%
  count(Outstate > p75.Outstate)
```

```
##   Outstate > p75.Outstate  n
## 1                   FALSE 11
## 2                    TRUE 54
```

```
# 90th:
college %>%
  filter(PrivElite == 1) %>%
  count(Outstate > p90.Outstate)
```

```
##   Outstate > p90.Outstate  n
## 1                   FALSE 23
## 2                    TRUE 42
```

```r
# colleges with the highest Out-of-State Tuition
college %>%
  select(PrivElite, Outstate) %>%
  arrange(desc(Outstate)) %>%
  head(10) # intresting, only half of the 10 most expensive are Priv/Elite
```

```
##                                     PrivElite Outstate
## Bennington College                          0    21700
## Massachusetts Institute of Technology       1    20100
## Gettysburg College                          0    19964
## Reed College                                0    19960
## Princeton University                        1    19900
## Yale University                             1    19840
## Amherst College                             1    19760
## Hamilton College                            0    19700
## Oberlin College                             0    19670
## Williams College                            1    19629
```

```r
# Histogram of Out-of-State Tuition for Private & Elite Colleges
PrivElite_subset <- college %>%
  filter(PrivElite == 1)
hist(PrivElite_subset$Outstate,
     main = "Distribution of Private & Elite Colleges' Out-of-State Tuition",
     xlab = "Out-of-State Tuition",
     breaks = 15)
```