Assignment8

John Bute

2024-10-01

Assignment 8

Q1: Load the dataset and set the row names

```
data <- read.csv("C:/Users/johnb/Desktop/Machine Learning/data/College.csv")
rownames(data) <- data[,1]
data <- data[,-1]
head(data, 3)
```

```
##                              Private Apps Accept Enroll Top10perc Top25perc
## Abilene Christian University     Yes 1660   1232    721        23        52
## Adelphi University               Yes 2186   1924    512        16        29
## Adrian College                   Yes 1428   1097    336        22        50
##                              F.Undergrad P.Undergrad Outstate Room.Board Books
## Abilene Christian University        2885         537     7440       3300   450
## Adelphi University                  2683        1227    12280       6450   750
## Adrian College                      1036          99    11250       3750   400
##                              Personal PhD Terminal S.F.Ratio perc.alumni Expend
## Abilene Christian University     2200  70       78      18.1          12   7041
## Adelphi University               1500  29       30      12.2          16  10527
## Adrian College                   1165  53       66      12.9          30   8735
##                              Grad.Rate
## Abilene Christian University        60
## Adelphi University                  56
## Adrian College                      54
```

Q2: Provide a nummerical summary of the data

```
summary(data)
```

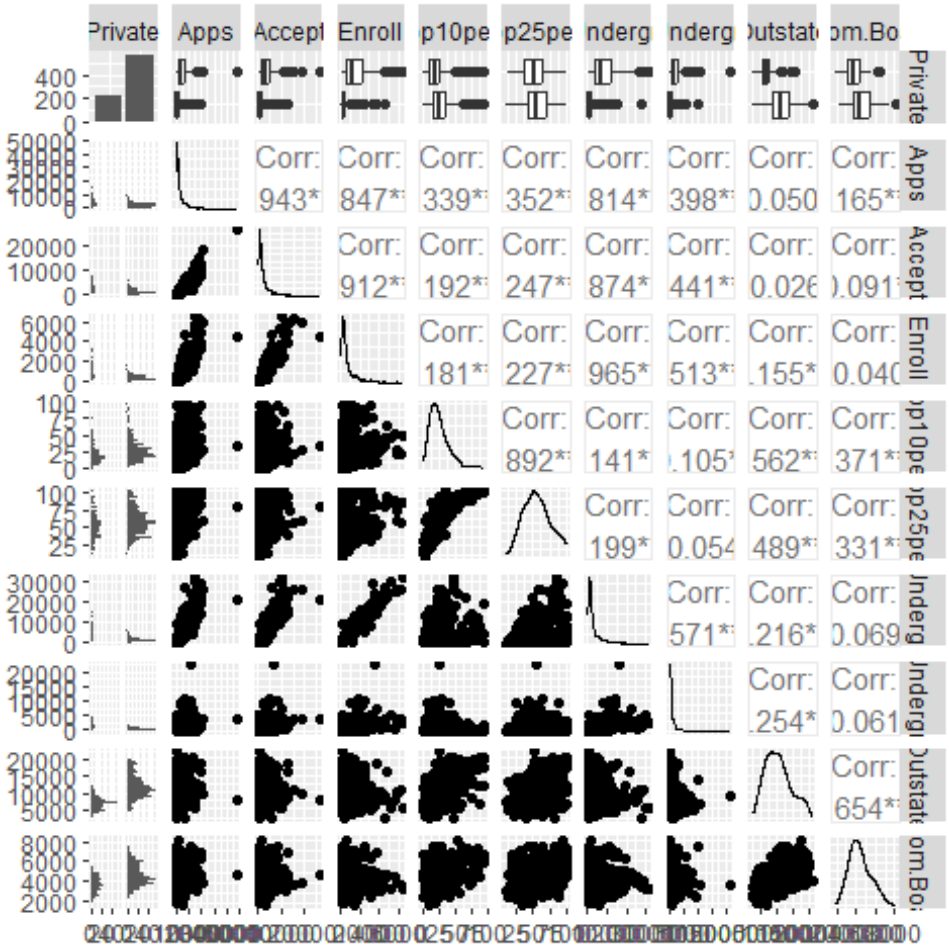
```
##    Private               Apps           Accept          Enroll
##  Length:777         Min.   :   81   Min.   :   72   Min.   :  35
##  Class :character   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242
##  Mode  :character   Median : 1558   Median : 1110   Median : 434
##                     Mean   : 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
```

From first glance, we can see that the maximum gradution rate value is 118%. There are some other important observations to look at but we will do so later in the report.

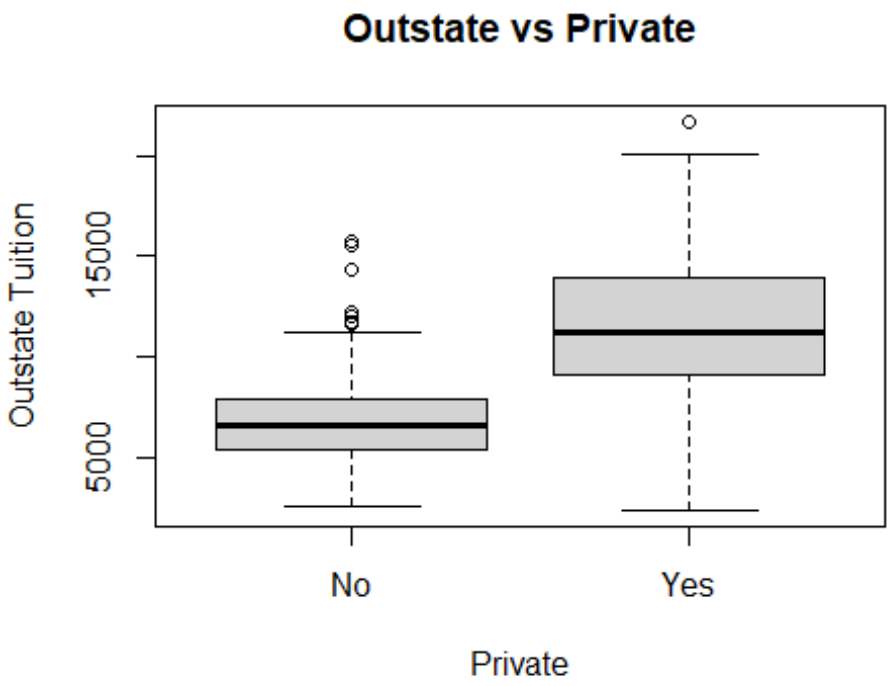
Q3: Produce a scatterplot matrix of the first 10 columns in the data

```
install.packages("GGally")

## package 'GGally' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\johnb\AppData\Local\Temp\Rtmp0qXBPC\downloaded_packages

library(GGally)
ggpairs(data[, 1:10])
```



Q4: Produce side-by-side boxplots of Outstate vs Private

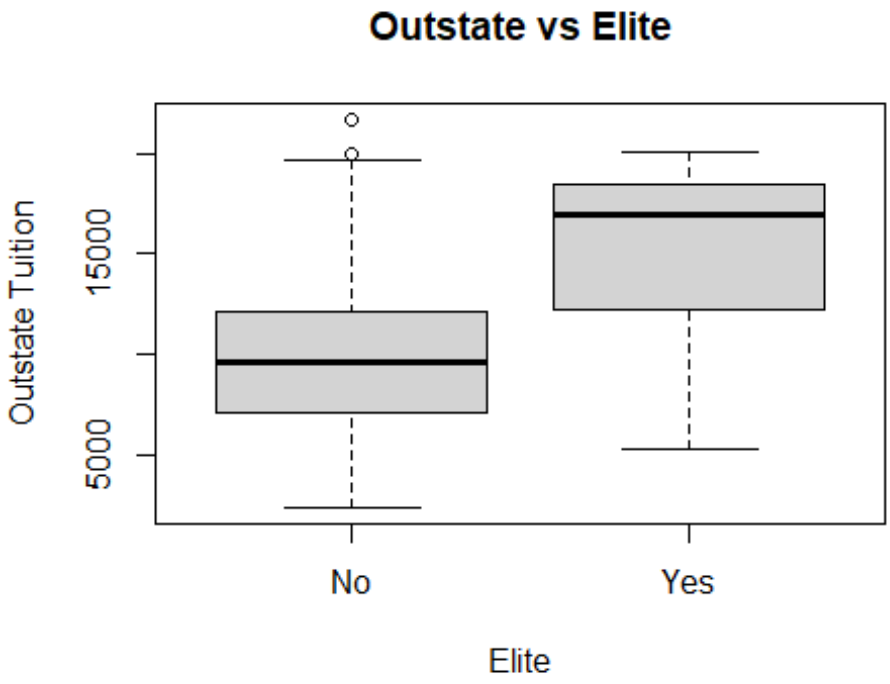
```
boxplot(Outstate ~ Private, data = data, main="Outstate vs Private",
xlab="Private", ylab="Outstate Tuition")
```



Based on the boxplots, the out-of-state tuition costs are generally higher for private schools than for public schools. This is not surprising as in general private schools tend to charge more

Q5: Create a new categorical variable, Elite, by binning the Top10perc variable. This variable divides universities into two groups: those for which Top10perc > 50 ("Yes"), and those for which that is not the case ("No"). How many elite universities are there? Produce side-by-side boxplots of Outstate versus Elite.

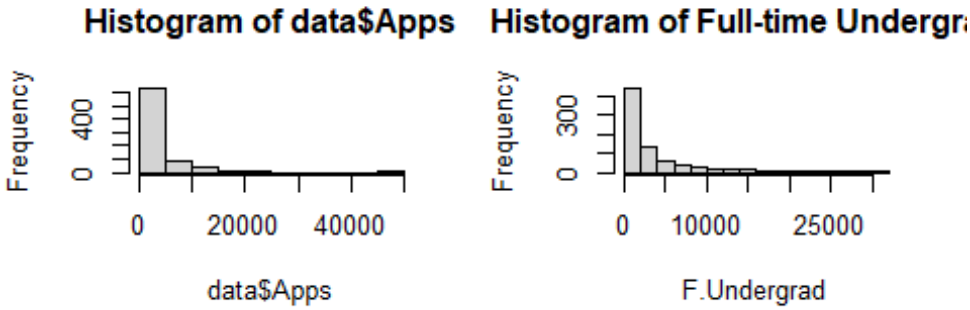
```
data$Elite <- ifelse(data$Top10perc > 50, "Yes", "No")
table(data$Elite)

##
##  No Yes
## 699  78

boxplot(Outstate ~ Elite, data = data, main="Outstate vs Elite", xlab="Elite", ylab="Outstate Tuition")
```

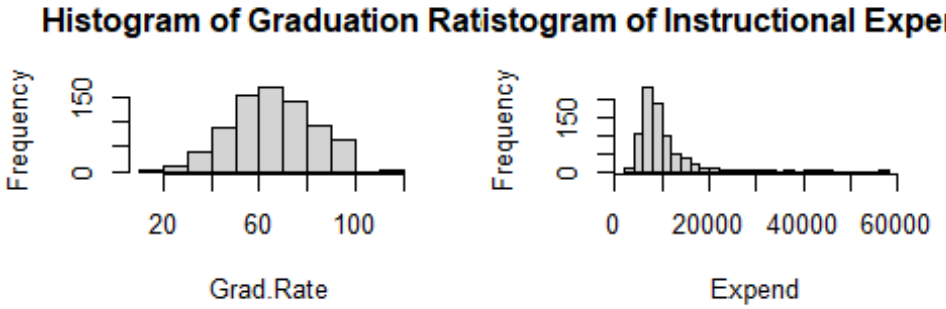


Some outliers for non-elite schools that charge as much and more as Elite schools for out-of-state tuition (perhaps reputation), but overall, there is a clear-cut minimum between elite and non-elite schools when it comes to out-of-state tuition

Q6: Produce histograms with differing numbers of bins for a few of the quantitative variables

```
par(mfrow = c(2,2))
hist(data$Apps)
hist(data$F.Undergrad, breaks=20, main="Histogram of Full-time Undergrads", xlab="F.Undergrad")
hist(data$Grad.Rate, breaks=15, main="Histogram of Graduation Rate", xlab="Grad.Rate")
hist(data$Expend, breaks=30, main="Histogram of Instructional Expenditure", xlab="Expend")
```



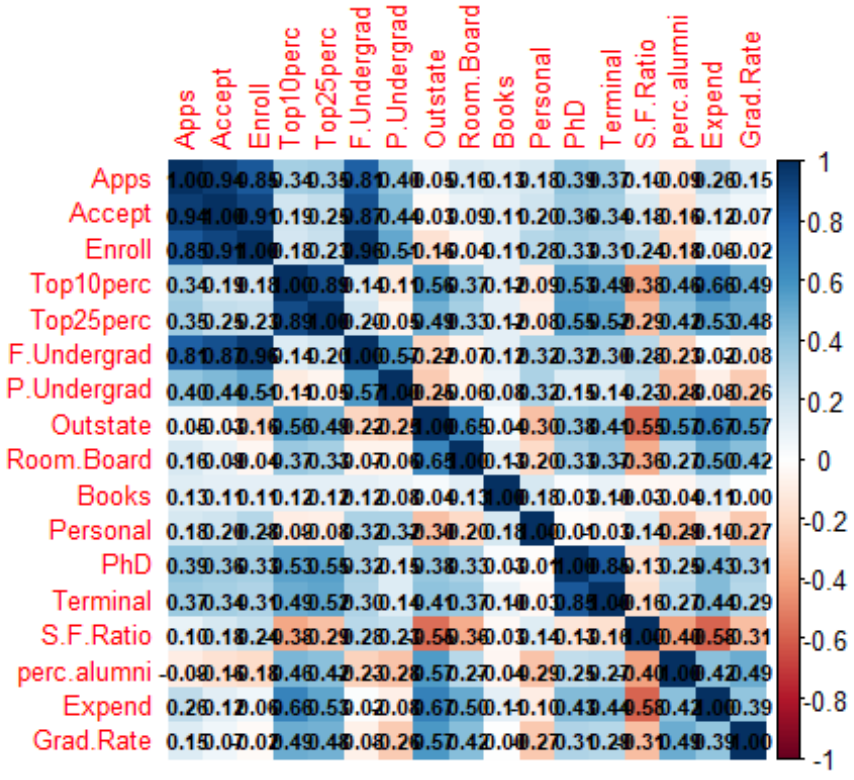


Q7: Continue exploring the data, and provide a brief summary of what you discove

```
install.packages("corrplot")
```

```
## package 'corrplot' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\johnb\AppData\Local\Temp\Rtmp0qXBPC\downloaded_packages
```

```
library(corrplot)
numeric_data <- data[, sapply(data, is.numeric)]
cor_matrix <- cor(numeric_data)
corrplot(cor_matrix, method = "color", tl.cex = 0.8, addCoef.col = "black", number.cex = 0.7)
```



Strong Positive Correlations:

Colleges that receive more applications tend to accept more students (Apps and Accept: 0.94). Institutions that accept more students tend to enroll more (Accept and Enroll: 0.91). Schools with a high percentage of top 10% high school students also have a large percentage of top 25% students (Top10perc and Top25perc: 0.89). Colleges with higher instructional expenditure per student tend to charge higher out-of-state tuition fees (Expend and Outstate: 0.67). Moderate Positive Correlations:

Schools with more faculty holding PhDs also have more faculty with terminal degrees (PhD and Terminal: 0.85). Higher instructional spending per student is moderately linked with higher graduation rates (Expend and Grad.Rate: 0.39). A greater percentage of top 10% high school graduates is associated with better graduation rates (Top10perc and Grad.Rate: 0.49). Weak or No Correlations:

The cost of books has little to no relationship with other variables in the dataset. Negative Correlations:

Schools with lower student-faculty ratios tend to spend more per student (S.F.Ratio and Expend: -0.58). More students per faculty member generally results in lower full-time undergrad student populations (S.F Ratio vs F.Undergrad is -0.41) and lower PHD-holding faculty members (PHD vs S.F.Ratio of -0.47) Schools with more out-of-state students tend to have Schools with more part-time undergraduates tend to have fewer PhD-holding faculty (P.Undergrad and PhD: -0.30).