Q8

Rahul Atre

2023-10-02

Q8 [Data Visualization and Data Exploration]

The file College.csv contains 18 measurements for 777 different universities and schools in the US. Load the data into a R data frame, and find a way to remove the Names column and place its values into the data frame's row names.

We can load the data into an R data frame (df) using the following command:

```
library(readr)
collegeData = read.csv("College.csv")
```

To remove the Names column and place the values into the data frame's row names, we can apply the procedure below:

```
rownames(collegeData) <- collegeData[,1] #Rename rownames
collegeData <- collegeData[,-1] #Delete Names column
```

1. Produce a numerical summary of the variables in the data set.

```
summary(collegeData)
```

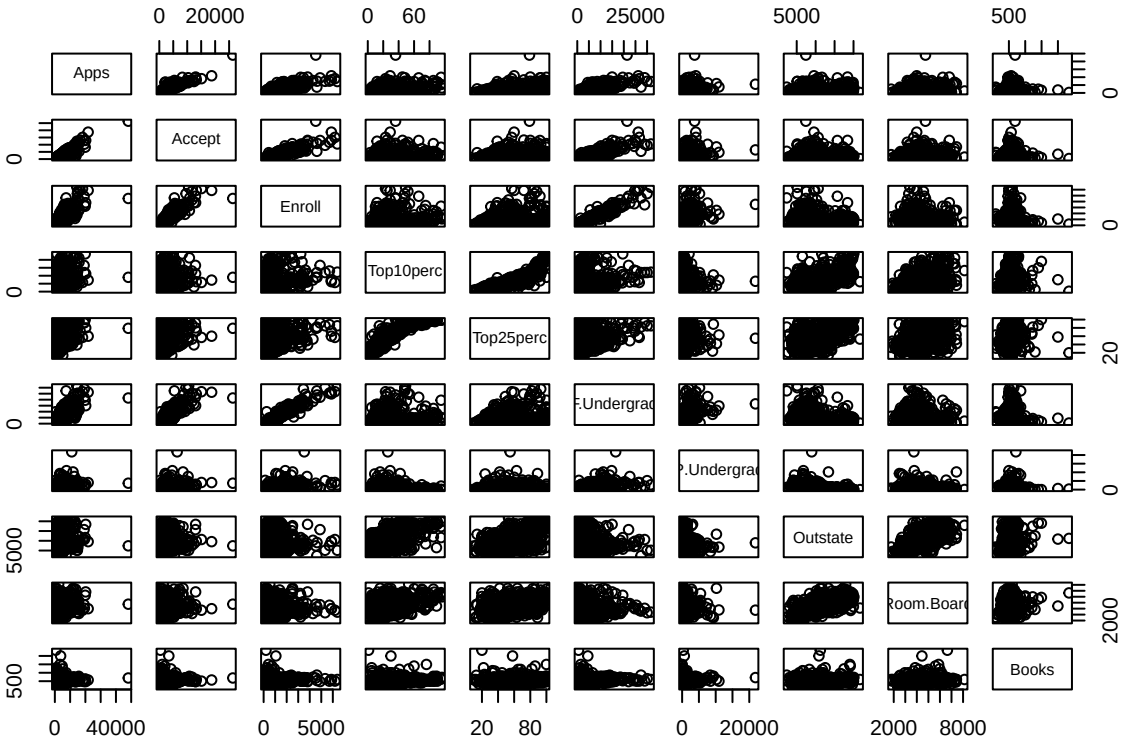
```
##    Private                Apps           Accept          Enroll
##  Length:777         Min.   :   81   Min.   :   72   Min.   :  35
##  Class :character   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242
##  Mode  :character   Median : 1558   Median : 1110   Median : 434
##                     Mean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

2. Produce a scatterplot matrix of the first ten columns in the data.

In R, we can use the function *pairs*() to create a scatterplot matrix. Since the first index is a categorical column, we will skip that and start from the 2nd, till the 11th quantitative column.

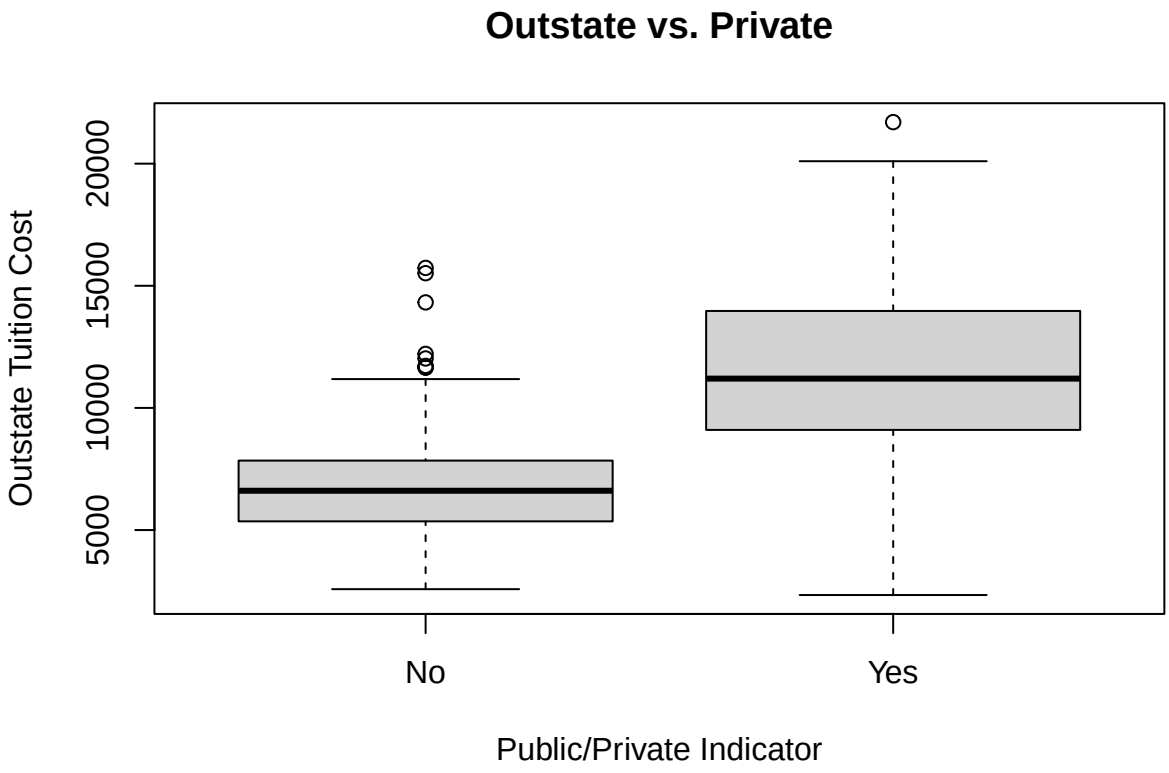
```
pairs(collegeData[, 2:11]) #Create scatterplot matrix with first 10 quantitative columns
```



3. Produce side-by-side boxplots of **Outstate** vs. **Private**.

In R, we can use the function *boxplot*() to create a boxplot comparison between the two given columns

```
boxplot(collegeData$Outstate ~ collegeData$Private, main = "Outstate vs. Private",
        ylab = "Outstate Tuition Cost", xlab = "Public/Private Indicator")
```



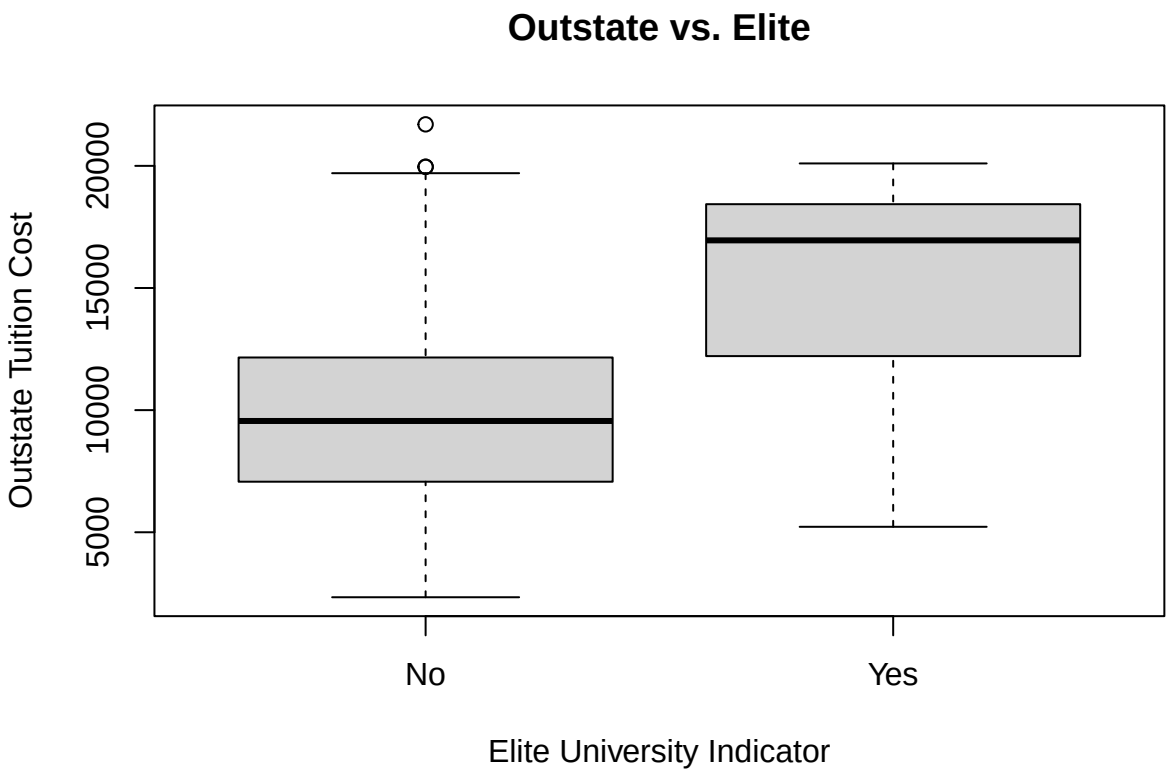
4. Create a new categorical variable, **Elite**, by binning the **Top10perc** variable. This variable divides universities into two groups: those for which **Top10perc** > 50 ("Yes"), and those for which that is not the case ("No"). How many elite universities are there? Produce side-by-side boxplots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

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

print(paste("The total number of elite universities are: ", table(collegeData$Elite)["Yes"]))
```

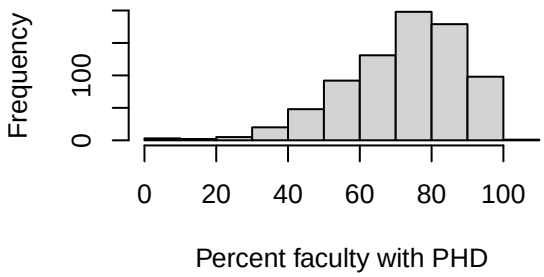
```
## [1] "The total number of elite universities are:  78"
```

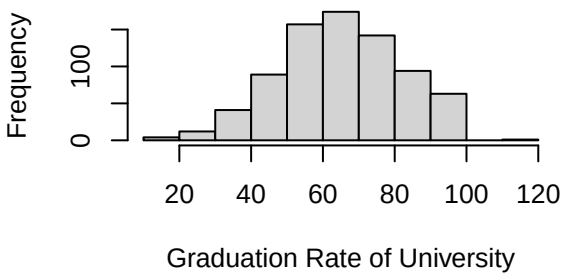
```
boxplot(collegeData$Outstate ~ collegeData$Elite, main = "Outstate vs. Elite",
        ylab = "Outstate Tuition Cost", xlab = "Elite University Indicator")
```

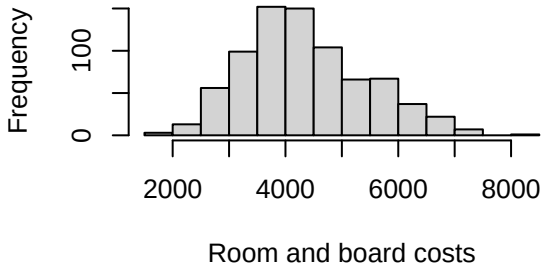


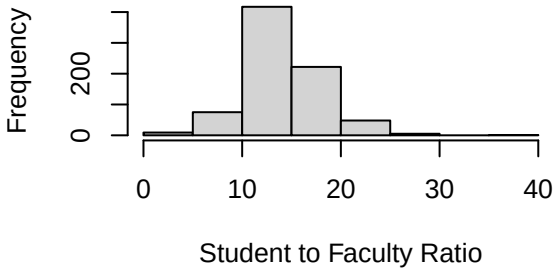
5. Produce histograms with differing numbers of bins for a few of the quantitative variables.

```
par(mfrow=c(2,2))
hist(collegeData$PhD, xlab = "Percent faculty with PHD", main = "")
hist(collegeData$Grad.Rate, xlab = "Graduation Rate of University", main = "")
hist(collegeData$Room.Board, xlab = "Room and board costs", main = "")
hist(collegeData$S.F.Ratio, xlab = "Student to Faculty Ratio", main = "")
```









6. Continue exploring the data, and provide a brief summary of what you discover.

College with the highest new students from top 10% of their high school class

```
row.names(collegeData)[which.max(collegeData$Top10perc)]
```

```
## [1] "Massachusetts Institute of Technology"
```

College with highest graduation rate

```
summary(collegeData$Grad.Rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   53.00   65.00   65.46   78.00  118.00
```

```
row.names(collegeData[collegeData$Grad.Rate>100, ])
```

```
## [1] "Cazenovia College"
```

```
collegeData['Cazenovia College', 'Grad.Rate']
```

```
## [1] 118
```

College with highest percentage of faculty with PhD's

```
summary(collegeData$PhD)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   62.00   75.00   72.66   85.00  103.00
```

```
row.names(collegeData[collegeData$PhD>100, ])
```

```
## [1] "Texas A&M University at Galveston"
```

```
collegeData['Texas A&M University at Galveston', 'PhD']
```

```
## [1] 103
```

Interestingly, the values for highest graduation rate and highest PhD percentage show erroneous data, that is, incorrect and invalid information in the data set. It is not possible for a college to have a grad. rate or PhD rate higher than 100%.

We can conclude that this data set is unclean and needs to be tidied before performing statistical analysis on it or use it as training data for a machine learning model.