M1 Individual Coding Assignment

Leonard Genders

2025-01-11

### Part A:

**Use the read.csv() function to read the data into R. Call the loaded data College. Note: I am setting stringsAsFactors=T arg to address ‘Private’ variable being char type for Part D since the dtype is character and not as factor type at loading.**

*Citation from textbook: James et al (2023, p. 49), “The stringsAsFactors = T argument tells R that any variable containing character strings should be interpreted as a qualitative variable, and that each distinct character string represents a distinct level for that qualitative variable.”*

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). An introduction to statistical learning with applications in R. (2nd ed.).\*

College <- read.csv('College.csv', stringsAsFactors = T)

**Interpretation:** We create an object called College and use the read.csv() function to read in the provided file from our current working directory with the argument stringsAsFactors set to ‘T’ for True.

### Part B:

**Look at the data using the View() function. Use rownames() to name the first row of the data as row.names. This rownames should be the names of the colleges.**

View(College) # look at the data  
  
rownames(College) <- College[, 1]  
  
# check that row names are names of the colleges  
View(College)  
  
# remove the first col in the data where the names are still stored  
College <- College[, -1]  
  
# check that col removed  
View(College)

**Interpretation:** Using the View() function allows us to see the College object without outputting into our report. We then use the rownames() function on College and output into College with a slice of all but the first columns by using [,1]. This is to drop the first column which was the college names and we lastly use the View() function again to confirm the changes from Part B.

### Part C:

**Use the summary() function to produce a numerical summary of the variables in the data set.**

summary(College)

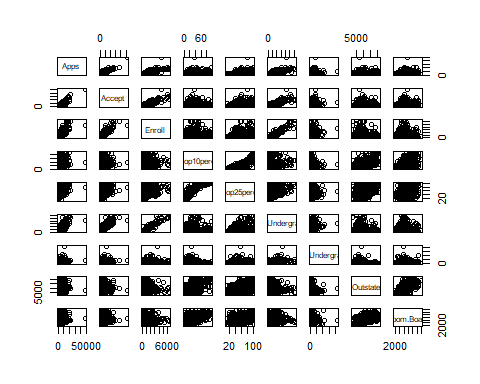
## Private Apps Accept Enroll Top10perc   
## No :212 Min. : 81 Min. : 72 Min. : 35 Min. : 1.00   
## Yes:565 1st Qu.: 776 1st Qu.: 604 1st Qu.: 242 1st Qu.:15.00   
## Median : 1558 Median : 1110 Median : 434 Median :23.00   
## Mean : 3002 Mean : 2019 Mean : 780 Mean :27.56   
## 3rd Qu.: 3624 3rd 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

**Interpretation:** Using the summary() function on College reveals that there are 565 private schools and 212 public schools in the College object for a total of 777 observations. Another highlight is that the mean percent of faculty with PhD’s is 72.66% and the average student to faculty ratio is 14.09. For the College object, we also learn that students spend on average $1,341 in personal spending. We can extract additional insights depending on the query requirement.

### Part D:

**Use the pairs() function to produce a scatterplot matrix of columns 2-10 or variables of the data.**

pairs(College[,2:10])

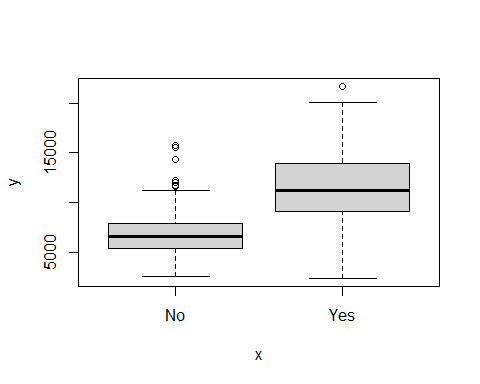


**Interpretation:** The pairs() function is used on the College object with a slicing operator to select the 2-10 columns to be used.

### Part E:

**Use the plot() or boxplot() function to produce side-by-side boxplots of Outstate versus Private.**

plot(College$Private, College$Outstate)



**Interpretation:** Two box plots are created side-by-side as required by the prompt for Part E. We can see that there are no lower outliers but the ‘No’ boxplot has a much smaller lower and upper quartiles between 5,000 and approximately 8,000 respectively. The ‘No’ whiskers extend from approximately 1,000 for the lower and 12,000 for the upper and have outliers up to above 15,000. The ‘Yes’ boxplot has larger lower and upper quartiles ranging from approximately 9,000 to 15,000. The ‘Yes’ whiskers extend approximately from 1,000 to 10,000 for the lower and 15,000 to 20,000 for the upper. There is a single upper outlier depicted at approximately 22,000. The median for the ‘No’ plot is around 6,000 and the median for the ‘Yes’ plot is approximately 12,000.

### Part F:

**Create a new qualitative variable, called Elite, by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.**

Elite <- rep("No", nrow(College)) # replicate "No" for each row in College  
Elite[College$Top10perc > 50] <- "Yes" # for Top10perc > 50, input "Yes"  
Elite <- as.factor(Elite) # change dtype to factor (Yes/No)  
College <- data.frame(College, Elite) # df adding Elite col to College

**Interpretation is commented for this R chunk**

**Use the summary() function to see how many elite universities there are.**

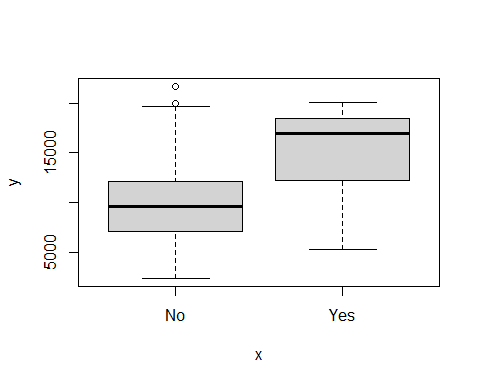
summary(College)

## 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

**Interpretation:** The summary() function for Part F creates new insights to the College data in that there are 78 colleges that fit the Elite criteria of more than 50% of the population of students coming from the 10%% of their high school classes.

**Now, use the plot() function to produce side-by-side boxplots of Outstate versus Elite.**

plot(College$Elite, College$Outstate)

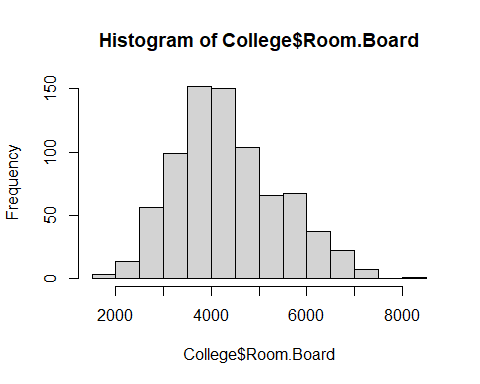


**Interpretation:** Two box plots are created side-by-side as required by the prompt for Part F. We can see that there are no lower outliers but the ‘No’ boxplot has a much smaller lower and upper quartiles between 7,500 and approximately 12,000 respectively. This is much larger than the boxplots from Part E. The ‘No’ whiskers extend from approximately 1,000 for the lower and 20,000 for the upper and have two outliers up to above 20,000. The ‘Yes’ boxplot has larger lower and upper quartiles ranging from approximately 12,000 to 19,500. The ‘Yes’ whiskers extend approximately from 6,000 to 13,000 for the lower and 18,000 to 20,000 for the upper. There are no upper outliers. The median for the ‘No’ plot is around 9,500 and the median for the ‘Yes’ plot is approximately 17,500.

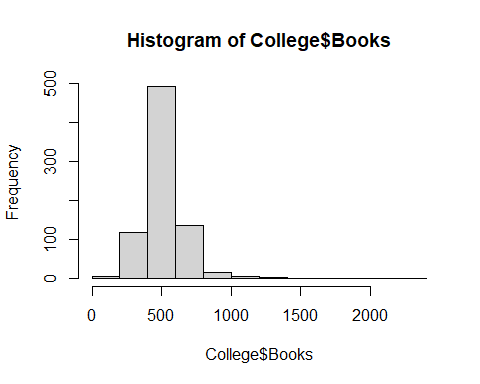
### Part G:

**Use the hist() function to produce histograms for a few of the quantitative variables: Room.Board, Books, Personal**

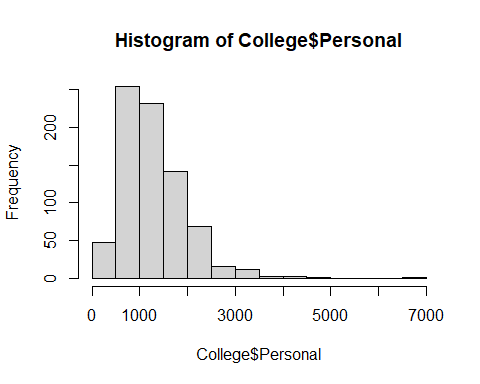
hist(College$Room.Board)



hist(College$Books)



hist(College$Personal)



**Interpretation:** The histogram of the Room.Board values show the highest frequency being the bin of values from approximately 3,500 to 4,000 dollars for 150 schools. The distribution appears to be right-skewed.

The histogram of the Books values shows the highest frequency being the bin of values for approximately 500 dollars for 500 schools. The distribution appears to be possibly slightly right-skewed.

The histogram of the Personal values show the highest frequency being the bin of values from 500 to 1,000 dollars for 250 schools.