Chapter 2 - Statistical Learning

Thalles Quinaglia Liduares

01/03/2022

Applied Exercise 2.1

Upload packages

```
library(dplyr)
```

This exercise relates to the College data set, which can be found in the file College.csv. It contains a number of variables for 777 different universities and colleges in the US.Before reading the data into R, it can be viewed in Excel or a text editor.

(a) Use the `read.csv()` function to read the data into R. Call the loaded data college. Make sure that you have the directory set to the correct location for the data.

```
setwd("C:\\Program Files\\R\\Machine Learning")

data<-read.csv("data.csv")
```

(b) Look at the data using the fix() function. You should notice that the first column is just the name of each university. We don't really want R to treat this as data. However, it may be handy to have these names for later. Try the following commands:

```
data<-data[,-1]
```

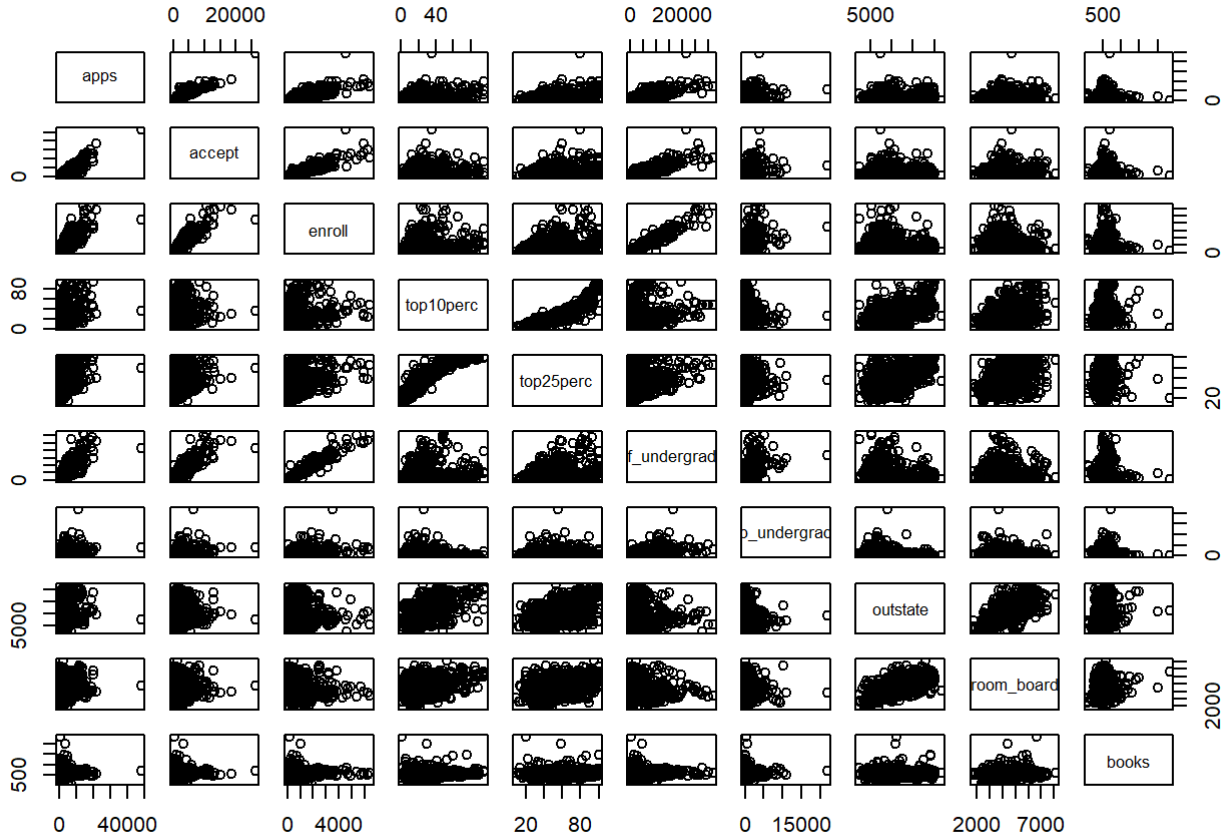
(c) i. Use the summary() function to produce a numerical summary of the variables in the data set.

```
summary(data)
```

```
##       apps           accept          enroll       top10perc       top25perc
##  Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00   Min.   :  9.0
##  1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0
##  Median : 1558   Median : 1110   Median : 434   Median :23.00   Median : 54.0
##  Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56   Mean   : 55.8
##  3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0
##  Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00   Max.   :100.0
##   f_undergrad     p_undergrad         outstate       room_board
##  Min.   :  139   Min.   :    1.0   Min.   : 2340   Min.   :1780
##  1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597
##  Median : 1707   Median :  353.0   Median : 9990   Median :4200
##  Mean   : 3700   Mean   :  855.3   Mean   :10441   Mean   :4358
##  3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050
##  Max.   :31643   Max.   :21836.0   Max.   :21700   Max.   :8124
##      books           personal          phd            terminal
##  Min.   :  96.0   Min.   : 250   Min.   :  8.00   Min.   : 24.0
##  1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.0
##  Median : 500.0   Median :1200   Median : 75.00   Median : 82.0
##  Mean   : 549.4   Mean   :1341   Mean   : 72.66   Mean   : 79.7
##  3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00   3rd Qu.: 92.0
##  Max.   :2340.0   Max.   :6800   Max.   :103.00   Max.   :100.0
##    s_f_ratio      perc_alumni        expend        grad_rate
##  Min.   : 2.50   Min.   : 0.00   Min.   : 3186   Min.   : 10.00
##  1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
##  Median :13.60   Median :21.00   Median : 8377   Median : 65.00
##  Mean   :14.09   Mean   :22.74   Mean   : 9660   Mean   : 65.46
##  3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
##  Max.   :39.80   Max.   :64.00   Max.   :56233   Max.   :118.00
```

ii. Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10].

```
first_ten_col<-data[,1:10]
pairs(first_ten_col)
```



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

```
#Is giving Error

#data$private<-as.factor(data$private)

#plot(data$private, data$outstate, xlab="Private University", ylab="Tuition in Dollars ($)")
```

iv. 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 %.

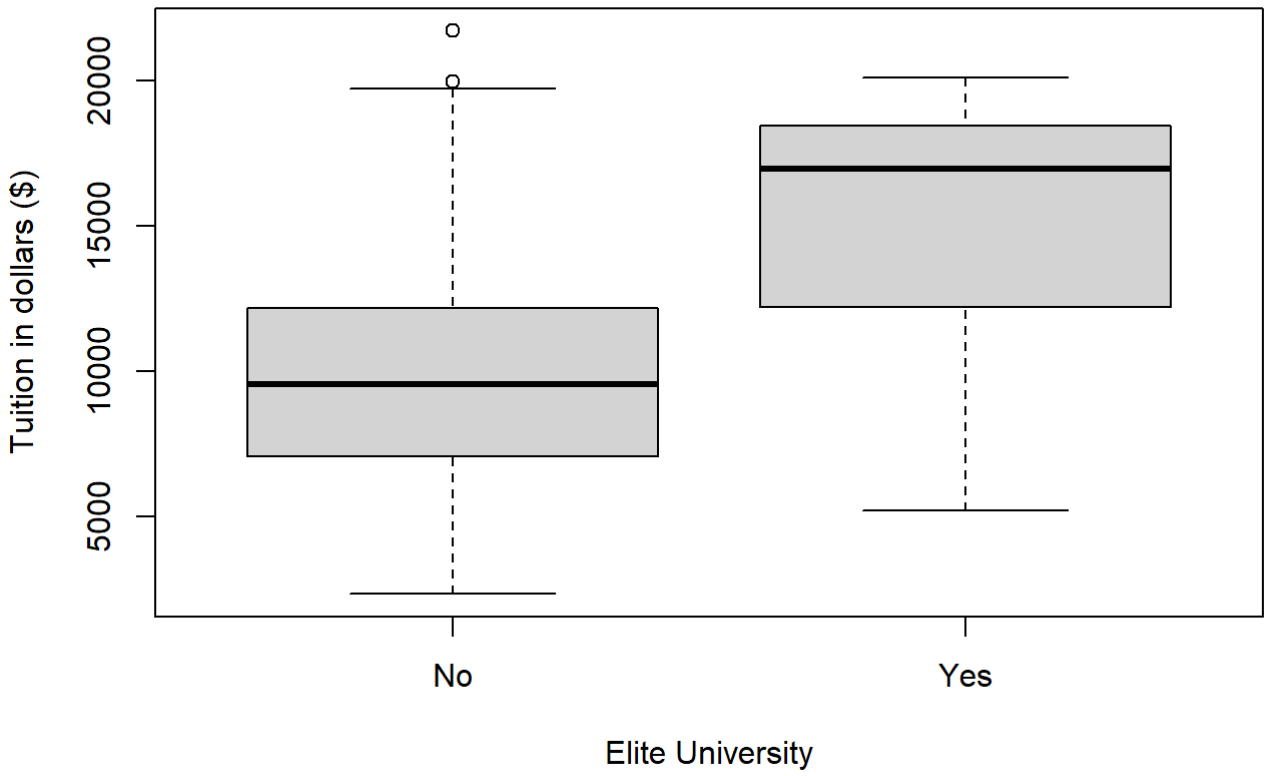
```
Elite<-rep("No",nrow(data))
Elite[data$top10perc >50]<-"Yes"
Elite<-as.factor(Elite)
data<-data.frame(data,Elite)
```

Use the summary() function to see how many elite universities there are. Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite.

```
summary(Elite)
```

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

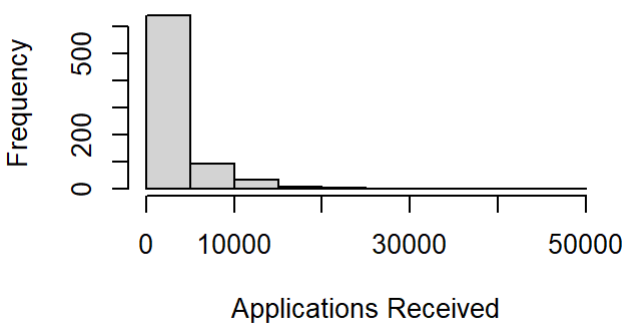
```
plot(data$Elite, data$outstate, xlab="Elite University", ylab="Tuition in dollars ($)")
```

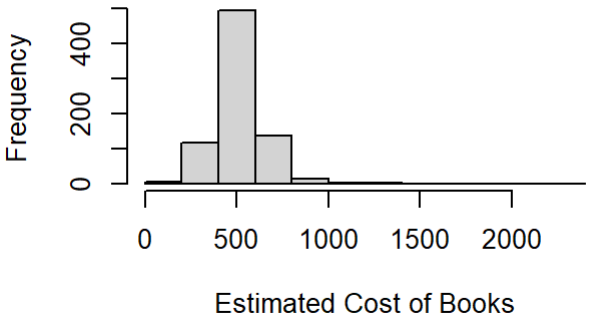


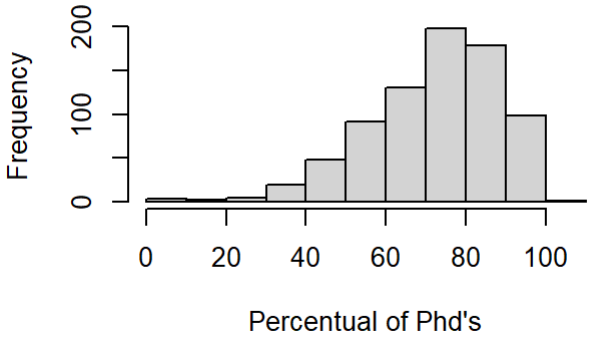
v. Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow=c(2,2)) useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.

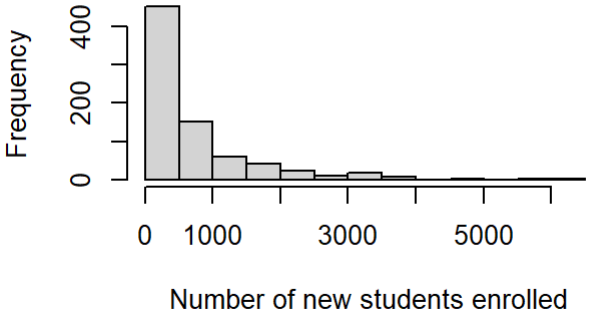
```
par(mfrow=c(2,2))

hist(data$apps, xlab="Applications Received", main="")
hist(data$books, xlab="Estimated Cost of Books", main="")
hist(data$phd, xlab="Percentual of Phd's", main="")
hist(data$enroll, xlab=" Number of new students enrolled", main="")
```









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

Lets see how many books cost above the average

```
books_above_avg<- data %>%
        filter(books >=mean(books)) %>%
        summarise(n=n())
books_above_avg
```

```
##     n
## 1 354
```

Hence, 354 books costs above the average price.