CptS 475/575: Data Science, Fall 2018

Assignment 2: R basics and Exploratory Data Analysis

Release Date: September 10, 2018 **Due Date**: September 17, 2018 (11:59 pm)

This assignment has **two exercises.** For questions that ask you to produce a specific plot, include that plot along with the code you used to generate it. You are required to use R Markdown to prepare your solution.

1. This exercise relates to the College data set, which can be found in the file College.csv on the course's public webpage (https://scads.eecs.wsu.edu/index.php/datasets/). The dataset contains a number of variables for 777 different universities and colleges in the US. The variables are

- Private : Public/private indicator
- Apps : Number of applications received
- Accept : Number of applicants accepted
- Enroll : Number of new students enrolled
- Top10perc : New students from top 10% of high school class
- Top25perc : New students from top 25% of high school class
- F.Undergrad : Number of full-time undergraduates
- P.Undergrad : Number of part-time undergraduates
- Outstate : Out-of-state tuition
- Room.Board : Room and board costs
- Books : Estimated book costs
- Personal : Estimated personal spending
- PhD : Percent of faculty with Ph.D.'s
- Terminal : Percent of faculty with terminal degree
- S.F.Ratio : Student/faculty ratio
- perc.alumni : Percent of alumni who donate
- Expend : Instructional expenditure per student
- Grad.Rate : Graduation rate

Before reading the data into R, you can view it in Excel or a text editor. For each of the following questions, include the code you used to complete the task as your response, along with any associated output.

(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.

(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:

```
> rownames (college )=college [,1]

> fix(college)
```

You should see that there is now a row.names column with the name of each university recorded. This means that R has given each row a name corresponding to the appropriate university. R will not try to perform calculations on the row names. However, we still need to eliminate the first column in the data where the names are stored. Try

```
> college =college [,-1]

> fix(college)
```

Now you should see that the first data column is Private. Note that another column labeled row.names now appears before the Private column. However, this is not a data column but rather the name that R is giving to each row.

(c)

i. Use the summary() function to produce a numerical summary of the variables in the data set. (Respond to this question with the mean graduation rate included in the summary result).

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].

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

iv. Create a new qualitative variable, called Top, by binning the Top25perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 25% of their high school classes exceeds 50%.

```
> Top=rep("No",nrow(college ))

> Top[college$Top25perc >50]=" Yes"

> Top=as.factor(Top)

> college=data.frame(college, Top)
```

Use the summary() function to see how many top universities there are. Now use the plot() or boxplot() function to produce side-by-side boxplots of Outstate with respect to

the two Top categories (Yes and No). Ensure that this figure has an appropriate title and axis labels.

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. Again, ensure that this figure has an appropriate title and axis labels.

vi. Continue exploring the data, and provide a brief summary of what you discover. You may use additional plots or numerical descriptors as needed. Feel free to think outside the box on this one but if you want something to point you in the right direction, look at the summary statistics for various features, and think about what they tell you. Perhaps try plotting various features from the dataset against each other and see if any patterns emerge.

2. This exercise involves the Auto.csv data set found on the course website. Make sure that the missing values have been removed from the data. To do this, consider the na.strings parameter of read.csv(), as well as the na.omit() function.

(a) Which of the predictors are quantitative, and which are qualitative?

(b) What is the range of each quantitative predictor? You can answer this using the range() function. Hint: consider using R's sapply() function to take the range of multiple features in a single function call.

(c) What is the mean and standard deviation of each quantitative predictor?

(d) Now remove the 25th through 75th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

(f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

Assignment 2

Sheryl Mathew (11627236)

14 September, 2018

Question 1

1. Read the College csv file

```
library(kableExtra)
college = read.csv("https://scads.eecs.wsu.edu/wp-content/uploads/2017/09/College.csv")
##fix (college)
kable(head(college), format = "latex", booktabs = T,
      caption="Table with the College Names from the CSV") %>%
kable_styling(latex_options = c("striped","hold_position","scale_down"))
```

Table 1: Table with the College Names from the CSV

X	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041	60
Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527	56
Adrian College	Yes	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735	54
Agnes Scott College	Yes	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016	59
Alaska Pacific University	Yes	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922	15
Albertson College	Yes	587	479	158	38	62	678	41	13500	3335	500	675	67	73	9.4	11	9727	55

```
rownames (college )=college [,1]
##fix(college)
kable(head(college), format = "latex", booktabs = T,
      caption="Table with the College Row Names assigned by R Markdown and
      College Names from the CSV") %>%
kable_styling(latex_options = c("striped","hold_position","scale_down"))
```

Table 2: Table with the College Row Names assigned by R Markdown and College Names from the CSV

	X	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Abilene Christian University	Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041	60
Adelphi University	Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527	56
Adrian College	Adrian College	Yes	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735	54
Agnes Scott College	Agnes Scott College	Yes	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016	59
Alaska Pacific University	Alaska Pacific University	Yes	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922	15
Albertson College	Albertson College	Yes	587	479	158	38	62	678	41	13500	3335	500	675	67	73	9.4	11	9727	55

```
college =college [,-1]
##fix(college)
kable(head(college), format = "latex", booktabs = T,
      caption="Table with the College Row Names assigned by R Markdown") %>%
kable_styling(latex_options = c("striped","hold_position","scale_down"))
```

Table 3: Table with the College Row Names assigned by R Markdown

	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041	60
Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527	56
Adrian College	Yes	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735	54
Agnes Scott College	Yes	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016	59
Alaska Pacific University	Yes	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922	15
Albertson College	Yes	587	479	158	38	62	678	41	13500	3335	500	675	67	73	9.4	11	9727	55

2. Summary of College

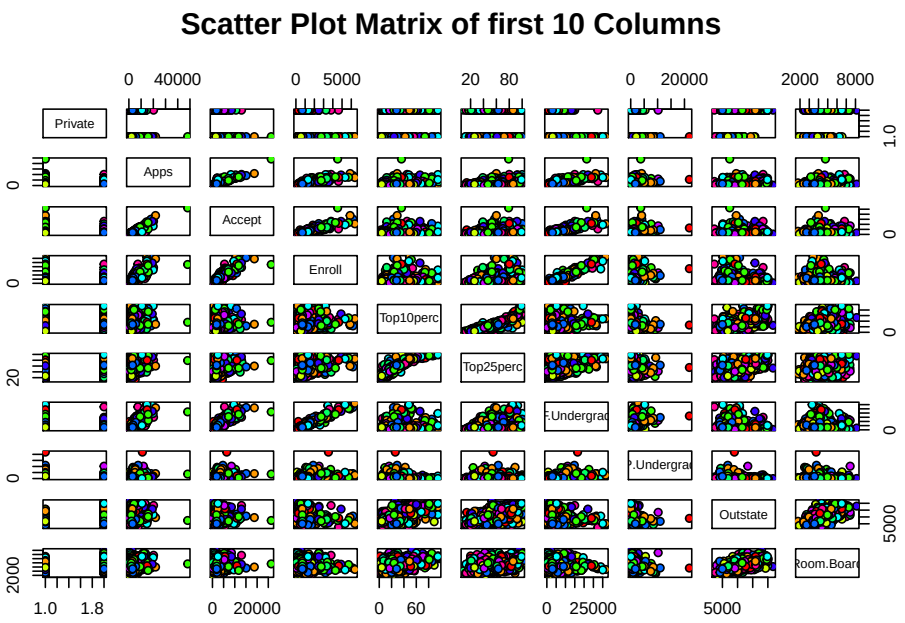
```
summary(college)
```

```
##  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
```

The mean graduation rate is **65.4633205**

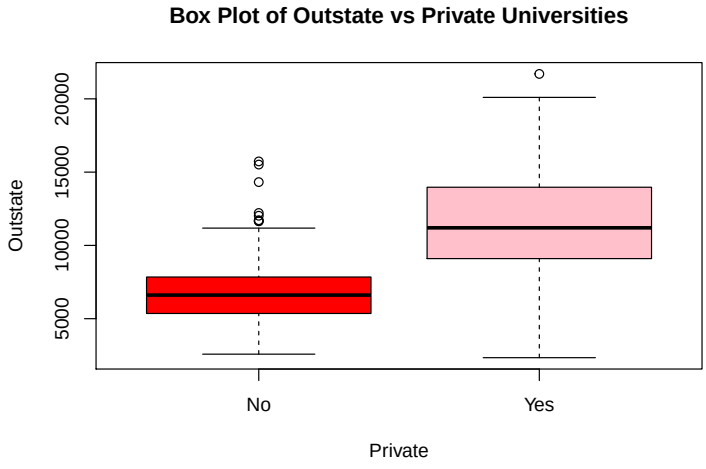
3. Scatter Plot Matrix of first 10 columns

```
pairs(college[,1:10],main="Scatter Plot Matrix of first 10 Columns",
      pch=21,bg=rainbow(10))
```



4. Box Plot of Outstate vs Private

```
plot(y=college$Outstate, x=college$Private, main="Box Plot of Outstate vs Private Universities",
     xlab="Private", ylab="Outstate", col=c("red","pink"))
```



5. Comparison of Outstate vs Top Universities

Getting The Top University Details

```
Top=rep("No",nrow(college ))
Top[college$Top25perc >50]=" Yes"
Top=as.factor(Top)
Top
```

```
##   [1]  Yes No   No    Yes No    Yes No    Yes  Yes No    Yes  Yes  Yes  Yes
##  [15] No   No    Yes No   No    Yes No    Yes  Yes No    Yes  Yes  Yes  Yes
##  [29] No    Yes  Yes  Yes No   No    Yes No    Yes  Yes No    Yes  Yes  Yes
##  [43] No    Yes  Yes No   No    Yes  Yes  Yes No   No   No   No    Yes  Yes
##  [57]  Yes No   No    Yes  Yes No   No    Yes  Yes No   No   No    Yes  Yes
##  [71]  Yes  Yes  Yes  Yes  Yes No   No    Yes  Yes  Yes  Yes  Yes  Yes No
##  [85]  Yes  Yes  Yes  Yes  Yes  Yes No    Yes No   No   No   No    Yes  Yes
##  [99] No    Yes No    Yes No   No    Yes No    Yes No    Yes  Yes  Yes No
## [113] No   No    Yes  Yes No    Yes  Yes No    Yes No    Yes  Yes  Yes  Yes
## [127] No   No   No    Yes  Yes  Yes  Yes  Yes No   No    Yes  Yes  Yes  Yes
## [141]  Yes  Yes No    Yes  Yes No   No    Yes  Yes  Yes  Yes  Yes  Yes  Yes
## [155] No   No   No   No    Yes  Yes No   No    Yes  Yes  Yes No    Yes No
## [169] No   No   No    Yes  Yes  Yes  Yes  Yes No   No   No   No   No   No
## [183] No   No    Yes  Yes No   No   No   No    Yes  Yes  Yes  Yes  Yes No
## [197]  Yes No   No    Yes  Yes  Yes No    Yes No    Yes No   No   No    Yes
## [211] No    Yes  Yes  Yes  Yes No   No    Yes No    Yes  Yes  Yes  Yes No
## [225]  Yes  Yes No    Yes  Yes  Yes  Yes No   No    Yes No   No   No    Yes
## [239]  Yes No    Yes  Yes  Yes  Yes No    Yes  Yes No    Yes No    Yes  Yes
## [253]  Yes  Yes  Yes  Yes  Yes  Yes  Yes  Yes  Yes  Yes No   No   No   No
## [267] No    Yes  Yes No    Yes No   No   No    Yes No   No    Yes  Yes  Yes
## [281] No   No    Yes  Yes  Yes No   No    Yes  Yes No   No   No    Yes No
## [295] No    Yes  Yes No    Yes  Yes No    Yes No   No   No   No    Yes  Yes
## [309]  Yes  Yes No   No    Yes  Yes No    Yes No   No    Yes No   No   No
## [323] No    Yes  Yes No    Yes  Yes  Yes  Yes  Yes No   No   No    Yes No
## [337] No    Yes  Yes  Yes No   No    Yes No    Yes  Yes  Yes  Yes No   No
## [351] No    Yes  Yes  Yes  Yes No   No   No   No    Yes No    Yes No   No
## [365]  Yes No    Yes  Yes No    Yes  Yes  Yes  Yes  Yes  Yes  Yes No   No
## [379] No    Yes No    Yes No    Yes No   No    Yes  Yes No   No    Yes No
## [393] No   No   No   No   No    Yes  Yes No    Yes  Yes No   No    Yes  Yes
## [407]  Yes  Yes No    Yes No   No    Yes No    Yes No   No    Yes No   No
## [421] No   No    Yes  Yes  Yes No   No    Yes  Yes  Yes  Yes  Yes  Yes  Yes
## [435]  Yes No    Yes  Yes  Yes No    Yes  Yes No    Yes No    Yes  Yes No
## [449] No   No   No   No   No    Yes No   No    Yes No    Yes  Yes  Yes  Yes
## [463]  Yes  Yes No   No   No   No    Yes  Yes No    Yes  Yes No    Yes No
## [477]  Yes No   No   No    Yes  Yes  Yes  Yes  Yes  Yes  Yes No   No   No
## [491] No   No   No   No   No    Yes  Yes  Yes  Yes No   No    Yes  Yes No
## [505] No    Yes  Yes  Yes  Yes  Yes  Yes  Yes  Yes  Yes  Yes  Yes  Yes  Yes
## [519] No    Yes  Yes  Yes  Yes No    Yes  Yes No    Yes  Yes No   No   No
## [533] No   No    Yes No   No   No   No    Yes  Yes No    Yes No   No    Yes
## [547]  Yes  Yes No    Yes  Yes  Yes  Yes No   No   No    Yes  Yes  Yes  Yes
## [561]  Yes  Yes  Yes  Yes No    Yes No   No   No    Yes  Yes No    Yes No
## [575]  Yes  Yes  Yes No    Yes  Yes No    Yes No    Yes  Yes  Yes No   No
## [589] No   No    Yes  Yes No    Yes No   No    Yes  Yes No    Yes No    Yes
## [603] No   No    Yes  Yes  Yes  Yes No    Yes  Yes  Yes  Yes  Yes  Yes  Yes
```

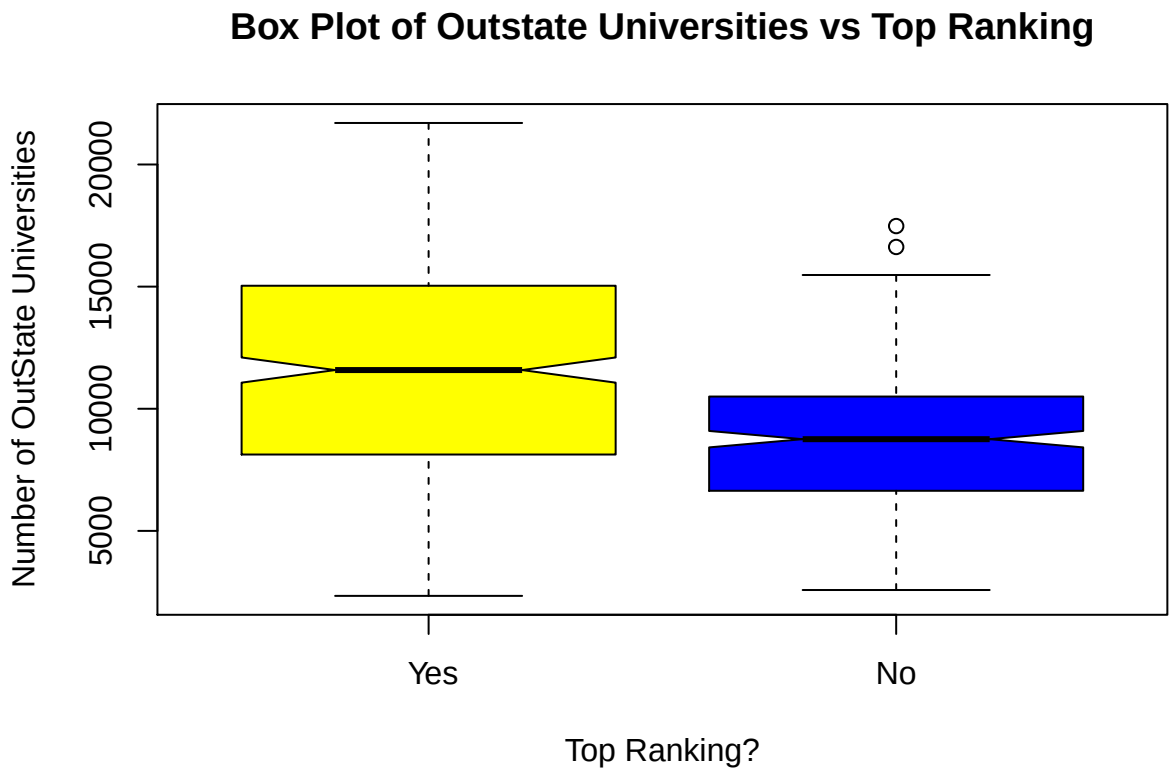
```
## [617] No   No    Yes  Yes  Yes No    Yes  Yes No    Yes No   No   No   No
## [631] No   No    Yes  Yes No   No    Yes  Yes No    Yes  Yes No    Yes  Yes
## [645]  Yes  Yes  Yes  Yes  Yes  Yes  Yes  Yes  Yes No    Yes  Yes  Yes  Yes
## [659] No    Yes  Yes  Yes  Yes  Yes  Yes  Yes  Yes No    Yes  Yes  Yes No
## [673] No    Yes No    Yes  Yes  Yes No   No   No    Yes  Yes  Yes No    Yes
## [687] No   No    Yes  Yes  Yes  Yes  Yes  Yes  Yes  Yes No    Yes No   No
## [701]  Yes No   No   No    Yes No   No    Yes  Yes  Yes  Yes No   No    Yes
## [715]  Yes No   No   No    Yes No    Yes  Yes No    Yes  Yes  Yes  Yes  Yes
## [729]  Yes No   No   No   No    Yes  Yes No   No    Yes  Yes No    Yes No
## [743] No    Yes No   No    Yes No   No    Yes  Yes  Yes  Yes  Yes No    Yes
## [757]  Yes  Yes  Yes No    Yes  Yes No    Yes No   No   No    Yes No    Yes
## [771]  Yes  Yes No   No    Yes  Yes  Yes
## Levels:  Yes No
```

```
college=data.frame(college, Top)
summary (Top)
```

```
##  Yes   No
##  449  328
```

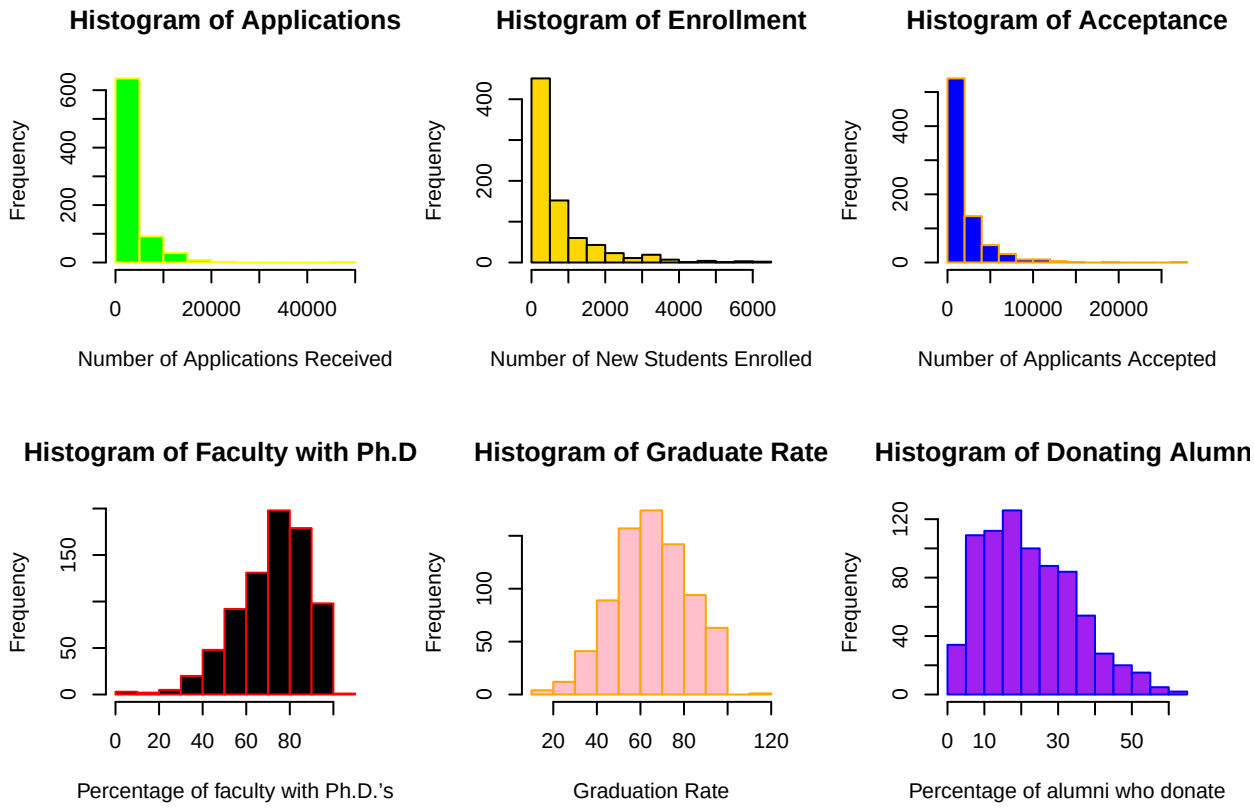
Plotting the Top University Details

```
boxplot(college$Outstate~Top,xlab="Top Ranking?", ylab="Number of OutState Universities", main ="Box Pl
```



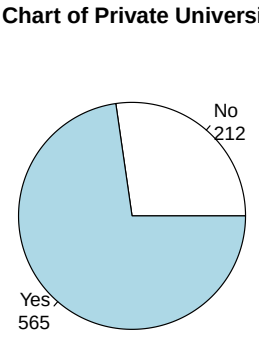
6. Use hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables.

```
par(mfrow=c(2,3))
hist(x=college$Apps,col="green",border="yellow",main="Histogram of Applications",
     xlab="Number of Applications Received")
hist(x=college$Enroll,col="gold",border="black",main="Histogram of Enrollment",
     xlab="Number of New Students Enrolled")
hist(x=college$Accept,col="blue",border="orange",main="Histogram of Acceptance",
     xlab="Number of Applicants Accepted")
hist(x=college$PhD,col="black",border="red",main="Histogram of Faculty with Ph.D.'s",
     xlab="Percentage of faculty with Ph.D.'s")
hist(x=college$Grad.Rate,col="pink",border="orange",main="Histogram of Graduate Rate",
     xlab="Graduation Rate")
hist(x=college$perc.alumni,col="purple",border="blue",main="Histogram of Donating Alumni",xlab="Percenta
```



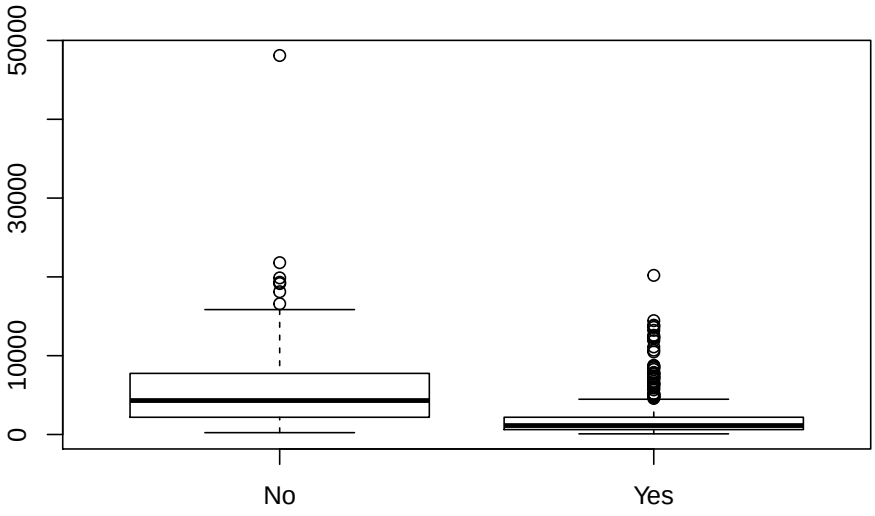
7. Data Exploration

```
mytable <- table(college$Private)
lbls <- paste(names(mytable), "\n", mytable, sep="")
pie(mytable, labels = lbls,
    main="Pie Chart of Private Universities")
```



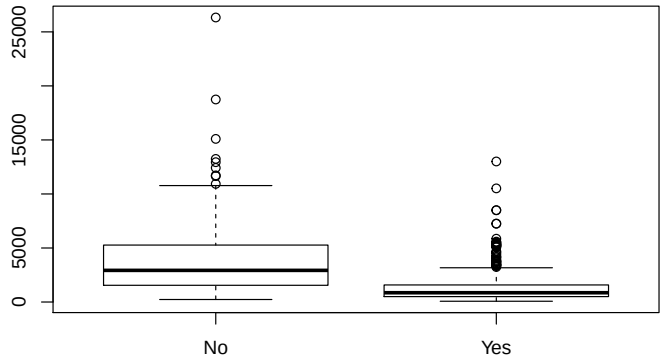
The number of Private Universities are more than Public Universities

```
plot(college$Private,college$Apps)
```



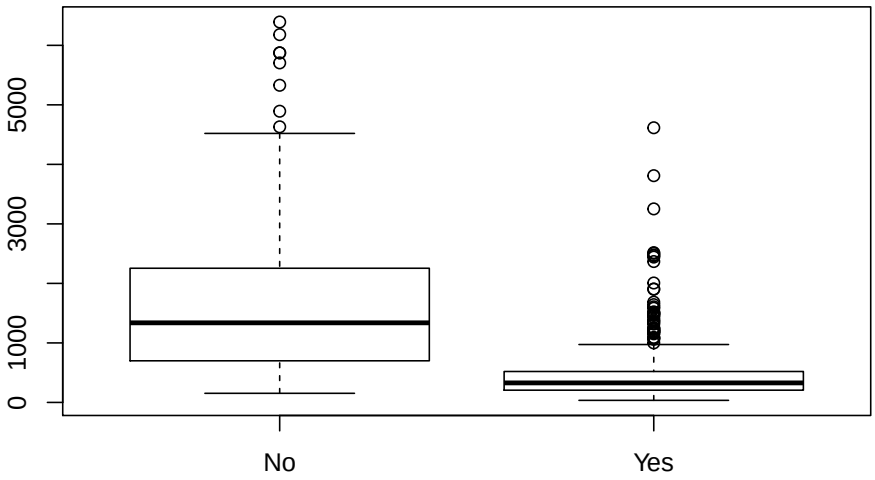
The number of applications sent to Public universities is more than Private.

```
plot(college$Private,college$Accept)
```



The number of acceptances by Public universities is more than Private.

```
plot(college$Private,college$Enroll)
```



The number of enrollment by Public universities is more than Private.

Conclusion: Though the number of Private Universities are more, their acceptance and enrollment rate is less when compared to Public Universities.

Question 2

1. Read the Auto csv file

```
auto=read.csv("https://scads.eecs.wsu.edu/wp-content/uploads/2017/09/Auto.csv",na.strings="?")
table(is.na(auto))
```

```
##
## FALSE  TRUE
##  3568     5
```

```
autoWithOmissions=subset(auto,is.na(auto$horsepower))
kable(autoWithOmissions, format = "latex", booktabs = T,
      caption="Table containing the NA for Horsepower in Auto CSV") %>%
kable_styling(latex_options = c("striped","hold_position","scale_down"))
```

Table 4: Table containing the NA for Horsepower in Auto CSV

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	name
33	25.0	4	98	NA	2046	19.0	71	1	ford pinto
127	21.0	6	200	NA	2875	17.0	74	1	ford maverick
331	40.9	4	85	NA	1835	17.3	80	2	renault lecar deluxe
337	23.6	4	140	NA	2905	14.3	80	1	ford mustang cobra
355	34.5	4	100	NA	2320	15.8	81	2	renault 18i

```
autoWithNoOmissions=na.omit(auto)
table(is.na(autoWithNoOmissions))
```

```
##
## FALSE
##  3528
```

2. Which of the predictors are quantitative, and which are qualitative?

```
kable(head(autoWithNoOmissions), format = "latex", booktabs = T,
      caption="Table with the first 5 rows of Auto.csv") %>%
kable_styling(latex_options = c("striped","hold_position","scale_down"))
```

Table 5: Table with the first 5 rows of Auto.csv

mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	name
18	8	307	130	3504	12.0	70	1	chevrolet chevelle malibu
15	8	350	165	3693	11.5	70	1	buick skylark 320
18	8	318	150	3436	11.0	70	1	plymouth satellite
16	8	304	150	3433	12.0	70	1	amc rebel sst
17	8	302	140	3449	10.5	70	1	ford torino
15	8	429	198	4341	10.0	70	1	ford galaxie 500

```
sapply(auto,class)
```

```
##          mpg    cylinders displacement   horsepower       weight
##    "numeric"    "integer"    "numeric"    "integer"    "integer"
## acceleration         year       origin         name
##    "numeric"    "integer"    "integer"     "factor"
```

Quantitative Predictors:: mpg, cylinders, displacement, horsepower, weight, acceleration, year, origin
Qualitative Predictors: name

3. What is the range of each quantitative predictor?

```
range(autoWithNoOmissions$mpg)
```

```
## [1]  9.0 46.6
```

```
rangeMatrix=sapply(list(autoWithNoOmissions$mpg,autoWithNoOmissions$cylinders,
                        autoWithNoOmissions$displacement,autoWithNoOmissions$horsepower,
                        autoWithNoOmissions$weight,autoWithNoOmissions$acceleration,
                        autoWithNoOmissions$year,autoWithNoOmissions$origin),range)
colnames(rangeMatrix) =c("mpg","cylinders","displacement","horsepower","weight","acceleration",
"year","origin")
rownames(rangeMatrix) =c("min","max")
print (rangeMatrix)
```

```
##      mpg cylinders displacement horsepower weight acceleration year origin
## min  9.0         3           68         46   1613          8.0   70      1
## max 46.6         8          455        230   5140         24.8   82      3
```

4. What is the mean and standard deviation of each quantitative predictor?

Mean

```
meanValues=sapply(list(autoWithNoOmissions$mpg,autoWithNoOmissions$cylinders,
                       autoWithNoOmissions$displacement,autoWithNoOmissions$horsepower,
                       autoWithNoOmissions$weight,autoWithNoOmissions$acceleration,
                       autoWithNoOmissions$year,autoWithNoOmissions$origin),mean)
meanValues
```

```
## [1]   23.445918    5.471939  194.411990  104.469388 2977.584184   15.541327
## [7]   75.979592    1.576531
```

Standard Deviation

```
standardDeviation=sapply(list(autoWithNoOmissions$mpg,autoWithNoOmissions$cylinders,
                       autoWithNoOmissions$displacement,autoWithNoOmissions$horsepower,
                       autoWithNoOmissions$weight,autoWithNoOmissions$acceleration,
                       autoWithNoOmissions$year,autoWithNoOmissions$origin),sd)
standardDeviation
```

```
## [1]   7.8050075   1.7057832 104.6440039  38.4911599 849.4025600   2.7588641
## [7]   3.6837365   0.8055182
```

5. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains after removing observations 25th through 75th?

Total number of Auto's after omitting null values

```
nrow(autoWithNoOmissions)
```

```
## [1] 392
```

Total number of autos after eliminating 25 to 75

```
till24=autoWithNoOmissions[1:24,]
nrow(till24)
```

```
## [1] 24
```

```
from76=autoWithNoOmissions[76:nrow(autoWithNoOmissions),]
nrow(from76)
```

```
## [1] 317
```

Combining the two separate data frames

```
combinedData=rbind(till24,from76)
nrow(combinedData)
```

```
## [1] 341
```

Range

```
rangeMatrix=sapply(list(combinedData$mpg,combinedData$cylinders,
                        combinedData$displacement,combinedData$horsepower,
                        combinedData$weight,combinedData$acceleration,
                        combinedData$year,combinedData$origin),range)
colnames(rangeMatrix) =c("mpg","cylinders","displacement","horsepower","weight","acceleration",
"year","origin")
rownames(rangeMatrix) =c("min","max")
print (rangeMatrix)
```

```
##      mpg cylinders displacement horsepower weight acceleration year origin
## min 11.0         3           68         46   1649          8.0   70      1
## max 46.6         8          455        230   4997         24.8   82      3
```

Mean

```
meanValues=sapply(list(combinedData$mpg,combinedData$cylinders,
                        combinedData$displacement,combinedData$horsepower,
                        autoWithNoOmissions$weight,combinedData$acceleration,
                        combinedData$year,combinedData$origin),mean)
meanValues
```

```
## [1]   24.195894    5.360704  187.167155  101.395894 2977.584184   15.650147
## [7]   76.683284    1.612903
```

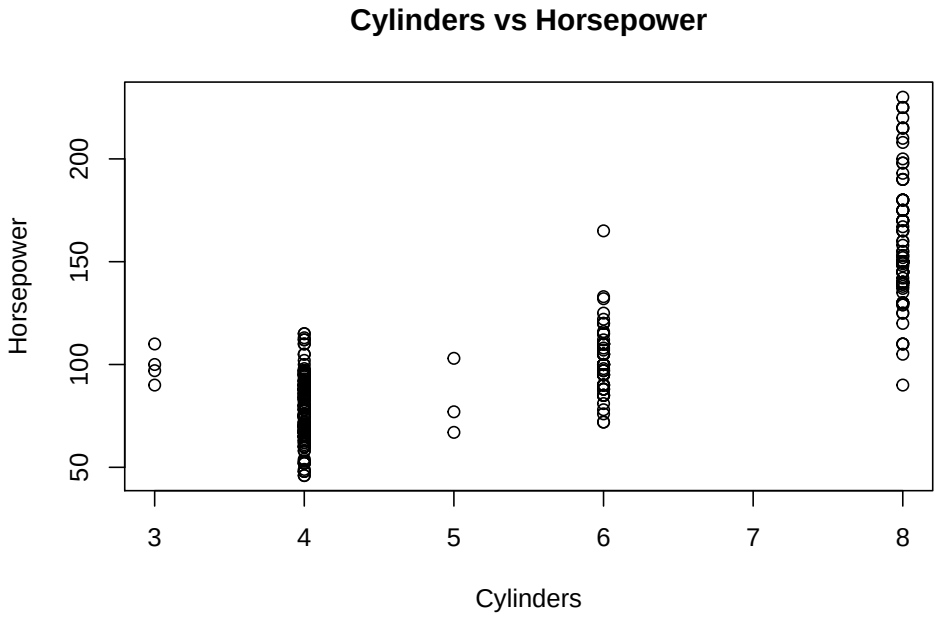
Standard Deviation

```
standardDeviation=sapply(list(combinedData$mpg,combinedData$cylinders,
                        combinedData$displacement,combinedData$horsepower,
                        combinedData$weight,combinedData$acceleration,
                        combinedData$year,combinedData$origin),sd)
standardDeviation
```

```
## [1]   7.7205330   1.6579873 101.1198397  36.2987423 799.6367540   2.7552156
## [7]   3.4247347   0.8169225
```

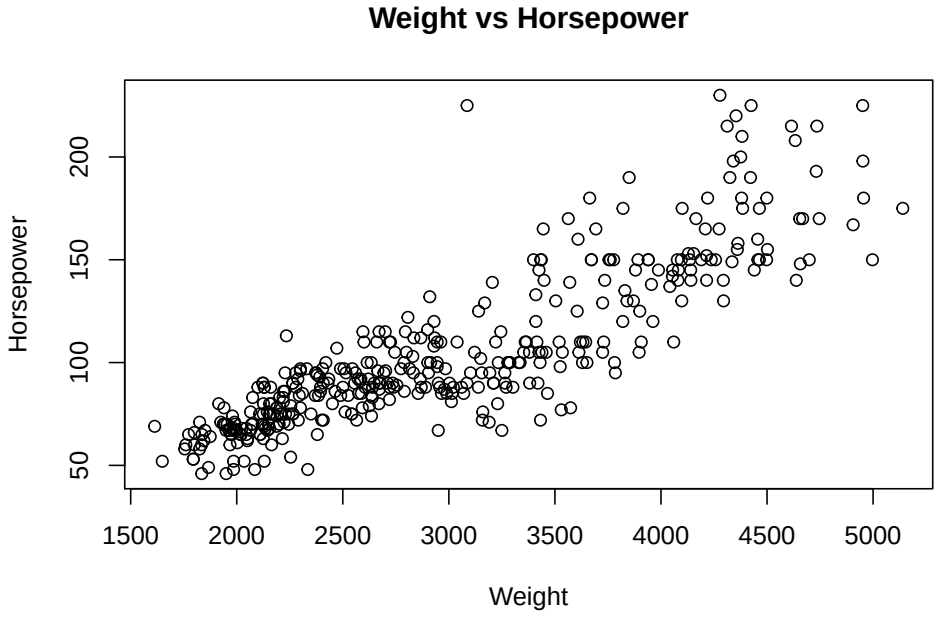
6. Exploring the data

```
plot(x=auto$cylinders,y=auto$horsepower,main="Cylinders vs Horsepower",xlab="Cylinders",ylab="Horsepowe
```



More the number of cylinders more the range of horsepower available

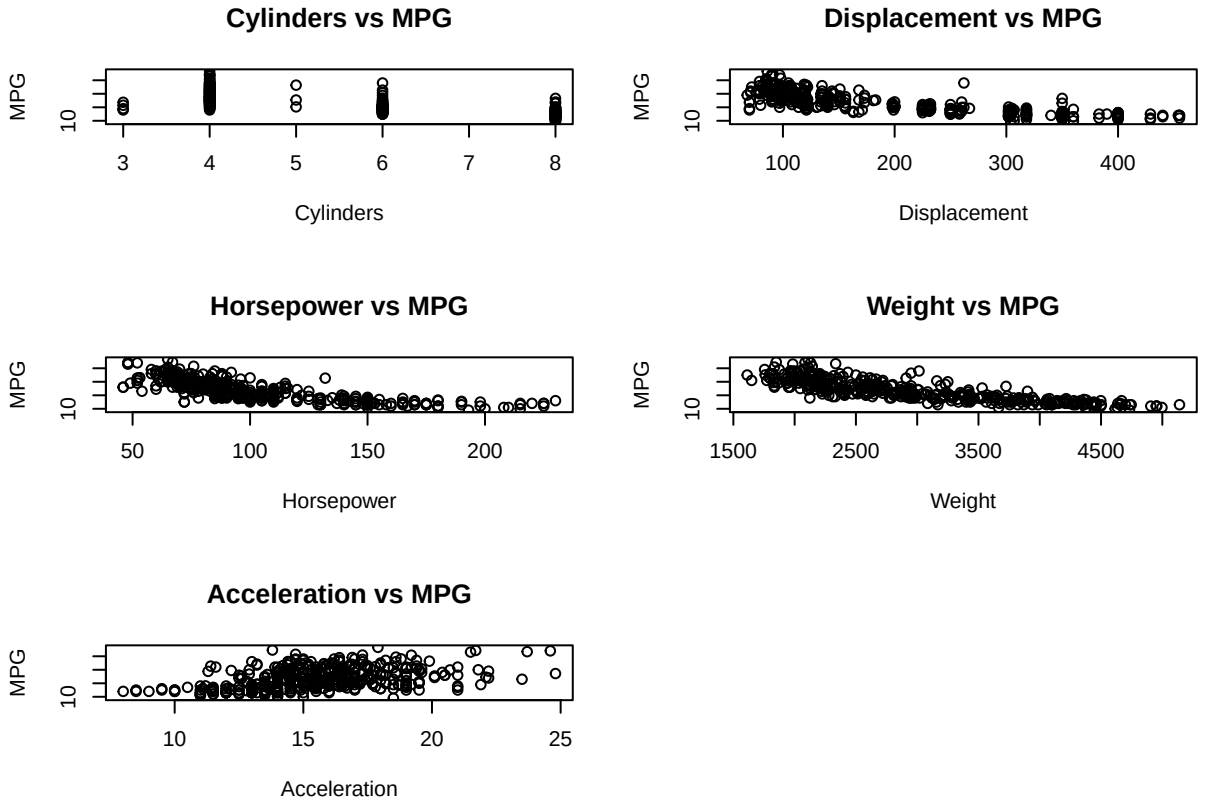
```
plot(x=auto$weight,y=auto$horsepower,main="Weight vs Horsepower",xlab="Weight",ylab="Horsepower")
```



As the weight of the vehicle increases more horsepower is needed.

7.Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer

```
par(mfrow=c(3,2))
plot(x=auto$cylinders,y=auto$mpg,main="Cylinders vs MPG",xlab="Cylinders",ylab="MPG")
plot(x=auto$displacement,y=auto$mpg,main="Displacement vs MPG",xlab="Displacement",ylab="MPG")
plot(x=auto$horsepower,y=auto$mpg,main="Horsepower vs MPG",xlab="Horsepower",ylab="MPG")
plot(x=auto$weight,y=auto$mpg,main="Weight vs MPG",xlab="Weight",ylab="MPG")
plot(x=auto$acceleration,y=auto$mpg,main="Acceleration vs MPG",xlab="Acceleration",ylab="MPG")
```



From the above plots we can see that MPG depends on **Horsepower** and **Weight** of the vehicle. If the weight or horsepower of the vehicle increases the MPG decreases. Therefore we can predict the MPG of a stationary vehicle.

If we want to predict the MPG of a moving vehicle we need to know the **Acceleration** of the vehicle at that time along with **Horsepower** and **Weight**. This is because when the acceleration increases MPG decreases.