ISLRChp2Applied-8

July 30, 2022

0.0.1 Exploratory Data Analysis of Universities and Colleges in United States in R

This data set contains a number of variables for **777** different universities and colleges in the United States. 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
> **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

We read the data by using the language of R. We pass this to the variable college. Just make sure that the directory is correct so that the data will be read.

[3]:
```
# Open the data set by using the function read.csv
college <- read.csv("H:/Programming/Jupyter Notebook/datasets/College.csv")

# View the data frame
View(college)
```

A data.frame: 777 × 19

X <chr>	Private <chr>	Apps <int>	Accept <int>	Enroll <int>	Top10perc <int>
Abilene Christian University	Yes	1660	1232	721	23
Adelphi University	Yes	2186	1924	512	16
Adrian College	Yes	1428	1097	336	22
Agnes Scott College	Yes	417	349	137	60
Alaska Pacific University	Yes	193	146	55	16
Albertson College	Yes	587	479	158	38
Albertus Magnus College	Yes	353	340	103	17
Albion College	Yes	1899	1720	489	37
Albright College	Yes	1038	839	227	30
Alderson-Broaddus College	Yes	582	498	172	21
Alfred University	Yes	1732	1425	472	37
Allegheny College	Yes	2652	1900	484	44
Allentown Coll. of St. Francis de Sales	Yes	1179	780	290	38
Alma College	Yes	1267	1080	385	44
Alverno College	Yes	494	313	157	23
American International College	Yes	1420	1093	220	9
Amherst College	Yes	4302	992	418	83
Anderson University	Yes	1216	908	423	19
Andrews University	Yes	1130	704	322	14
Angelo State University	No	3540	2001	1016	24
Antioch University	Yes	713	661	252	25
Appalachian State University	No	7313	4664	1910	20
Aquinas College	Yes	619	516	219	20
Arizona State University Main campus	No	12809	10308	3761	24
Arkansas College (Lyon College)	Yes	708	334	166	46
Arkansas Tech University	No	1734	1729	951	12
Assumption College	Yes	2135	1700	491	23
Auburn University-Main Campus	No	7548	6791	3070	25
Augsburg College	Yes	662	513	257	12
Augustana College IL	Yes	1879	1658	497	36
⋮	⋮	⋮	⋮	⋮	⋮
Westfield State College	No	3100	2150	825	3
Westminster College MO	Yes	662	553	184	20
Westminster College	Yes	996	866	377	29
Westminster College of Salt Lake City	Yes	917	720	213	21
Westmont College	No	950	713	351	42
Wheaton College IL	Yes	1432	920	548	56
Westminster College PA	Yes	1738	1373	417	21
Wheeling Jesuit College	Yes	903	755	213	15
Whitman College	Yes	1861	998	359	45
Whittier College	Yes	1681	1069	344	35
Whitworth College	Yes	1121	926	372	43
Widener University	Yes	2139	1492	502	24
Wilkes University	Yes	1631	1431	434	15
Willamette University	Yes	1658	1327	395	49
William Jewell College	Yes	663	547	315	32
William Woods University	Yes	469	435	227	17
Williams College	Yes	4186	1245	526	81
Wilson College	Yes	167	130	46	16
Wingate College	Yes	1239	1017	383	10
Winona State University	No	3325	2047	1301	20

We 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.

[4]:
```
# Change the index
rownames(college) <- college[, 1]
college <- college[, -1]

View(college)
```

	Private	Apps	Accept	Enroll	Top10perc
	<chr>	<int>	<int>	<int>	<int>
Abilene Christian University	Yes	1660	1232	721	23
Adelphi University	Yes	2186	1924	512	16
Adrian College	Yes	1428	1097	336	22
Agnes Scott College	Yes	417	349	137	60
Alaska Pacific University	Yes	193	146	55	16
Albertson College	Yes	587	479	158	38
Albertus Magnus College	Yes	353	340	103	17
Albion College	Yes	1899	1720	489	37
Albright College	Yes	1038	839	227	30
Alderson-Broaddus College	Yes	582	498	172	21
Alfred University	Yes	1732	1425	472	37
Allegheny College	Yes	2652	1900	484	44
Allentown Coll. of St. Francis de Sales	Yes	1179	780	290	38
Alma College	Yes	1267	1080	385	44
Alverno College	Yes	494	313	157	23
American International College	Yes	1420	1093	220	9
Amherst College	Yes	4302	992	418	83
Anderson University	Yes	1216	908	423	19
Andrews University	Yes	1130	704	322	14
Angelo State University	No	3540	2001	1016	24
Antioch University	Yes	713	661	252	25
Appalachian State University	No	7313	4664	1910	20
Aquinas College	Yes	619	516	219	20
Arizona State University Main campus	No	12809	10308	3761	24
Arkansas College (Lyon College)	Yes	708	334	166	46
Arkansas Tech University	No	1734	1729	951	12
Assumption College	Yes	2135	1700	491	23
Auburn University-Main Campus	No	7548	6791	3070	25
Augsburg College	Yes	662	513	257	12
Augustana College IL	Yes	1879	1658	497	36
Westfield State College	No	3100	2150	825	3
Westminster College MO	Yes	662	553	184	20
Westminster College	Yes	996	866	377	29
Westminster College of Salt Lake City	Yes	917	720	213	21
Westmont College	No	950	713	351	42
Wheaton College IL	Yes	1432	920	548	56
Westminster College PA	Yes	1738	1373	417	21
Wheeling Jesuit College	Yes	903	755	213	15
Whitman College	Yes	1861	998	359	45
Whittier College	Yes	1681	1069	344	35
Whitworth College	Yes	1121	926	372	43
Widener University	Yes	2139	1492	502	24
Wilkes University	Yes	1631	1431	434	15
Willamette University	Yes	1658	1327	395	49
William Jewell College	Yes	663	547	315	32
William Woods University	Yes	469	435	227	17
Williams College	Yes	4186	1245	526	81
Wilson College	Yes	167	130	46	16
Wingate College	Yes	1239	1017	383	10
Winona State University	No	3325	2047	1301	20

A data.frame: 777 × 18

We get the numerical summaries of the variables in the data set.

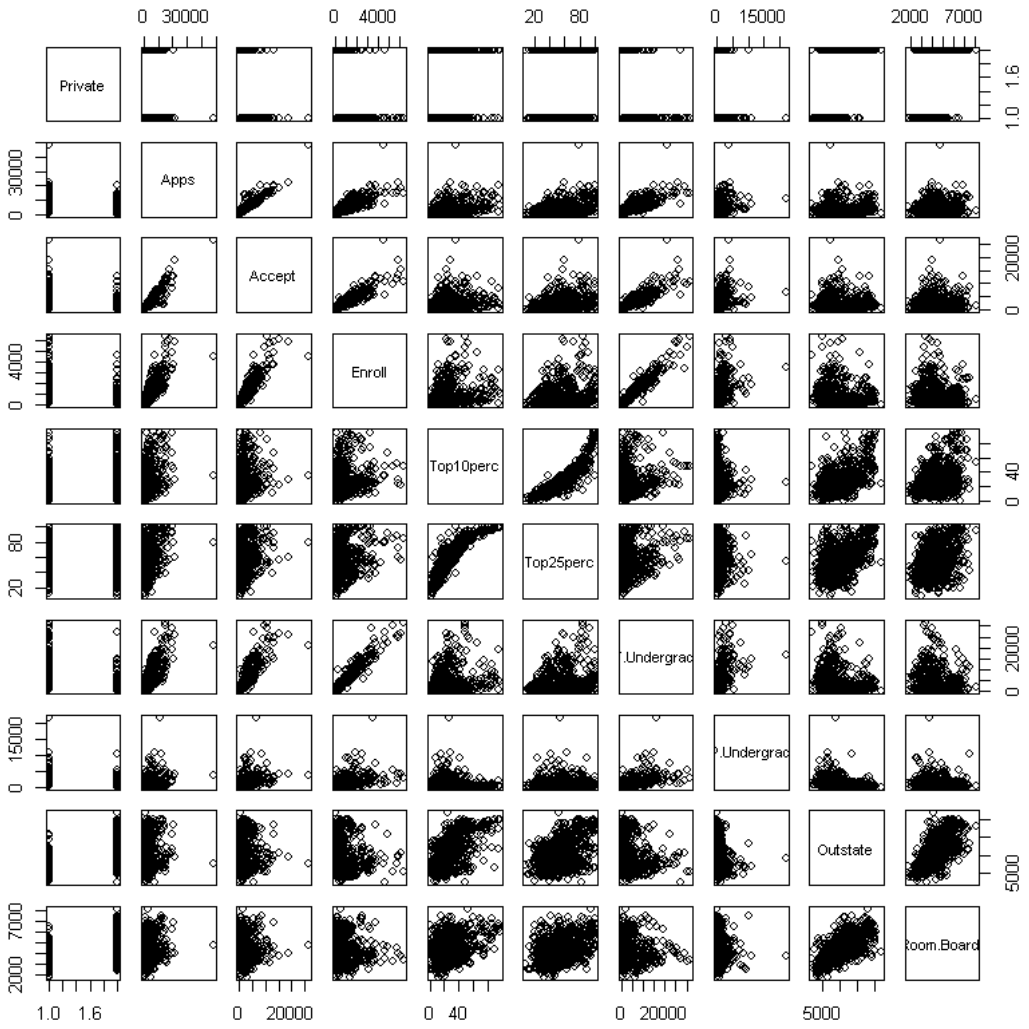
[5]:
```
summary(college)
```

```
  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
```

We produce pairs plot to generate an overall insight from the data set visually. It can be done for the first ten columns only.

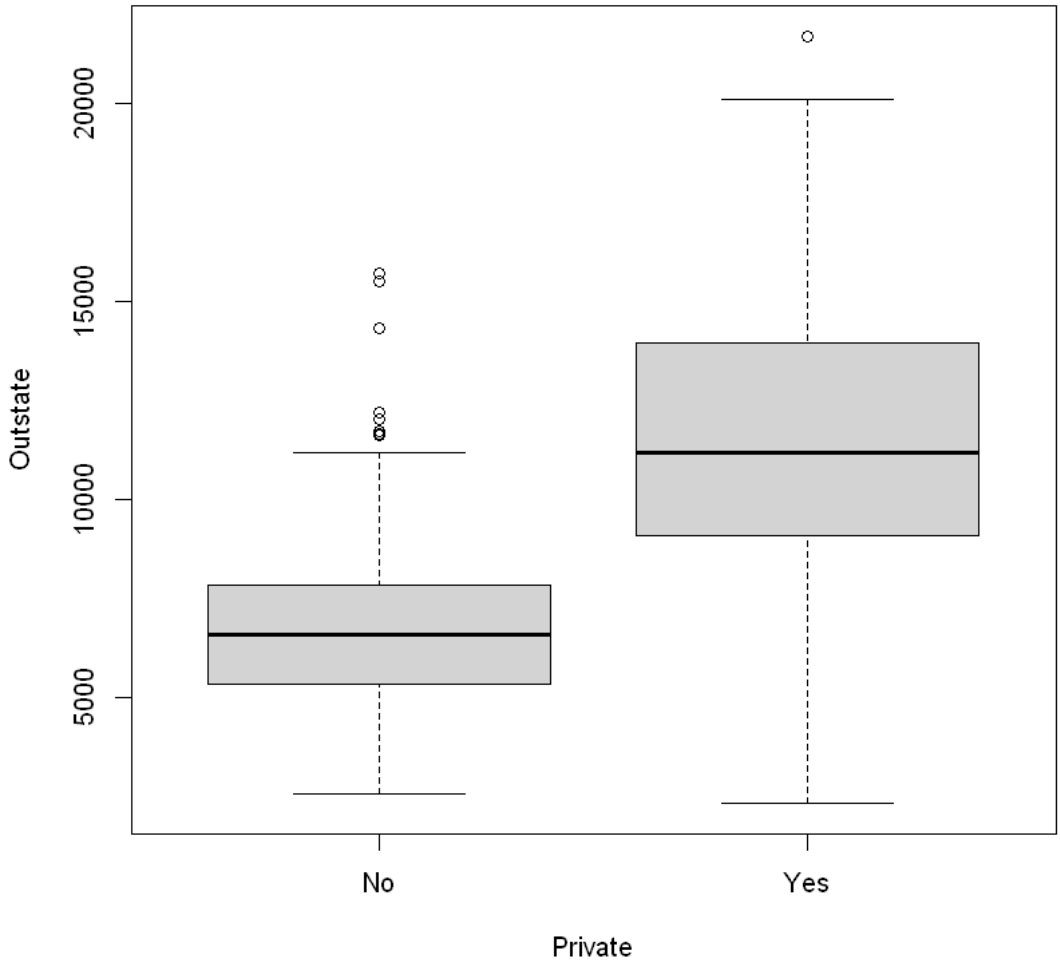
[6]:
```
college[,1] = as.factor(factor(college[,1]))

pairs(college[, 1:10])
```



We produce side-by-side boxplots of **Outstate** versus **Private** columns.

[7]:
```
# Create a boxplot of Outstate versus Private
boxplot(Outstate ~ Private, data = college)
```



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 %.

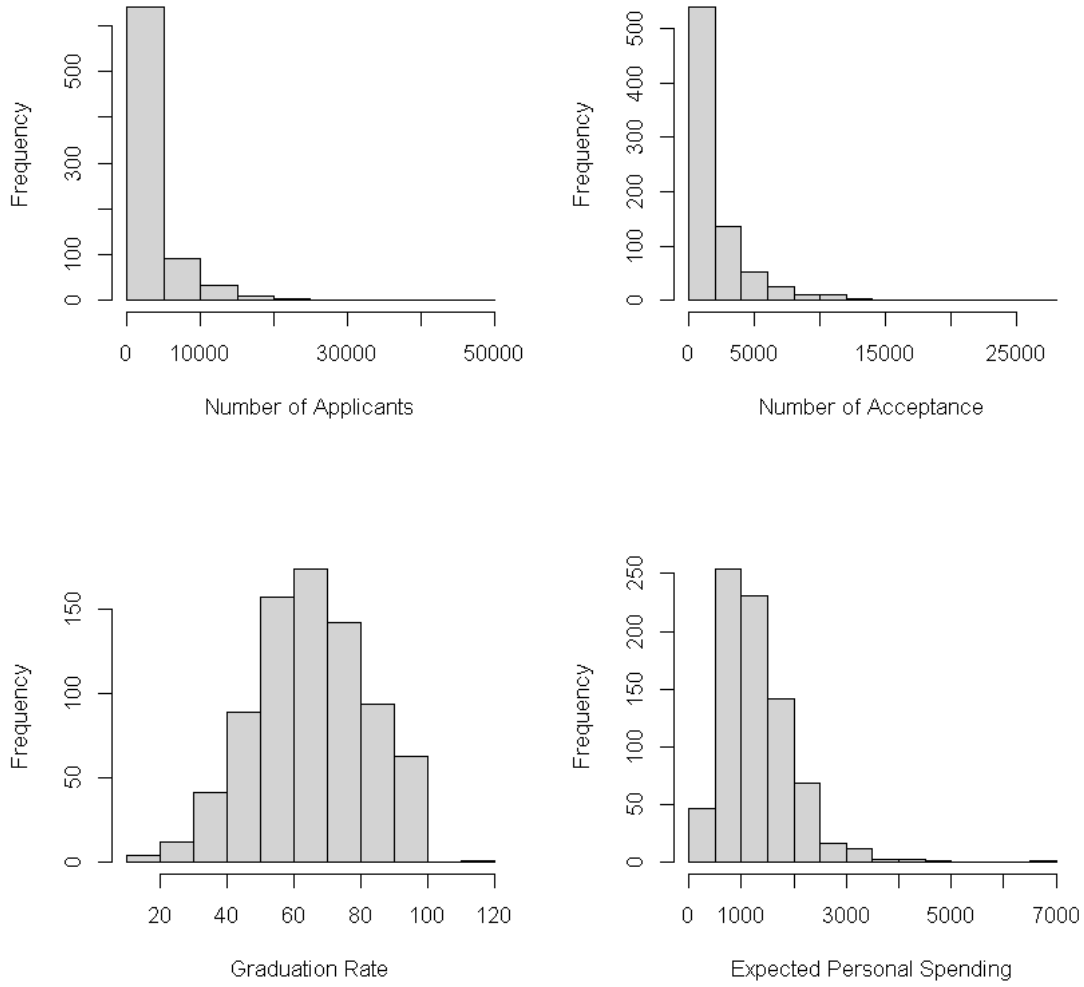
[8]:
```
# Create a new variable column Elite where the top10perc of newly accepted␣
 ↪students reaches more than 50. Set this to Yes and factorize
Elite <- rep("No", nrow(college))
Elite[college$Top10perc > 50] <- "Yes"
Elite <- as.factor(Elite)
college <- data.frame(college , Elite)

# Print the summary of 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      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
```

[10]:
```r
# Create a 2 by 2 plots
par(mfrow = c(2, 2))

# Analyze four of the columns in the dataset by using the histogram function
hist(college$Apps, xlab = "Number of Applicants", main  = "")
hist(college$Accept, breaks = 12, xlab = "Number of Acceptance", main  = "")
hist(college$Grad.Rate, xlab = "Graduation Rate", main = "")
hist(college$Personal, breaks = 12, xlab = "Expected Personal Spending", main =␣
↪"")
```



By exploring the data, I have come up with different questions:

- What are the top 10 universities and colleges with the highest number of new students from top 10% of high school class?

  Masschusetts Institute of Technology is at the top with students belonging to their top 10% class in high school. Following schools are **Harvey Mudd College, University of California Berkeley, Yale University, Duke University, Harvard University, Princeton University, Georgia Insitute of Technology, Brown University, and Dartmouth college**. However, this does not necessarily guarantees that these schools are the best 10 schools in United States.

[23]:
```
# Subset the top 10 universities and college by order
top10 = college[order(college$Top10perc, decreasing = TRUE), ]
```

```
# Print the univerisities or college
head(row.names(top10), 10)
```

1. 'Massachusetts Institute of Technology' 2. 'Harvey Mudd College' 3. 'University of California at Berkeley' 4. 'Yale University' 5. 'Duke University' 6. 'Harvard University' 7. 'Princeton University' 8. 'Georgia Institute of Technology' 9. 'Brown University' 10. 'Dartmouth College'

- What are thetop 10 universities and colleges with the lowest rate of accepted students?

  Princeton University is at the top with students belonging to their top 10% class in high school. Following schools are **Harvard University, Yale University, Amherst College, Brown University, Georgetown University, Dartmouth College, Duke University, Columbia University, and Williams College**. However, this does not necessarily guarantees that these schools are the best 10 schools in United States.

[44]:
```
# Get the acceptance rate column by dividing the number of accepted students to␣
↪the number of applicants
college$accept_rate = college$Accept / college$Apps

# Subset
top10lwst = college[order(college$accept_rate, decreasing = FALSE), ]
head(row.names(top10lwst), 10)
```

1. 'Princeton University' 2. 'Harvard University' 3. 'Yale University' 4. 'Amherst College' 5. 'Brown University' 6. 'Georgetown University' 7. 'Dartmouth College' 8. 'Duke University' 9. 'Columbia University' 10. 'Williams College'

- What are thetop 10 universities and colleges with the highest rate of accepted students?

  Emporia State Univeristy is at the top with students belonging to their top 10% class in high school. Following schools are **Mayville State University, MidAmerica Nazarene College, Southwest Baptist University, University of Wisconsin-Superior, Wayne State College, Arkansas Tech University, Southwestern Adventist College, Pikeville College, and Mount Marty College**. However, this does not necessarily guarantees that these schools are the best 10 schools in United States.

[47]:
```
# Get the acceptance rate column by dividing the number of accepted students to␣
↪the number of applicants
college$accept_rate = college$Accept / college$Apps

# Subset
top10lwst = college[order(college$accept_rate, decreasing = TRUE), ]
head(row.names(top10lwst), 10)
```

1. 'Emporia State University' 2. 'Mayville State University' 3. 'MidAmerica Nazarene College' 4. 'Southwest Baptist University' 5. 'University of Wisconsin-Superior' 6. 'Wayne State College' 7. 'Arkansas Tech University' 8. 'Southwestern Adventist College' 9. 'Pikeville College' 10. 'Mount Marty College'

- What is the top univeristy or college with the highest number of enrolment?

  Rutgers at New Brunswick

[46]:
```
# Get the college or university with the highest number of applications received
row.names(college)[which.max(college$Apps)]
```

'Rutgers at New Brunswick'

We get the relationship of Outstate tuition and the Graduation Rate. The figure shows that the higher the outstate tuition, the higher the rate of graduation. However, it does not necessarily guarantees correlation nor causation.

[50]:
```
plot(Grad.Rate ~ Outstate, data = college)
```