Tutorial 1 Solutions

STAT 4040/7040

1. [ISL] Q 2.8 solutions:

   a. As this data is part of the ISLR package, I will load in the data from that package. It is good practice to use the *rm(list = ls())* to clear the work-space (working memory) so variables are not overwritten.

   ```
   rm(list = ls())
   library(ISLR)
   data(College)
   college <- College
   ```

   b. I used the *head()* command to look at the first six rows of the data. You can also try using *fix()* to examine the data. Note that the row names already are the college names! They must have made some changes to the files since the textbook!

```
head(college)
```

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

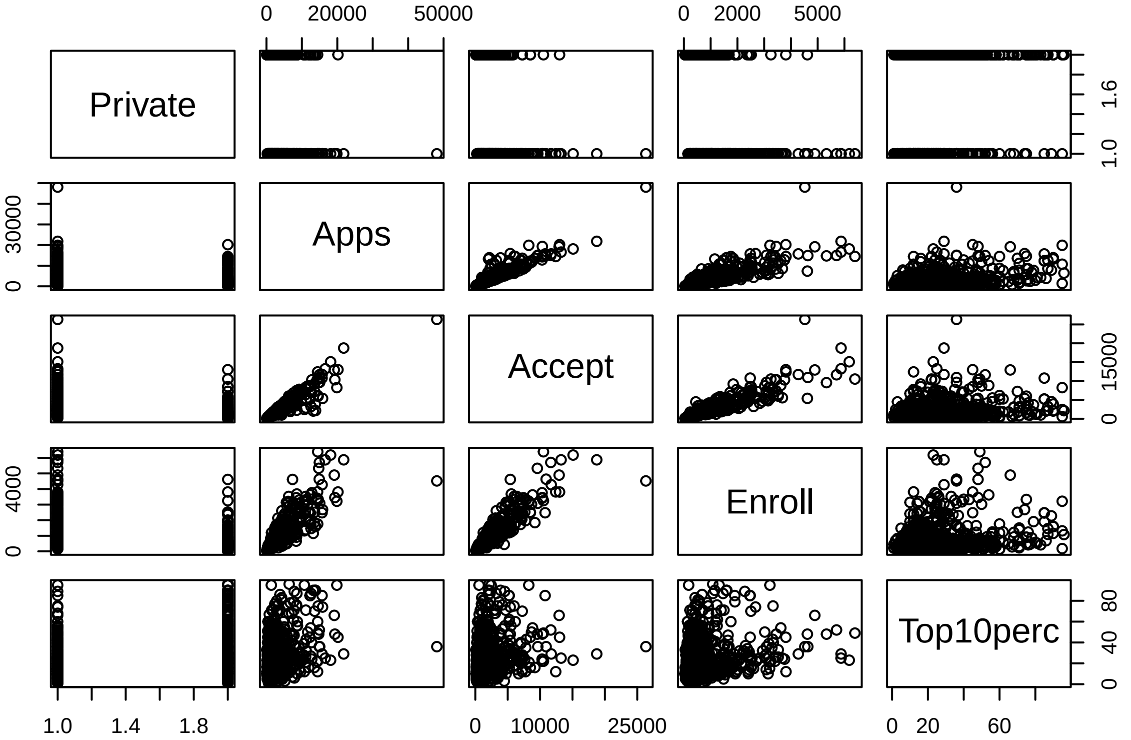
```
## Abilene Christian University        60
## Adelphi University                  56
## Adrian College                      54
## Agnes Scott College                 59
## Alaska Pacific University           15
## Albertson College                   55
```

c. Let's do some investigation of the data through summary statistics and visualizations. For the *pairs()* command I just used the first 5 variables, as the the scatter plots become quite small for viewing on paper, but on a computer screen you can "blow-up" the graph so you can examine more variables.

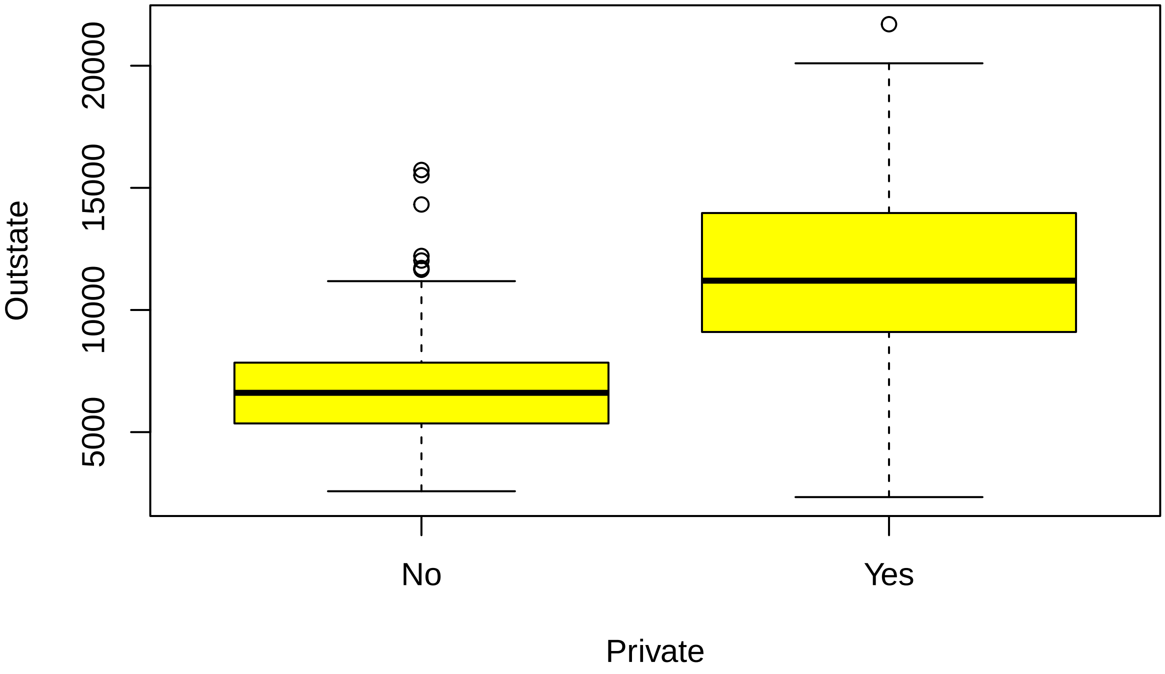
```
summary(college)
```

```
##  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
```

```
pairs(college[,1:5])
```



```
plot(college$Outstate ~ college$Private, col="yellow", xlab="Private",
     ylab="Outstate")
```



Let's create the new variable *Elite*:

```
Elite  <- rep ("No",nrow(college))
Elite [college$Top10perc >50]="Yes"
Elite  <- as.factor (Elite)
college <- data.frame(college, Elite)
summary(Elite)
```

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

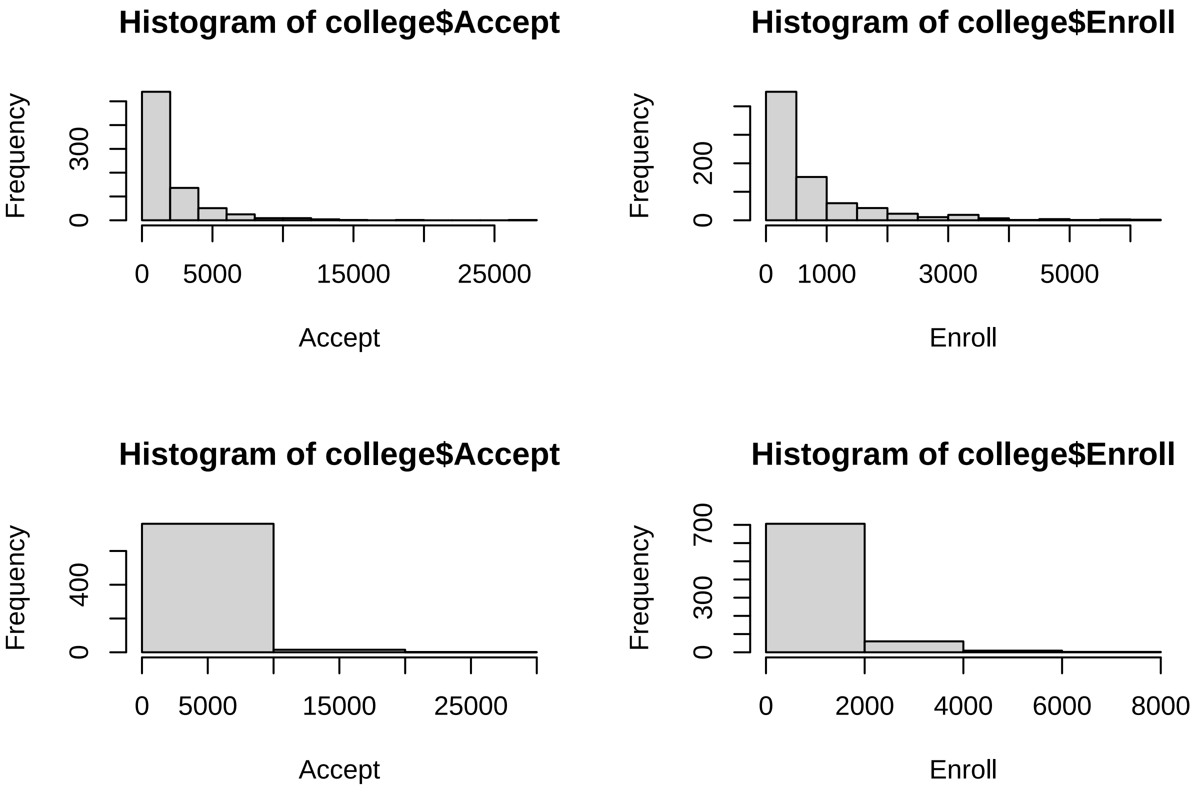
From the summary, we see that there are 78 elite universities in the data set. Let's look at the first six elite universities:

```
rownames(college[college$Elite=="Yes",])[1:6]
```

```
## [1] "Agnes Scott College"          "Amherst College"
## [3] "Barnard College"              "Birmingham-Southern College"
## [5] "Bowdoin College"              "Brown University"
```

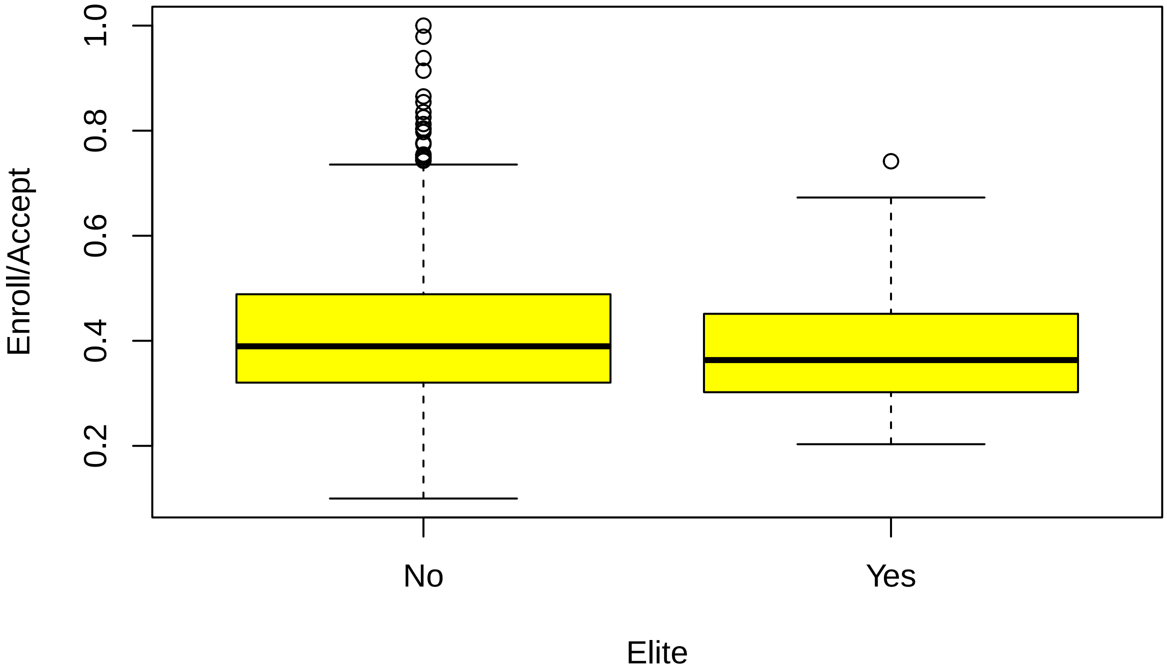
Let's look at histograms of the number of students accepted and the number enrolled. Both variables are quite right skewed.

```
par(mfrow=c(2,2))
hist(college$Accept, xlab="Accept")
hist(college$Enroll, xlab="Enroll")
hist(college$Accept, nclass=3, xlab="Accept")
hist(college$Enroll, nclass=3, xlab="Enroll")
```



I prefer to use the default number of bins that R provides. We can also specify exact break points for the bins. See *help(hist)* for more details. Finally let's examine the ratio *Enroll/Accept* (probability of enrollment) against whether the university is elite or not.

```
plot(college$Enroll/college$Accept ~ college$Elite, col="yellow",
     xlab="Elite", ylab="Enroll/Accept")
```



It seems that the probability of enrollment is slightly lower for elite universities. This may be due to the cost of some of these universities compared to non-elite ones.

2. [ISL] Q 2.8 plots via *ggplot2*:

The package *ggplot2* is part of the *Tidyverse* collection of R-packages https://www.tidyverse.org. Note, that for ANU computers, these packages may not be installed and you cannot install them. However, if you have your own machine you can install them. You only have to install them once. To install just the *ggplot2* package and an extension package use:

```
install.packages("ggplot2")
install.packages("GGally") # an extension to ggplot2 - needed for the ggpairs
```

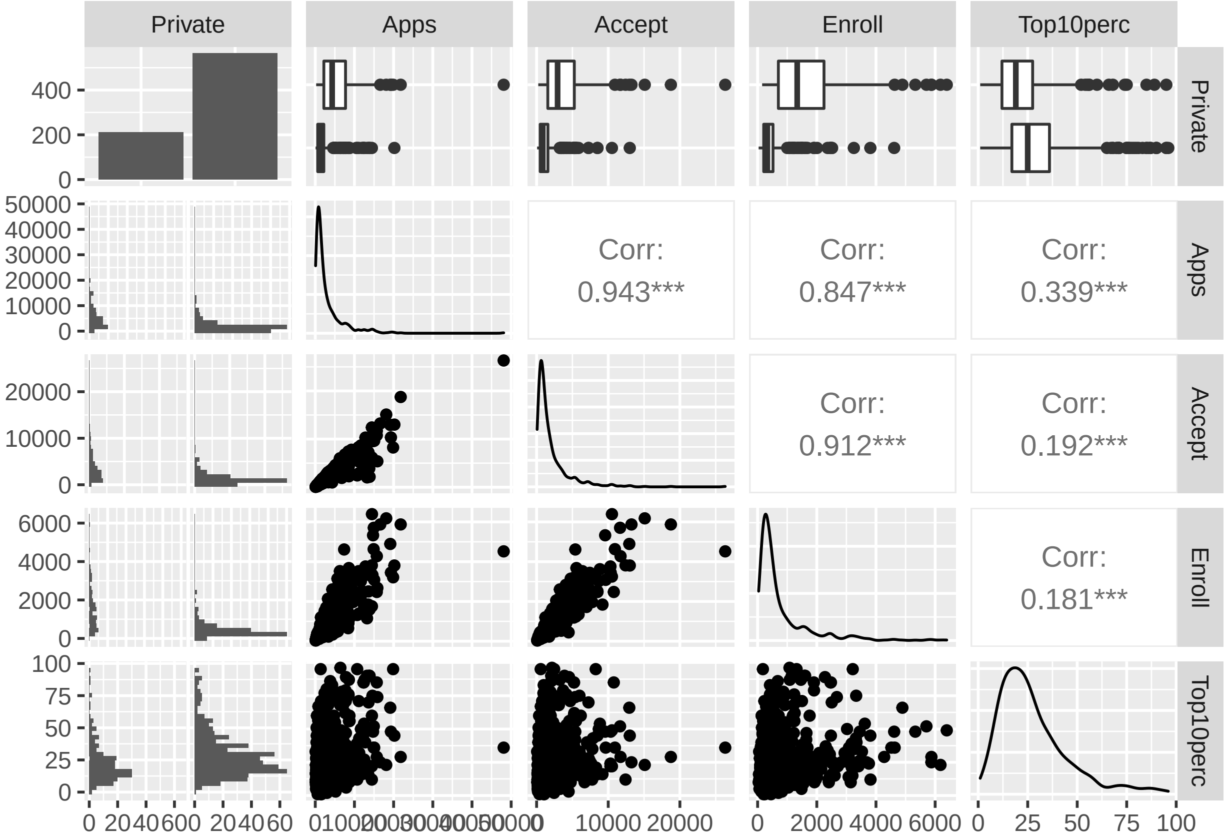
```
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

- Let's make a pairs plot

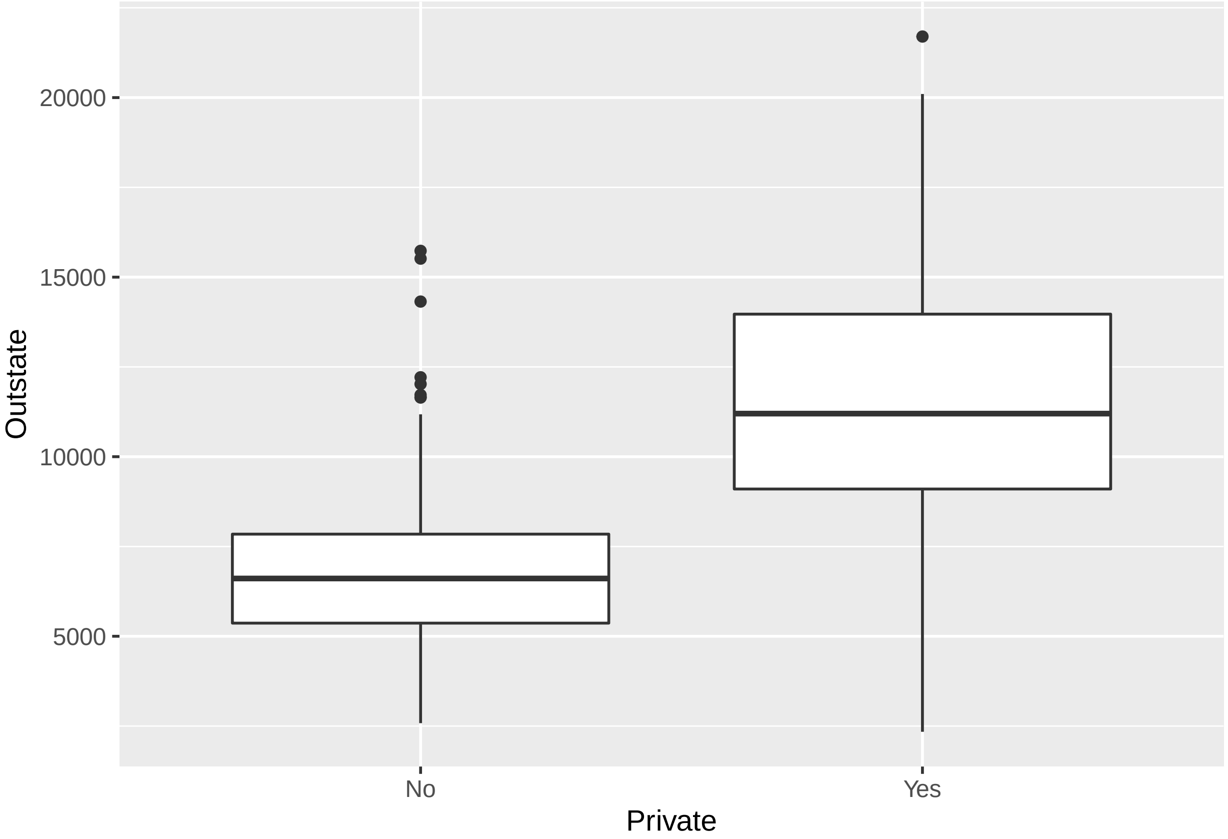
```
ggpairs(college, columns=1:5)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



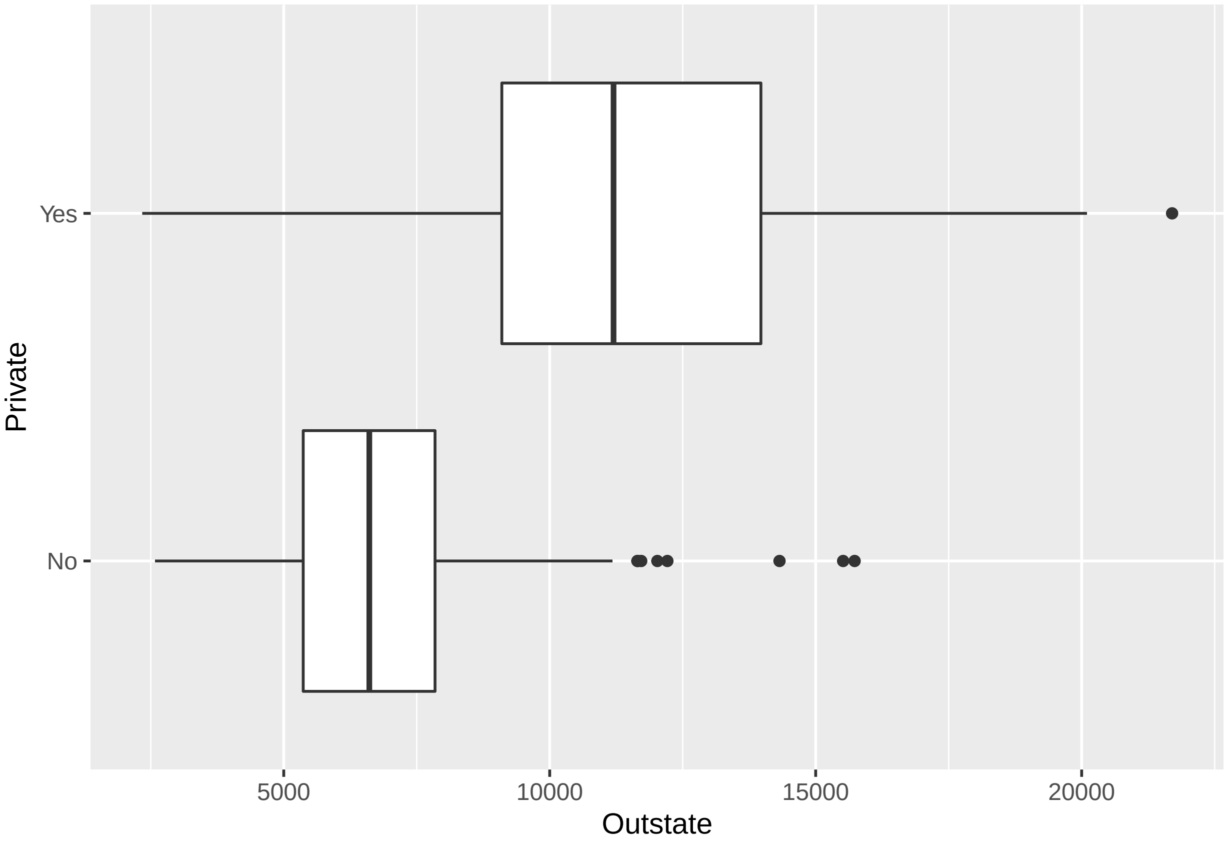
- Let's make the boxplot (https://ggplot2.tidyverse.org/reference/geom_boxplot.html)

```r
p <- ggplot(college, aes(Private, Outstate))
p + geom_boxplot()
```



- Note that *aes()* will be for the *aes(x-axis, y-axis)*. Orientation follows the discrete axis (i.e. the factor variable).

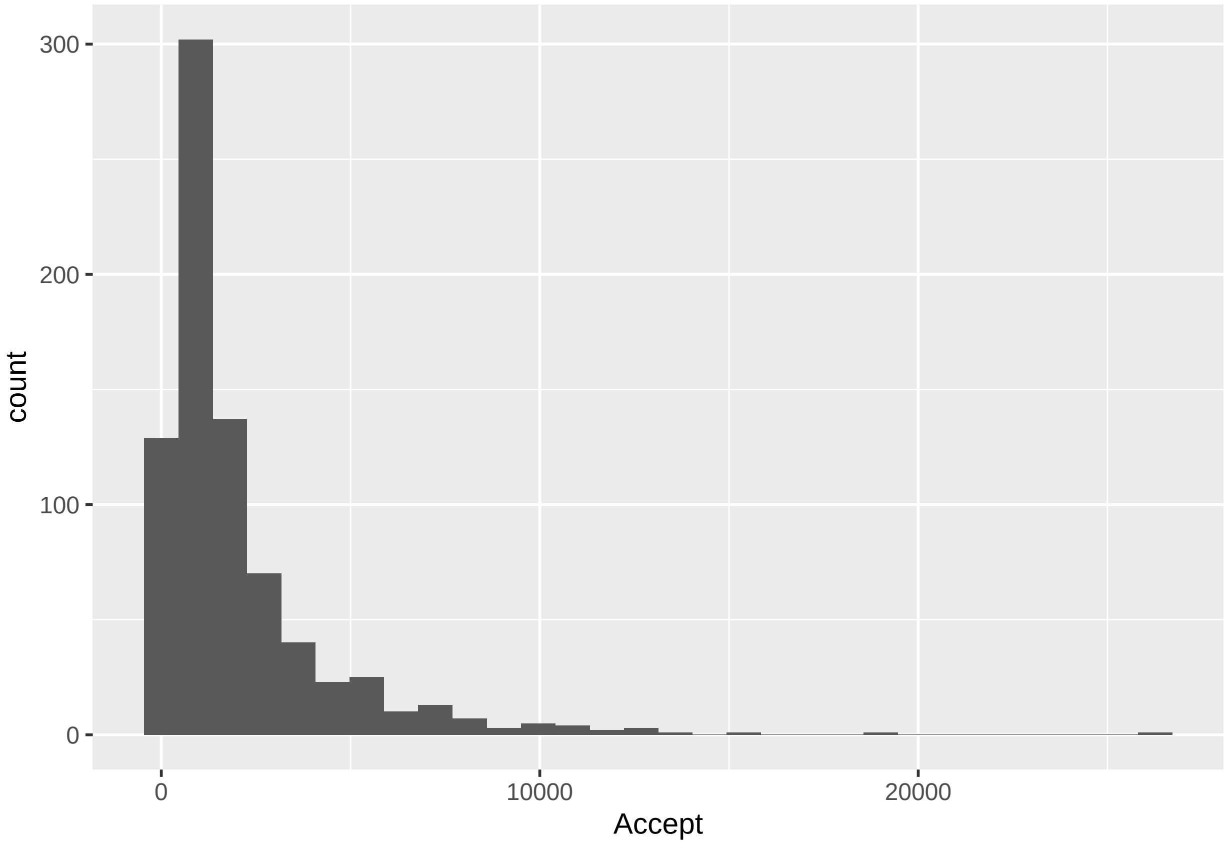
```
p <- ggplot(college, aes(Outstate, Private))
p + geom_boxplot()
```



- The two histograms (https://ggplot2.tidyverse.org/reference/geom_histogram.html):

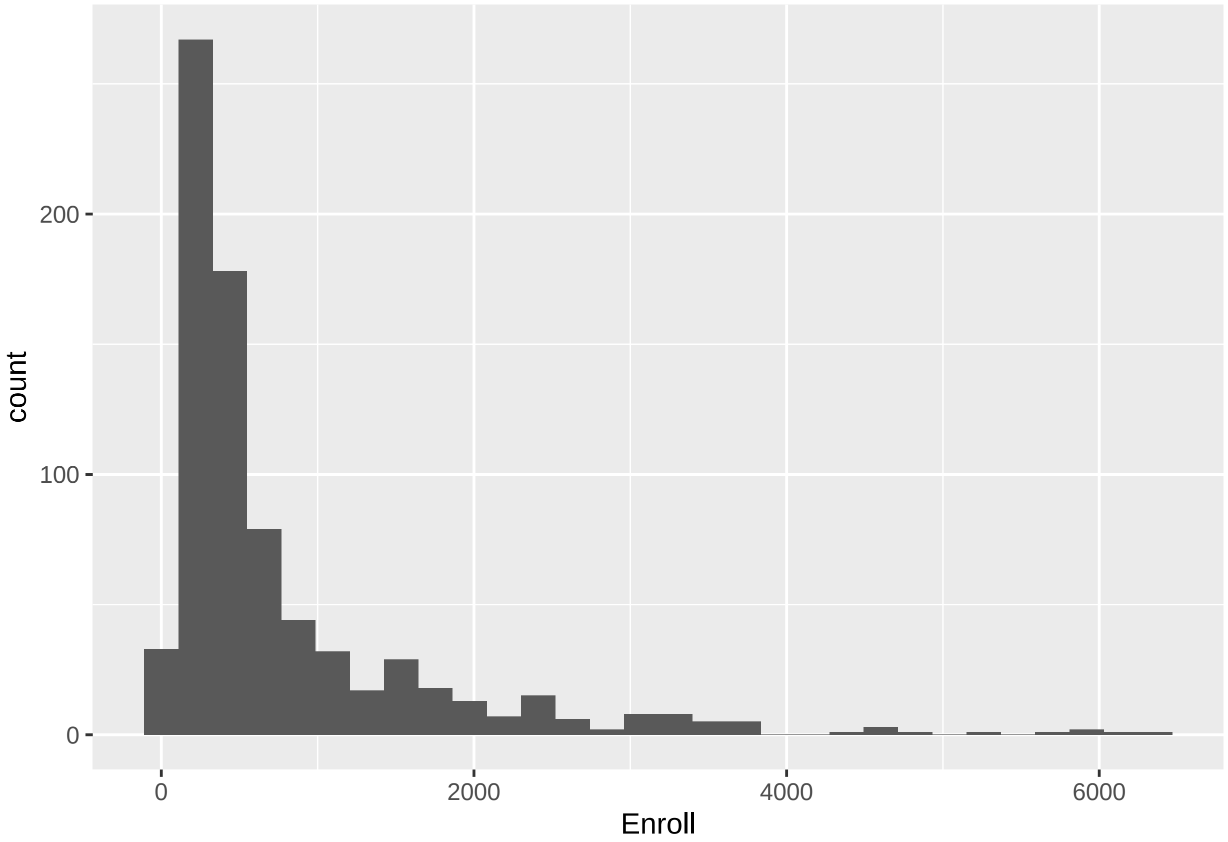
```
p2 <- ggplot(college, aes(Accept))
p2 + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```r
p3 <- ggplot(college, aes(Enroll))
p3 + geom_histogram()
```

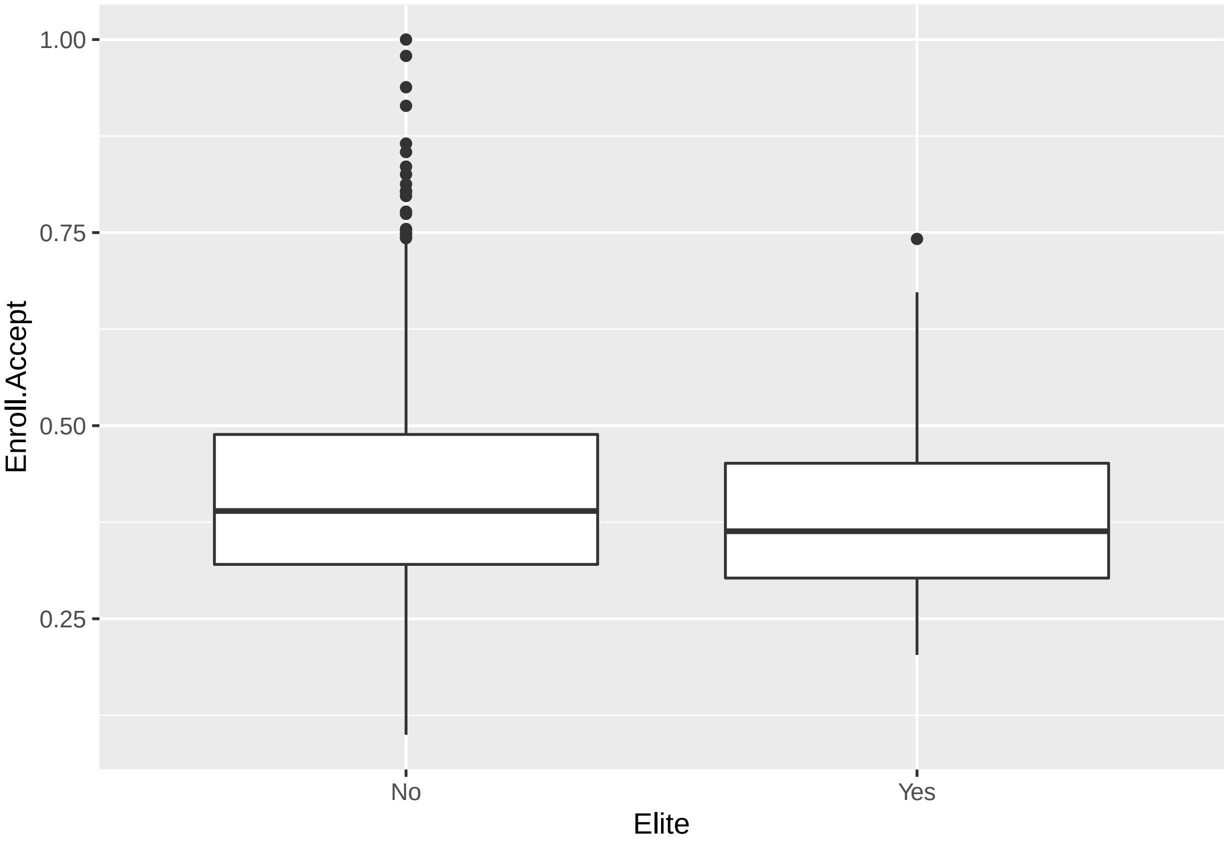
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



- Add the ratio to the data frame and change the name. Then make the plot:

```
college <- data.frame(college, college$Enroll/college$Accept)
names(college)[20] <- "Enroll.Accept"

p4 <- ggplot(college, aes(Elite, Enroll.Accept))
p4 + geom_boxplot()
```


1.00
0.75
0.50
0.25
Enroll.Accept
No
Yes
Elite