STATS415_HW1_Xinye Xu

Q3

Numerical summarys: From the dataset of College.csv, there are 19 variables and 777 observations. Most universiteis are private (565). From Apps, the average Number of applications received is 3002, accepted is 2019, number of new enrolll is 780. The mean of new students from top 10 % of high school class is 27.56, and 55.8 for top 25%. Mean of number of full-time undergraduates = 3700, 855.3 for part-time. As for the mean of different costs, the mean of Out-of-state tuition is $10441, Room and board costs is 4358, books is 549.4 and 1341 for personal. 79% for the mean of percent of faculty with terminal degree. The mean of Student/faculty ratio is 14.09, percent of alumni who donate is 22.74%.And mean of Instructional expenditure per student is 9660, Graduation rate is 65.46%.

```
College = read.csv('http://www-bcf.usc.edu/~gareth/ISL/College.csv', header = T, na.strings = '?')
# str(College)
colMeans(College[,-c(1,2)]) # summary(College) too much inform
```

```
##        Apps      Accept      Enroll   Top10perc   Top25perc F.Undergrad
##  3001.63835  2018.80438   779.97297    27.55856    55.79665  3699.90734
## P.Undergrad    Outstate  Room.Board       Books    Personal         PhD
##   855.29858 10440.66924  4357.52638   549.38095  1340.64221    72.66023
##    Terminal   S.F.Ratio perc.alumni      Expend   Grad.Rate
##    79.70270    14.08970    22.74389  9660.17117    65.46332
```

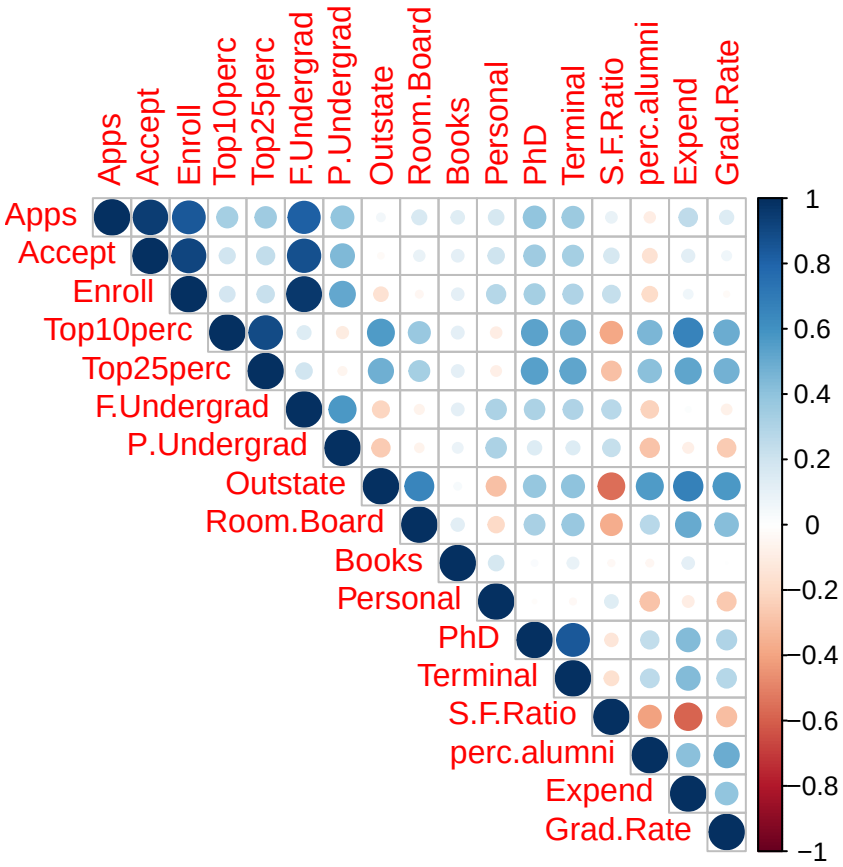
Based on the correlation plot below, Apps, Accept, Enroll and F.Undergrad are highly postive correlated obviously, and its cor(Apps, Accept) = 0.94. cor(Enroll, Accept) = 0.85 tions), cor(Enroll, F.Undergrad) = 0.96. Plus, Top10perc and Top25perc are highly posi related, PhD and S.F.Ratio are also related heavily. Intereting thing is taht Top10perc, Expend and Outstate are also related.cor(Top10perc, Expend) = 0.66, cor(Outstate, Expend) = 0.67. It may indicate that schools with more students from Top10perc high school have higer Outstate tuition but they spend more for Instructional expenditure per student.

```
corr <- cor(College[,-c(1,2)]) # exclude factors in first and second column
# round(corr, 2) # matrix directly
require("corrplot")
```

```
## Loading required package: corrplot
```

```
## corrplot 0.84 loaded
```

```
corrplot(corr, type = "upper")
```



From the histograms plot, expect for Top25perc, which seems to be quite normal, other plots are either left or right skewed. We look further to the Top10perc including the factor of Private. From the boxplot, it suggests Private school might have large number of outliers which are laying larger than the 75% quantile of the Top10perc. From the histogram, it is right skewed so it does not suggest a symmetric norml distribution.

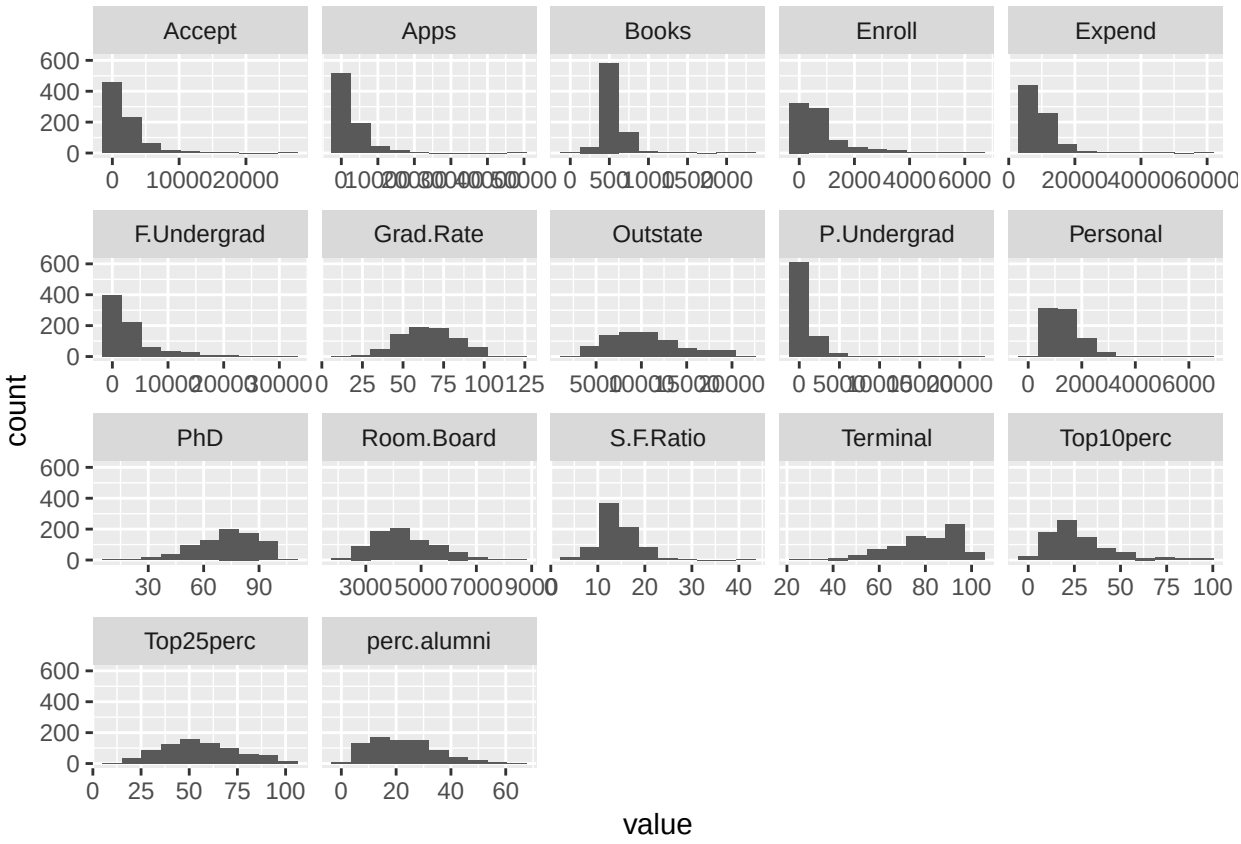
```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.4.4
```

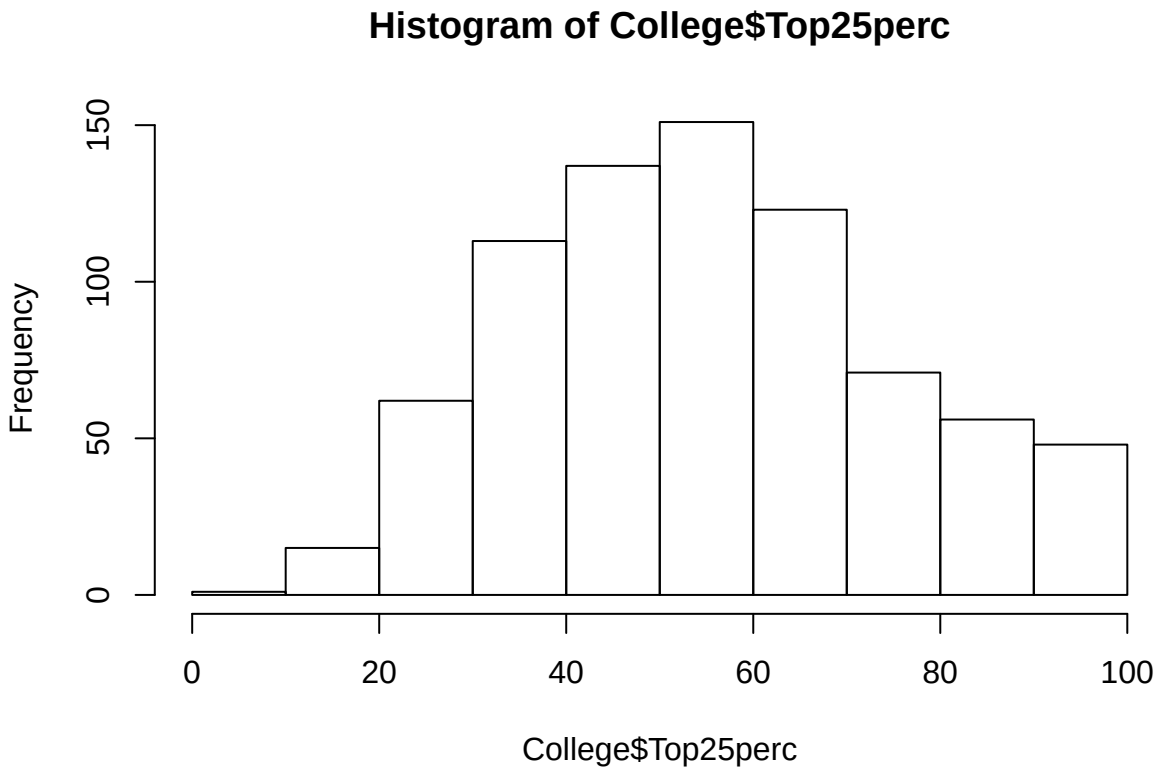
```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

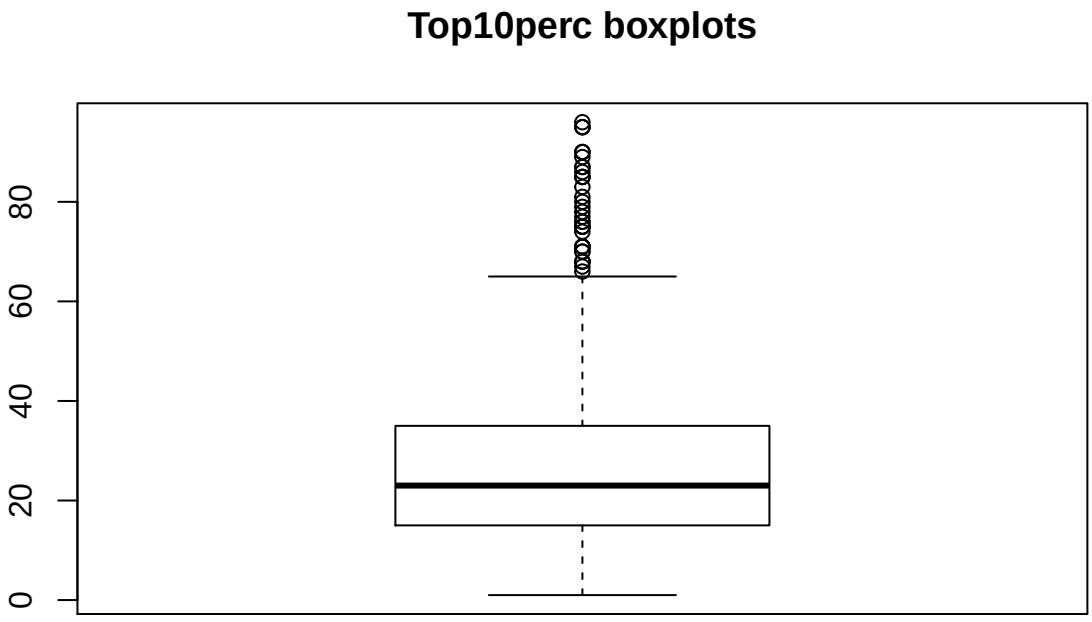
```
ggplot(gather(College[,-c(1,2)]), aes(value)) +
    geom_histogram(bins = 10) +
    facet_wrap(~key, scales = 'free_x')
```



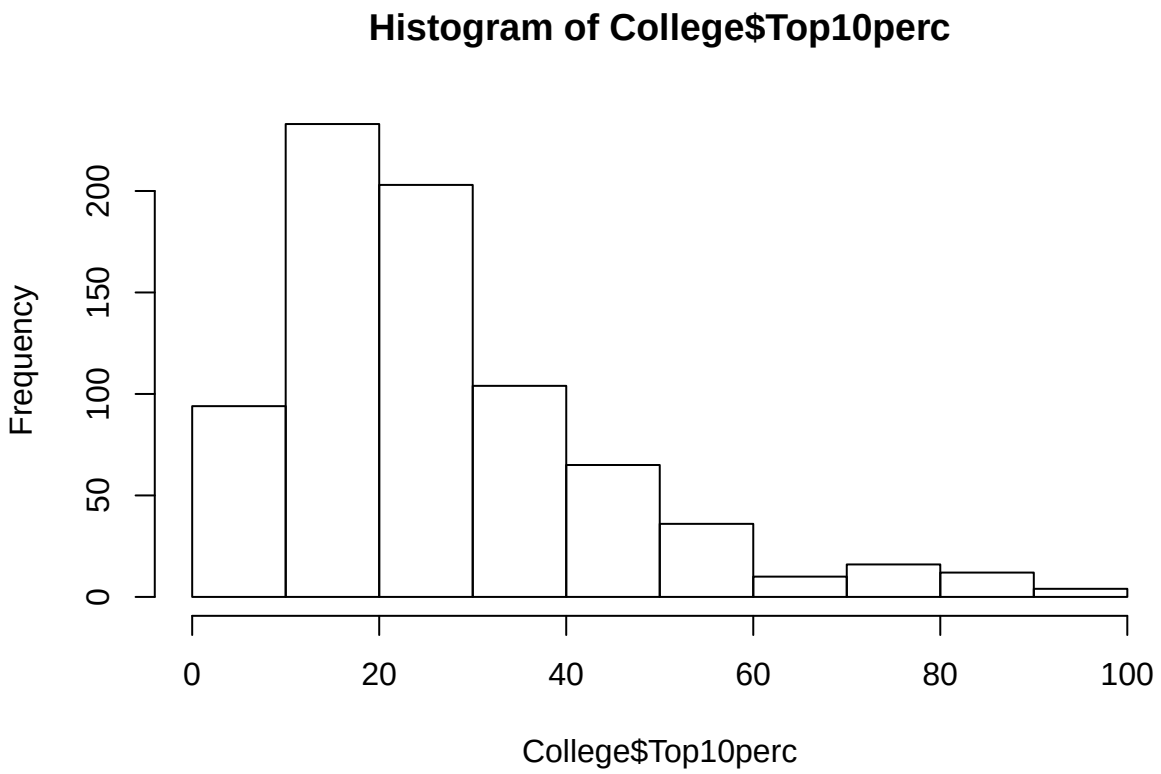

```
# The scales = 'free_x' is necessary unless your data is all of a similar scale.

hist(College$Top25perc)
```




```
boxplot(College$Top10perc, main = 'Top10perc boxplots')
```




```
hist(College$Top10perc)
```



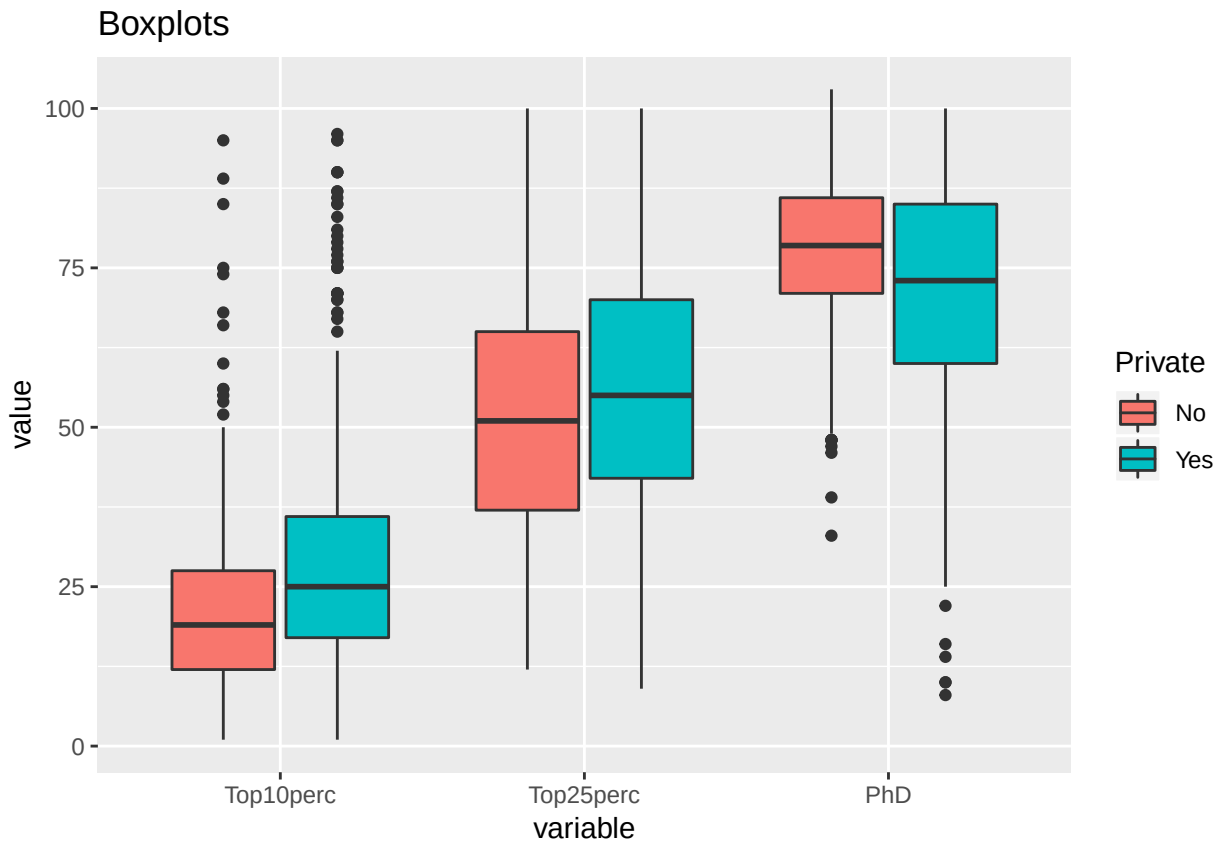
From the side-by-side boxplots, it suggests Private school might have higher Top10perc and Top25perc rates.But for the Phd rate, private school seems have less percentage of people holding phd. We look further to the relationship between Outstate and Expend. We can see clear trends that higher Outstate tuition can bring hgiher instructional expenditure for student. Also, from the scatterplot, private school usually means higher instructional expenditure.

```
require(ggplot2)
library(reshape2) # multi side-by-side boxplot
```

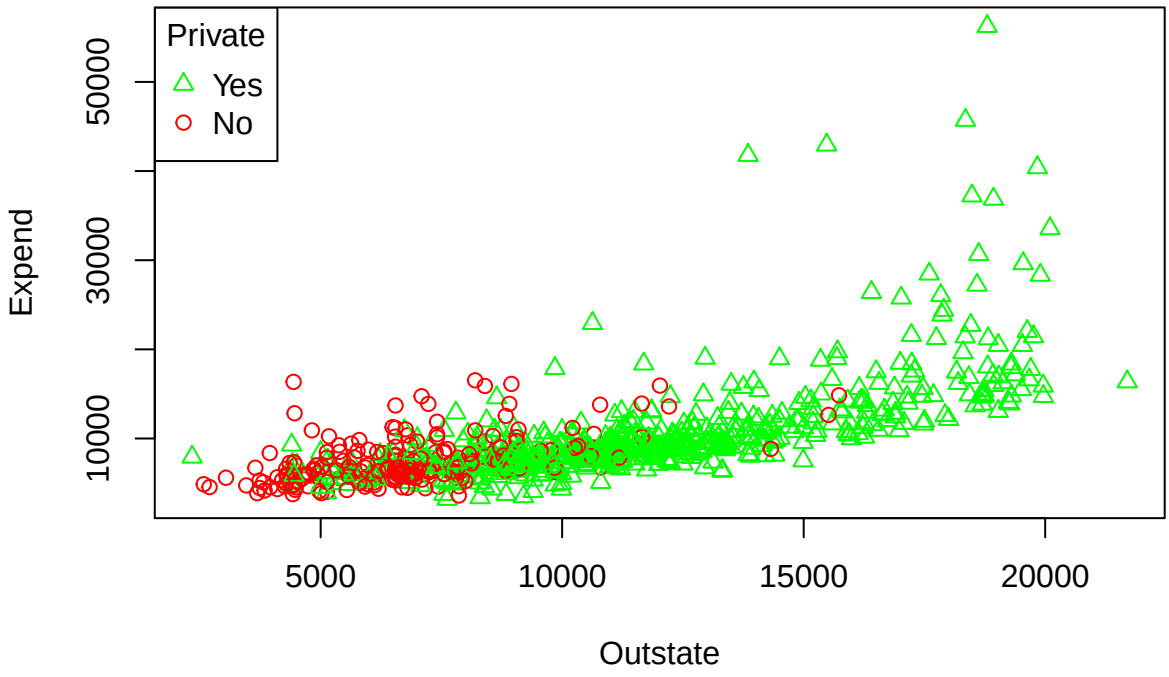
```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
df2<- melt(College[,c('Private','Top10perc','Top25perc','PhD')],id.var=c("Private"))
ggplot(df2, aes(variable, value)) + geom_boxplot(aes(fill=Private)) + labs(title = "Boxplots")
```




```
# boxplot(College$Top10perc ~ College$Private, main = 'Top10perc boxplots')
plot(College$Outstate, College$Expend, xlab = 'Outstate', ylab = 'Expend',
     main = 'Outstate vs Expend',
     col = c('red', 'green')[College$Private],
     pch = c(1:2)[College$Private])
legend('topleft', legend = unique(College$Private),
        title = 'Private',
        col = c('red', 'green')[unique(College$Private)],
        pch = c(1:2)[unique(College$Private)])
```




```
lm_m <- lm(Expend ~ Outstate, data = College)
summary(lm_m)$coef
```

```
##                Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept) 542.8696946 385.92914940  1.406656  1.599302e-01
## Outstate      0.8732488   0.03449491 25.315293 1.629891e-103
```