Homework 1 Solutions

Weijing Tang

January 21, 2018

Problem 1

- (a) Examples:
  - Categorical variable: age above 21 or below 21
  - Ordinal variable: school year, i.e. freshman, sophomore, junior, senior, graduate (must have order)
  - Interval variable: birthdate
  - Ratio variable: age
- (b) The students who have taken Stats 415 in the University of Michigan.
- (c) The students who are enrolled in US college.

Problem 2

- (a)
  - For a term j that occurs in every document, $g_j = n$, then $f^\star_{ij} = f_{ij} \log(\frac{n}{n}) = 0$. This transformation makes all such terms' weights to be 0.
  - For a term j that occurs in only one document, $g_j = 1$, then $f^\star_{ij} = f_{ij} \log(\frac{n}{1}) = f_{ij} \log n$. This transformation makes all such terms' weights to increase by a factor of $\log n$.
- (b) This transformation reflects the observation with terms that occur in every document do not have any power to distinguish one document from another, while those that are relatively rare do.

Problem 3

```
college = read.csv("College.csv")
rownames(college) = college[,1]
college = college[,-1]
# head(college)
```

We can get a good six number summary of the data using the `summary` command.

```
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 range of number of part-time undergraduates is from 1 to 21836. The mean graduation rate is 65.46%. Note that there exists one college whose graduation rate is more than 100, We would better double check how they calculate the graduation rate and keep an eye on it.

```
summary(college$Accept/college$Apps)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1545  0.6756  0.7788  0.7469  0.8485  1.0000
```

```
college[which.max(college$Accept/college$Apps),]
```

```
##                          Private Apps Accept Enroll Top10perc Top25perc
## Emporia State University      No 1256   1256    853        43        79
##                          F.Undergrad P.Undergrad Outstate Room.Board Books
## Emporia State University        3957         588     5401       3144   450
##                          Personal PhD Terminal S.F.Ratio perc.alumni
## Emporia State University     1888  72       75      19.3           4
##                          Expend Grad.Rate
## Emporia State University   5527        50
```

```
college[which.min(college$Accept/college$Apps),]
```

```
##                       Private  Apps Accept Enroll Top10perc Top25perc
## Princeton University      Yes 13218   2042   1153        90        98
##                       F.Undergrad P.Undergrad Outstate Room.Board Books
## Princeton University        4540         146    19900       5910   675
##                       Personal PhD Terminal S.F.Ratio perc.alumni Expend
## Princeton University     1575  91       96       8.4          54  28320
##                       Grad.Rate
## Princeton University        99
```

Emporia State University has the highest rate of applicants accepted, which received 1256 applications in total and accepted all the applications, while Princeton University has the lowest rate, 0.1545.

We use `cor` to calculate pairwise correlations. The inputs must be numeric.

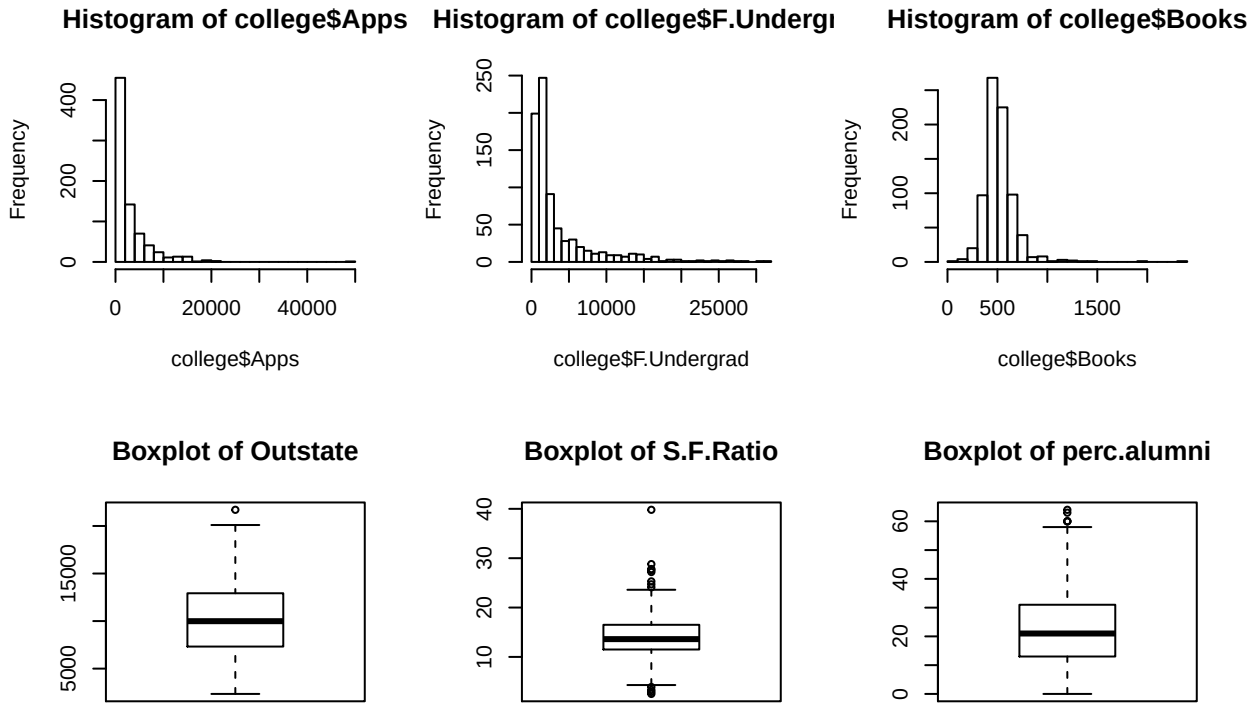
```
round(cor(college[,-1]),2)
```

```
##              Apps Accept Enroll Top10perc Top25perc F.Undergrad
## Apps         1.00   0.94   0.85      0.34      0.35        0.81
## Accept       0.94   1.00   0.91      0.19      0.25        0.87
## Enroll       0.85   0.91   1.00      0.18      0.23        0.96
## Top10perc    0.34   0.19   0.18      1.00      0.89        0.14
## Top25perc    0.35   0.25   0.23      0.89      1.00        0.20
## F.Undergrad  0.81   0.87   0.96      0.14      0.20        1.00
## P.Undergrad  0.40   0.44   0.51     -0.11     -0.05        0.57
## Outstate     0.05  -0.03  -0.16      0.56      0.49       -0.22
## Room.Board   0.16   0.09  -0.04      0.37      0.33       -0.07
## Books        0.13   0.11   0.11      0.12      0.12        0.12
## Personal     0.18   0.20   0.28     -0.09     -0.08        0.32
## PhD          0.39   0.36   0.33      0.53      0.55        0.32
## Terminal     0.37   0.34   0.31      0.49      0.52        0.30
## S.F.Ratio    0.10   0.18   0.24     -0.38     -0.29        0.28
## perc.alumni -0.09  -0.16  -0.18      0.46      0.42       -0.23
## Expend       0.26   0.12   0.06      0.66      0.53        0.02
## Grad.Rate    0.15   0.07  -0.02      0.49      0.48       -0.08
##              P.Undergrad Outstate Room.Board Books Personal   PhD Terminal
## Apps                0.40     0.05       0.16  0.13     0.18  0.39     0.37
## Accept              0.44    -0.03       0.09  0.11     0.20  0.36     0.34
## Enroll              0.51    -0.16      -0.04  0.11     0.28  0.33     0.31
## Top10perc          -0.11     0.56       0.37  0.12    -0.09  0.53     0.49
## Top25perc          -0.05     0.49       0.33  0.12    -0.08  0.55     0.52
## F.Undergrad         0.57    -0.22      -0.07  0.12     0.32  0.32     0.30
## P.Undergrad         1.00    -0.25      -0.06  0.08     0.32  0.15     0.14
## Outstate           -0.25     1.00       0.65  0.04    -0.30  0.38     0.41
## Room.Board         -0.06     0.65       1.00  0.13    -0.20  0.33     0.37
## Books               0.08     0.04       0.13  1.00     0.18  0.03     0.10
## Personal            0.32    -0.30      -0.20  0.18     1.00 -0.01    -0.03
## PhD                 0.15     0.38       0.33  0.03    -0.01  1.00     0.85
## Terminal            0.14     0.41       0.37  0.10    -0.03  0.85     1.00
## S.F.Ratio           0.23    -0.55      -0.36 -0.03     0.14 -0.13    -0.16
## perc.alumni        -0.28     0.57       0.27 -0.04    -0.29  0.25     0.27
## Expend             -0.08     0.67       0.50  0.11    -0.10  0.43     0.44
## Grad.Rate          -0.26     0.57       0.42  0.00    -0.27  0.31     0.29
##              S.F.Ratio perc.alumni Expend Grad.Rate
## Apps              0.10       -0.09   0.26      0.15
## Accept            0.18       -0.16   0.12      0.07
## Enroll            0.24       -0.18   0.06     -0.02
## Top10perc        -0.38        0.46   0.66      0.49
## Top25perc        -0.29        0.42   0.53      0.48
## F.Undergrad       0.28       -0.23   0.02     -0.08
## P.Undergrad       0.23       -0.28  -0.08     -0.26
## Outstate         -0.55        0.57   0.67      0.57
## Room.Board       -0.36        0.27   0.50      0.42
## Books            -0.03       -0.04   0.11      0.00
## Personal          0.14       -0.29  -0.10     -0.27
```

```
## PhD             -0.13        0.25   0.43      0.31
## Terminal        -0.16        0.27   0.44      0.29
## S.F.Ratio        1.00       -0.40  -0.58     -0.31
## perc.alumni     -0.40        1.00   0.42      0.49
## Expend          -0.58        0.42   1.00      0.39
## Grad.Rate       -0.31        0.49   0.39      1.00
```

From the pairwise correlation matrix above, we find that the number of applications received, accepted, the number of new students enrolled and the number of full-time undergraduates are pairwise highly positive-correlated. Also the correlation between the percent of faculty with Ph.D.'s and the percent of faculty with terminal degree is around 0.85.

```
par(mfrow=c(2,3))
hist(college$Apps,breaks = 30)
hist(college$F.Undergrad, breaks = 30)
hist(college$Books, breaks = 30)
boxplot(college$Outstate, main = "Boxplot of Outstate")
boxplot(college$S.F.Ratio, main = "Boxplot of S.F.Ratio")
boxplot(college$perc.alumni, main = "Boxplot of perc.alumni")
```



From the histograms above, the distribution of the number of applications received is right-skewed and so is the distribution of the number of full-time undergraduates. And there is no obvious skewness of the distribution of estimated personal spending. There are some possible outliers showed on the boxplot of Student/faculty ratio.

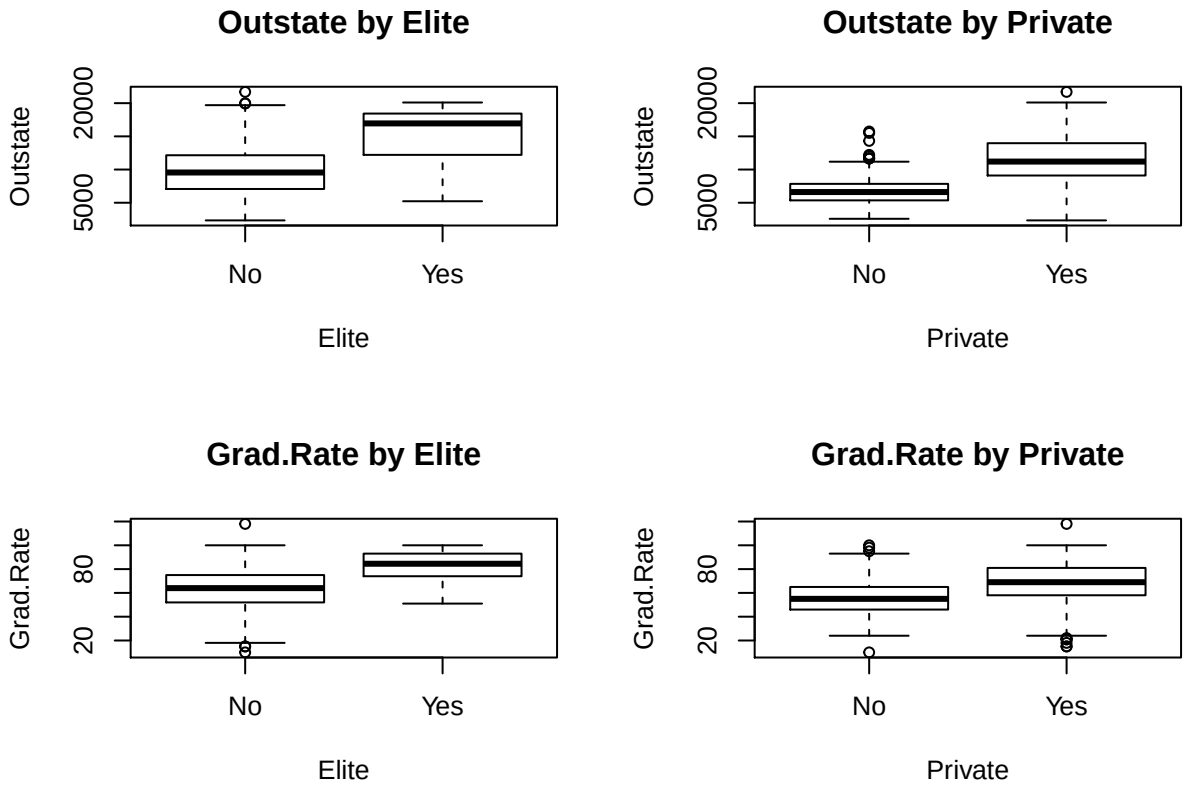
```
Elite = rep('No', nrow(college))
Elite[college$Top10perc >50] = 'Yes'
Elite = as.factor(Elite)
```

```
college = data.frame(college, Elite)
summary(college$Elite)
```

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

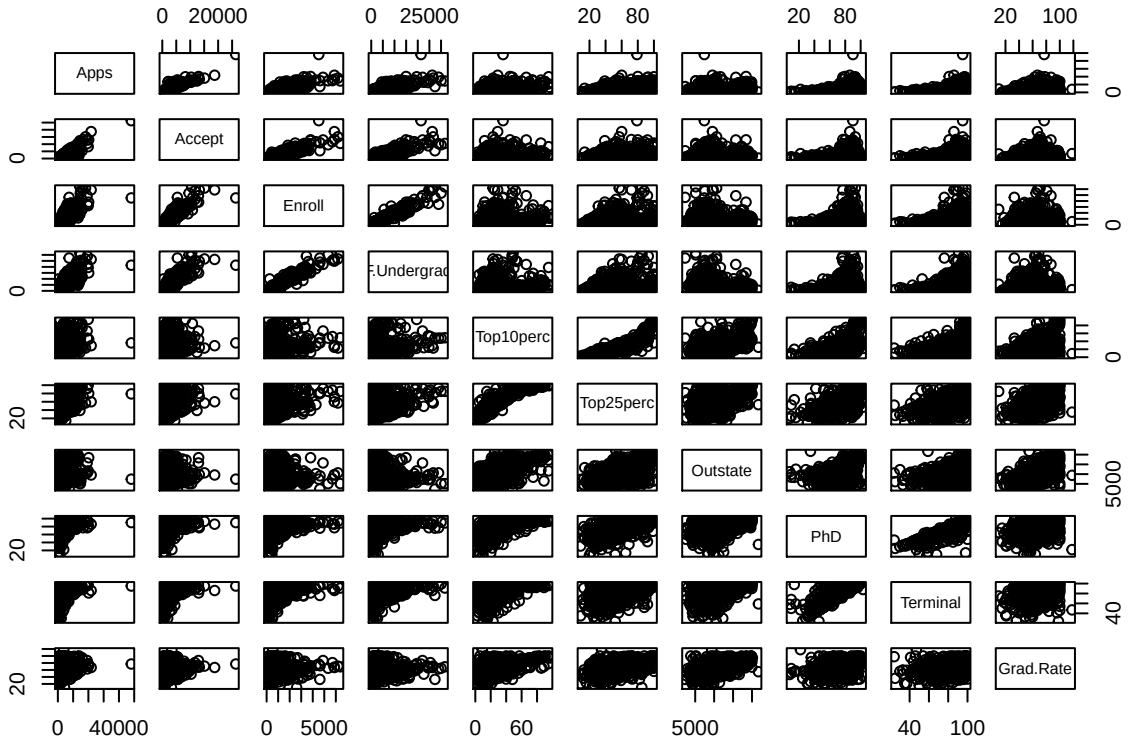
There are 78 colleges whose proportion of students coing from the top 10% of their high school classes exceeds 50%.

```
par(mfrow=c(2,2))
plot(college$Outstate~college$Elite, main = 'Outstate by Elite',
     ylab = 'Outstate', xlab = 'Elite')
plot(college$Outstate~college$Private,main = 'Outstate by Private',
     ylab = 'Outstate', xlab = 'Private')
plot(college$Grad.Rate~college$Elite, main = 'Grad.Rate by Elite',
     ylab = 'Grad.Rate', xlab = 'Elite')
plot(college$Grad.Rate~college$Private, main = 'Grad.Rate by Private',
     ylab = 'Grad.Rate', xlab = 'Private')
```



There is a huge difference between the median and the range variation of Out-of-State tuition of Private and Public college. Also the median and the range variation of graduation rate of the colleges, whose proportion of students coming from the top 10% of their high school classes exceeds 50%, are higher than that of other colleges.

```
pairs(subset(college, select = c("Apps","Accept","Enroll","F.Undergrad","Top10perc",
                                 "Top25perc","Outstate","PhD","Terminal","Grad.Rate")))
```



Top10perc, Top25perc and Outstate seem to be useful predictors of Grad.Rate. Some patterns we found in pairwise correlation matrix can also be supported by these scatter plots.