Stat415-homework1

Homework1.1

(a) Name different kinds variables related to age

Answer: categorical variable: age over 23 and age below 23

ordinal variable: age below 18, age between 18-20, age between 21-23,age between 24-26 and age over 26

interval variable: student's age

ratio variable: respective percentage of students whose age ranges from 18 to 30

(b) Name a population about which we could plausibly make inferences

Answer: all the boys from Stat-415 course

(c) Name a population about which we could not make valid inferences

Answer: All the students from Stat-415 course whose age is 23

Homework1.2

(a) What is the effect of this transformation?

Answer: If it occurs in one document, original frequency would multiply by log(n) to get updated frequency through this transformation. If it occurs in every document, updated frequency would become zero directly through this transformation.

(b) What might be the purpose of this transformation?

Answer:

We can use this transformation to measure how much information the word provides and decide if it is common or rare in all documents. If the word is quite common in all documents, then it should be irrelevant or non-significant. Thus this transformation would diminish its weight. If the word is not common in all the documents but it appears quite often in certain documents, then it is very likely to be important. As a result, it can be considered as a key word and its weight would be increased by this transformation. With the help of this transformation, we can filter key words in the text.

Homework1.3

3.Perform exploratory data analysis of the data set and report your results. Comment on any interesting or significant features.Try to contain some numerical summaries for each variable and give multivariate numerical summaries (pairwise correlation) & graphical summaries.

(a) Use the read.csv() function to read the data into R.

```
data=read.csv("College.csv",header=T)
```

(b) Look at the data using the fix() function.

```
#give each row a name corresponding to the appropriate university
rownames(data) = data[,1]
fix(data)
#eliminate the first column of names in the data matrix before performi
ng numerical operations
data = data[,-1]
```

(c) Numerical summaries for all variables

```
attach(data)
# treat data$Private as a categorical variable
as.factor(Private)

summary(data)

## 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
```

Analysis:

According to the summary result, 212 samples are public universities, while others are private universities. The number of Applications received ranges from 81 to 48094. The median and mean of number of applications are 1558 and 3002 separately. The number of accepted applicants ranges from 72 to 26330 and the number of enrolled students ranges from 35 to 6392.

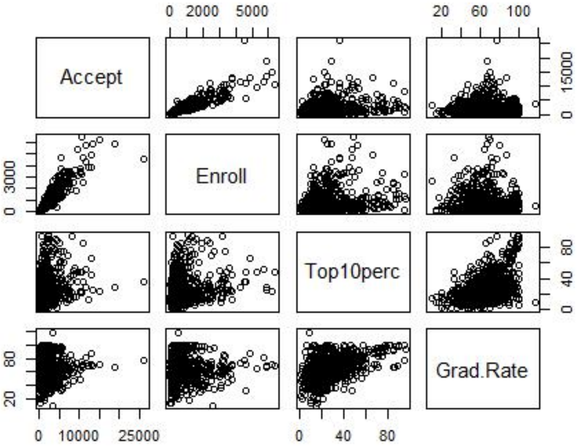
The percent of new students from Top 10% ranges from 1% to 96%, while the percent of new students from Top 25% ranges from 9% to 100%. The number of full time undergraduates ranges from 139 to 31643, while the number of part time undergraduates ranges from 1 to 21936.

The cost could be divided into several parts. Out state tuition could range from 2340 to 21700. Cost for Room.Board ranges from 1780 to 8124. Cost for books ranges from 96 to 2340,while personal spending ranges from 250 to 6800. Variables related to cost indicates that the total cost would be quite different among students and universities.

There is also more information about faculty and educational resource.The percent of faculty with PhD's ranges from 8% to 103%.However, the percent should not overcome 100 thus it could be indicated that some errors may exist here and filtration may be needed. Percent of faculty with terminal degree ranges from 24% to 100%. S.F.Ratio ranges from 2.5 to 39.8. Percent of alumni who donate ranges from 0% to 64%. Instructional expenditure per student ranges from 3186 to 56233 and graduation rate ranges from 10% to 118%. Similarly, this percentage should not over come thus the data may need to be dealt with before further analysis. In general, there is much difference between universities in faculty and educational resource.

(d) A scatter plots matrix can be produced with the pairs() function.Recall you can select a subset of variables to plot.

```
data1=data[,3:5]
Scatter_data=data.frame(data1,Grad.Rate)
pairs(Scatter_data)
```

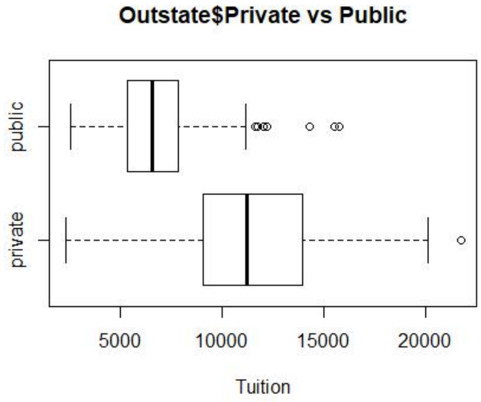


```
cor(Scatter_data)

##                Accept      Enroll Top10perc   Grad.Rate
## Accept    1.00000000  0.91163666 0.1924469  0.06731255
## Enroll    0.91163666  1.00000000 0.1812935 -0.02234104
## Top10perc 0.19244693  0.18129353 1.0000000  0.49498923
## Grad.Rate 0.06731255 -0.02234104 0.4949892  1.00000000
```

According to the result, it is obvious that there is positive relationship between number of applicants accepted and Number of new students enrolled and the correlation is 0.91163666. Besides, there is positive relationship between Graduate Rate and Top10 percent and the correlation is 0.49498923. It indicates that better students would lead to higher graduate rate.

(e) Produce side-by-side box plots if one of the variables is categorical.

```
attach(data)

Private_private=which(Private=="Yes")
Private_public=which(Private=="No")
private=Outstate[Private_private]
public=Outstate[Private_public]
cate_outstate=c("private","public")
boxplot(private,public,names=cate_outstate,horizontal=TRUE,main="Outsta
te$Private vs Public",xlab="Tuition")
```

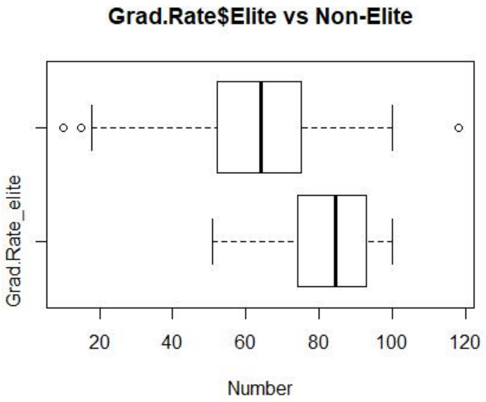


According to this side by side plot, we can find that out state tuition for public universities would be lower compared with the tuition for private universities. To be more specific, the median tuition for public universities would be around 6000-7000, while the median tuition for private universities has been over 10000.

(f) You can create new variables by transforming original variables.

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

##  No Yes
## 699  78
```

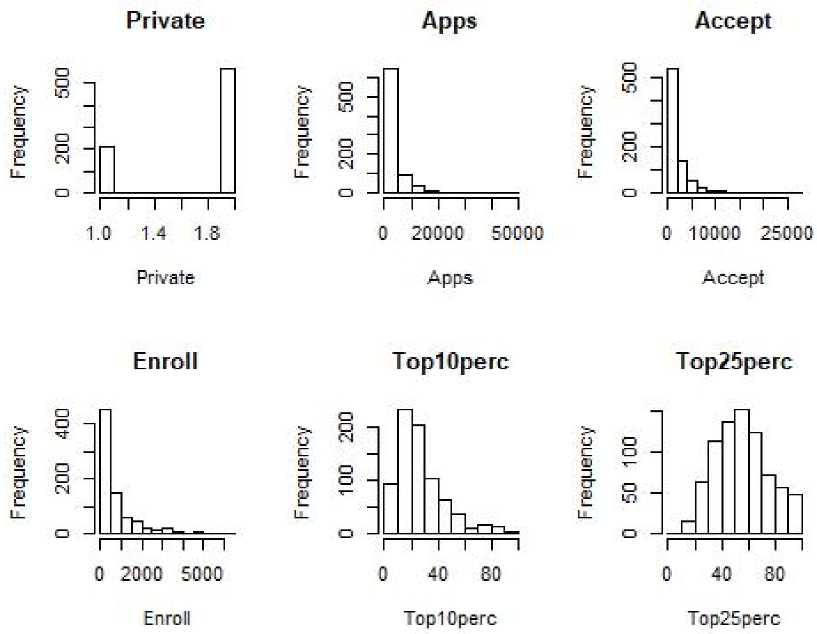
```r
# compare percent of faculty with Grad.Rate between Elite and non-Elite
Grad.Rate_Elite=which(Elite=="Yes")
Grad.Rate_NonElite=which(Elite=="No")
Grad.Rate_elite=Grad.Rate[Grad.Rate_Elite]
Grad.Rate_nonelite=Grad.Rate[Grad.Rate_NonElite]
cate_Grad.Rate=c("Grad.Rate_elite","Grad.Rate_nonelite")
boxplot(Grad.Rate_elite,Grad.Rate_nonelite,names=cate_Grad.Rate,horizontal=TRUE,main="Grad.Rate$Elite vs Non-Elite",xlab="Number")
```



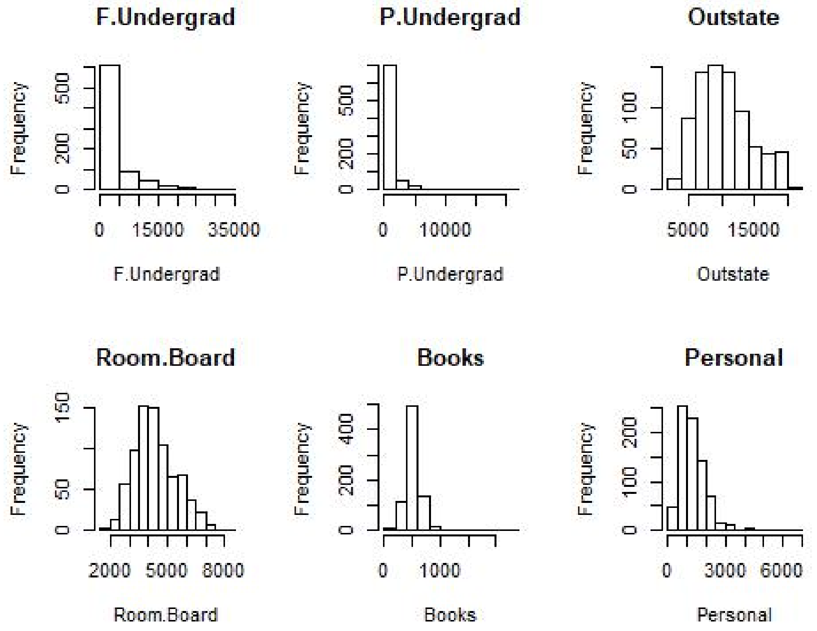
According to the side-by-side box plots, it is obvious that Elite universities have higher graduate rate.

(g) The hist() function produces histograms.

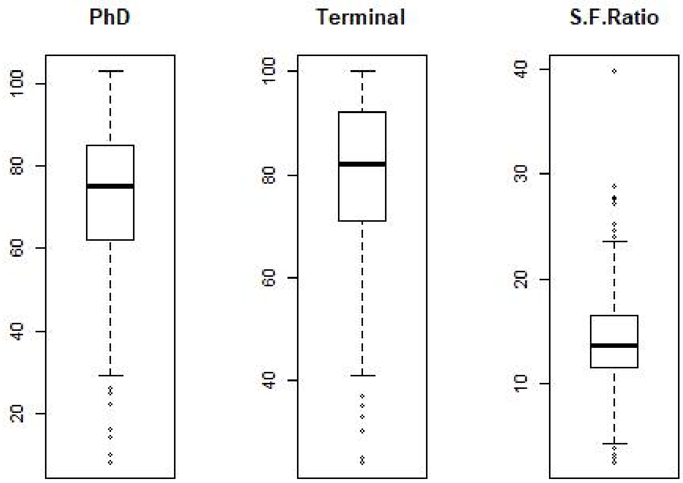
```r
par(mfrow=c(2,3))
attach(data)

Private=as.numeric(Private)
hist(Private,main="Private")
hist(Apps,main="Apps")
hist(Accept,main="Accept")
hist(Enroll,main="Enroll")
hist(Top10perc,main="Top10perc")
hist(Top25perc,main="Top25perc")
```



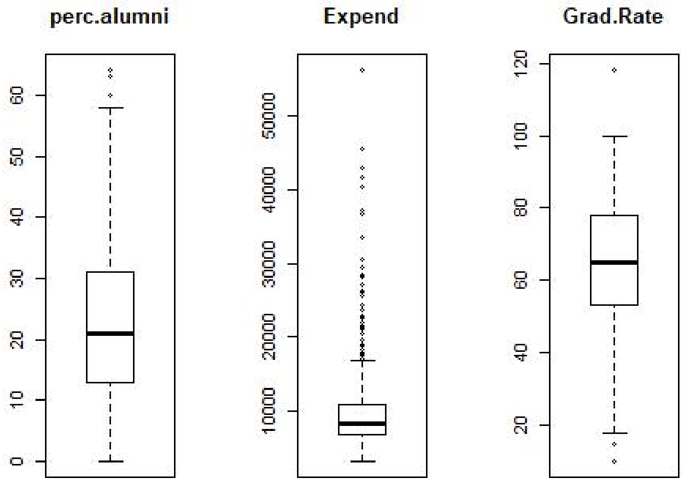
```
par(mfrow=c(2,3))
attach(data)

hist(F.Undergrad,main="F.Undergrad")
hist(P.Undergrad,main="P.Undergrad")
hist(Outstate,main="Outstate")
hist(Room.Board,main="Room.Board")
hist(Books,main="Books")
hist(Personal,main="Personal")
```



```r
attach(data)

par(mfrow=c(1,3))
boxplot(PhD,main="PhD")
boxplot(Terminal,main="Terminal")
boxplot(S.F.Ratio,main="S.F.Ratio")
```



```r
attach(data)

par(mfrow=c(1,3))
boxplot(perc.alumni,main="perc.alumni")
boxplot(Expend,main="Expend")
boxplot(Grad.Rate,main="Grad.Rate")
```


perc.alumni
Expend
Grad.Rate