HW1

## (8)

### (a)

* 我們用read.csv成功讀取資料，header=T 意思為讀取第一列(row)為變數名稱。

## X Private Apps Accept Enroll Top10perc Top25perc  
## 1 Abilene Christian University Yes 1660 1232 721 23 52  
## 2 Adelphi University Yes 2186 1924 512 16 29  
## 3 Adrian College Yes 1428 1097 336 22 50  
## 4 Agnes Scott College Yes 417 349 137 60 89  
## 5 Alaska Pacific University Yes 193 146 55 16 44  
## 6 Albertson College Yes 587 479 158 38 62  
## F.Undergrad P.Undergrad Outstate Room.Board Books Personal PhD Terminal  
## 1 2885 537 7440 3300 450 2200 70 78  
## 2 2683 1227 12280 6450 750 1500 29 30  
## 3 1036 99 11250 3750 400 1165 53 66  
## 4 510 63 12960 5450 450 875 92 97  
## 5 249 869 7560 4120 800 1500 76 72  
## 6 678 41 13500 3335 500 675 67 73  
## S.F.Ratio perc.alumni Expend Grad.Rate  
## 1 18.1 12 7041 60  
## 2 12.2 16 10527 56  
## 3 12.9 30 8735 54  
## 4 7.7 37 19016 59  
## 5 11.9 2 10922 15  
## 6 9.4 11 9727 55

### (b)

* 可利用rownames()來命名每個row，並且在命名完後刪掉第一行(Column)。
* fix()的圖無法在Markdown呈現，放在文後。

rownames(college)=college[,1]  
fix(college)  
  
college =college [,-1]  
fix(college)

### (c)

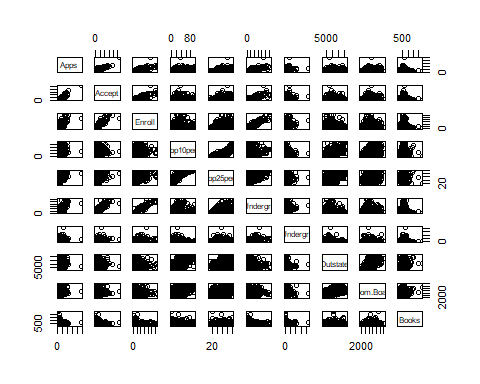
#### <1>

* summary()可看出每個變數的屬性，如此題Private為character， 其餘變數為數值資料，則列出簡單的敘述統計

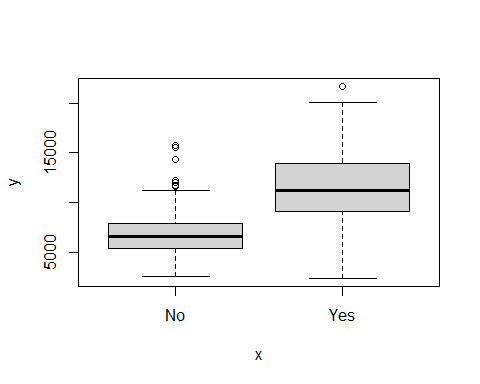
## Private Apps Accept Enroll   
## Length:777 Min. : 81 Min. : 72 Min. : 35   
## Class :character 1st Qu.: 776 1st Qu.: 604 1st Qu.: 242   
## Mode :character Median : 1558 Median : 1110 Median : 434   
## Mean : 3002 Mean : 2019 Mean : 780   
## 3rd Qu.: 3624 3rd Qu.: 2424 3rd Qu.: 902   
## Max. :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

#### <2>

* 第一個變數Private為character，故從第二個變數開始到第11個變數執行， pairs()可畫出數值型資料兩兩變數間的散佈圖。



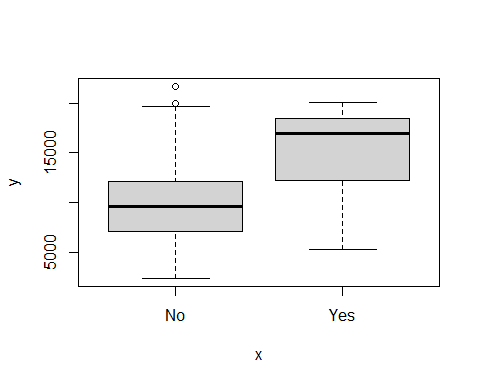
#### <3>

* 將Private的outcome讀做factor後放在x，Outstate的值放在y即可畫出分別的盒狀圖, 觀察下圖可發現Private為No的Outstate資料明顯低於Yes且Variance也明顯較低，資料較集中。 其代表的意思為對於別州或國際學生的學費在私立大學明顯比公立的大學貴。 

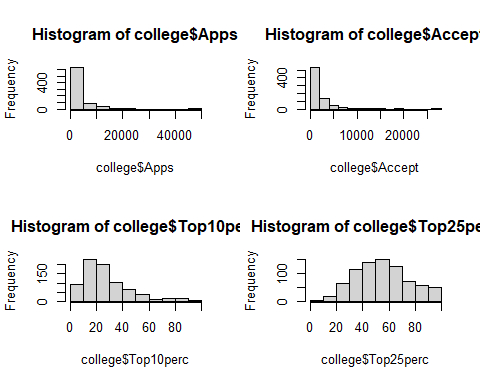
#### <4>

* 將大學以Top10perc是否大於50分為兩群，使用summary()後呈現累積次數表等同table()

## No Yes   
## 699 78

* 由此盒狀圖可看到Yes組的明顯比No組的來得高， 代表在私立大學收到顯著比公立大學還要多，來自前10%高中的學生。 

#### <5>

* 上方兩張圖代表App的Received 跟Accepted 的數量，明顯的呈現右偏，代表此App在大多學校裡並不流行
* 下方兩張圖代表新生來自Top 10% , 25%的高中的數量，可明顯看到Top 10 為右偏，Top 25 較對稱， 符合我們對於成績優秀學生人數較少的直覺。 

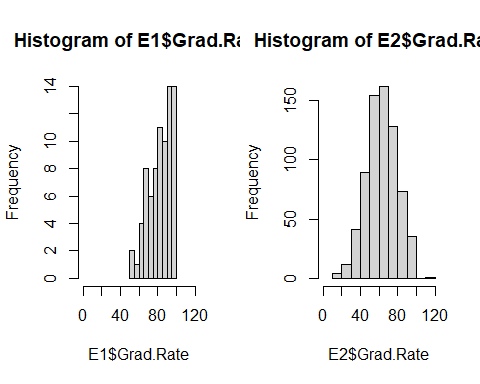
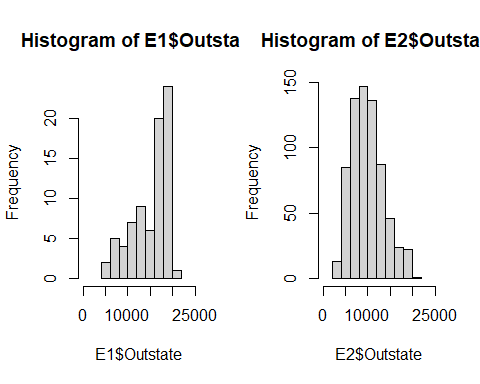
#### (d)

* 我們藉由前面對於Elite的定義(Yes/No)分為E1,E2兩個組群。

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

* 明顯可看為Elite大學的Graduation rate為左偏分配，且集中100左右，而非Elite大學的Graduation rate的分配較為對稱，平均落在60%左右，由畢業率的差異，可見此Elite變數可有效分別出成績較優秀的學生。 
* 明顯可看為Elite大學的Out-of-state tuition 為左偏分配，且集中20000左右， 而非Elite大學的Graduation rate的分配則略為右偏，平均落在10000左右， 顯示出被判斷為Elite的大學對於別州、國際學生的收費較非Elite的大學更貴。 

## (10)

### (a)

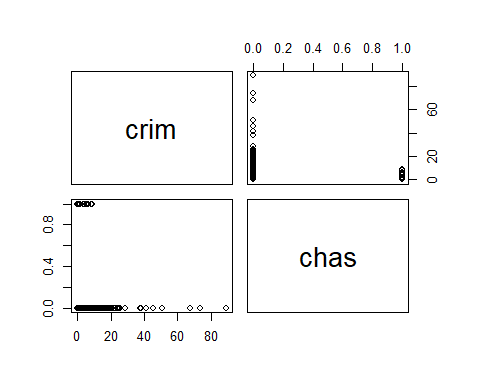
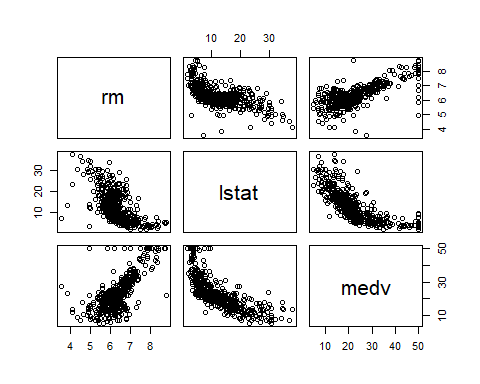
* 總共有506個row，14個columns，row代表觀測值，columns代表變數

##   
## Attaching package: 'MASS'

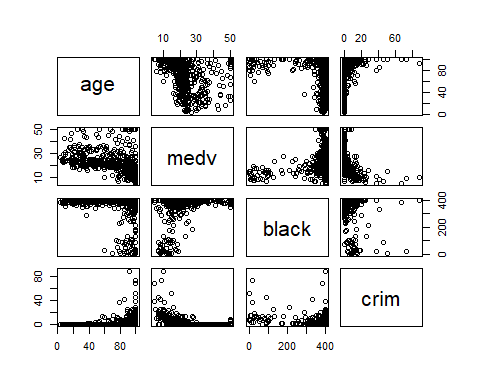
## The following object is masked from 'package:dplyr':  
##   
## select

## 'data.frame': 506 obs. of 14 variables:  
## $ crim : num 0.00632 0.02731 0.02729 0.03237 0.06905 ...  
## $ zn : num 18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...  
## $ indus : num 2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...  
## $ chas : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ nox : num 0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...  
## $ rm : num 6.58 6.42 7.18 7 7.15 ...  
## $ age : num 65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...  
## $ dis : num 4.09 4.97 4.97 6.06 6.06 ...  
## $ rad : int 1 2 2 3 3 3 5 5 5 5 ...  
## $ tax : num 296 242 242 222 222 222 311 311 311 311 ...  
## $ ptratio: num 15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...  
## $ black : num 397 397 393 395 397 ...  
## $ lstat : num 4.98 9.14 4.03 2.94 5.33 ...  
## $ medv : num 24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...

### (b)

* rm 代表平均每間住宅有幾個房間
* Istat 代表人口中屬於較低階級的比例
* medv 代表區域房價的中位數
* crim 代表 平均的犯罪率
* chas 代表 是否在Charles River邊
* age 代表 1940年之前建造的自有住房的比例。
* tax 代表 每10,000美元的財產稅率。
* ptration 代表各區的師生比例
* 由圖一可見隨著平均房間數增加，低端人口比例顯著的遞減，推論可能是因為房間數的增加代表此區住戶多為一整個家庭，也因此能組成一個家庭的人的收入不大可能會是低端人口，也因此房價亦隨著平均房間數的增加而升高呈現正相關。
* 由圖二可明顯看到在河畔的犯罪件數明顯低上不少，代表河邊住的可能為較高端的人口，也因此治安較佳。 

### (c)

* 由圖可見，age代表舊房屋的比例，而隨著age的比例升高，crim的件數也增加，推論舊城區的地方治安較差。
* medv與crim呈現負相關，由之前的推論房價低的地方低端人口的比例較高， 由以上推論可得房價較低的地方低端人口較多影響治安較差。
* 從black與age可看出大部分的黑人族群都住在舊城區，也因此與crim的關係與age相似，在黑人較多的地區，犯罪件數較多。 

### (d)

* 列出crim,tax,ptratio大於90百分位數的觀測值，發現各變數所列的出的觀測值皆不重複，且各組之間的所篩選出的觀測值有連續性 (E.g. crim取出的資料為第370480筆,tax取出的資料為第489493筆)

## [1] 368 374 375 376 377 379 380 381 382 385 386 387 388 389 393 395 399 401 402  
## [20] 404 405 406 407 408 410 411 412 413 414 415 416 417 418 419 420 421 423 426  
## [39] 427 428 435 436 437 438 439 441 445 469 470 478 480

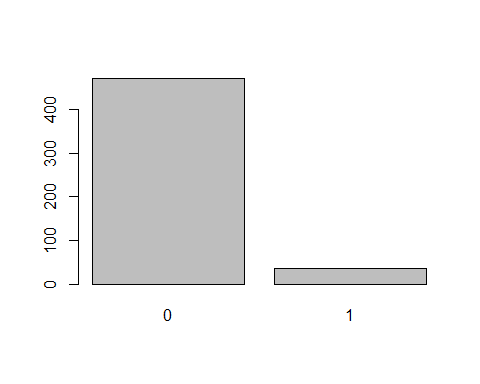
## [1] 489 490 491 492 493

## [1] 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32  
## [20] 33 34 35 55 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142  
## [39] 355 356 502 503 504 505 506

### (e)

* 由table可看出有35區在Charles River旁。

##   
## 0 1   
## 471 35



### (f)

* ptratio的中位數為19.05，代表平均的師生比例為一個老師需要教導約19個學生。

## [1] 19.05

### (g)

* 首先發現第399,406筆資料為最小的medv(中位數房價)，接著以crim,black,age等變數來看他們與平均值的差異， 由以下結果可發現，該兩區的犯罪件數明顯比平均高，黑人比例亦比平均高，且兩者皆為舊城區。

## [1] 399 406

## [1] "medv Min(399): 5 v.s. Average: 22.53"

## [1] "crim Min(399): 38.3518 v.s. Average: 3.61"

## [1] "black Min(399): 396.9 v.s. Average: 356.67"

## [1] "age Min(399): 100 v.s. Average: 68.57"

## [1] "medv Min(406): 5 v.s. Average: 22.53"

## [1] "crim Min(406): 67.9208 v.s. Average: 3.61"

## [1] "black Min(406): 384.97 v.s. Average: 356.67"

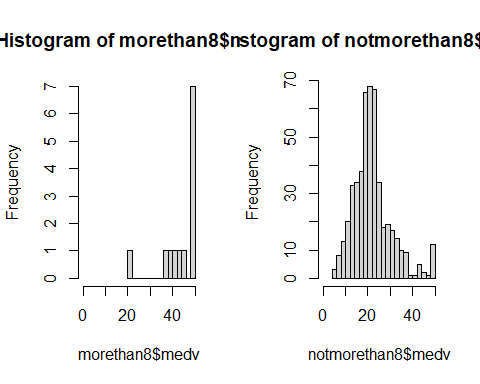
## [1] "age Min(406): 100 v.s. Average: 68.57"

### (h)

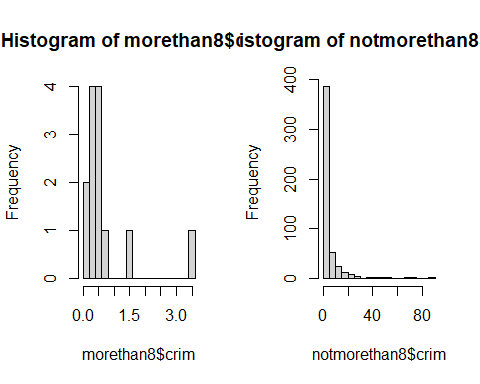
* 列出平均房間大於7間、8間的區數

## [1] 64

## [1] 13

* 以平均房間數8間做為分組標準，畫出兩組的中位數房價直方圖，由圖可見， 大於7間的房價為左偏分配大多落在30以上，而小於8間的房價為右偏分配，且平均約略落在20-30之間， 由以上敘述可推論房間數較多的住宅區有較高的房價。 

## [1] "medv morethan8: 44.2 v.s. notmorethan8: 21.96"

* 平均房數大於8間的區域的犯罪率明顯比小於8間的還要小。 

## [1] "crim morethan8: 0.72 v.s. notmorethan8: 3.69"