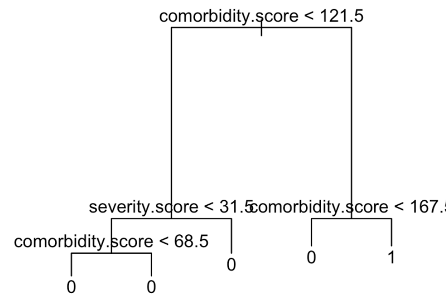
1.

(a)

```
> tree_health_data = tree(readmit30~.-readmit30, healthdata)
> plot(tree_health_data)
> text(tree_health_data,pretty=0)
> tree_health_data
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

 1) root 4382 4702.0 0 ( 0.77225 0.22775 )
   2) comorbidity.score < 121.5 3084 2440.0 0 ( 0.86511 0.13489 )
     4) severity.score < 31.5 2423 1609.0 0 ( 0.89682 0.10318 )
       8) comorbidity.score < 68.5 1337  616.7 0 ( 0.93867 0.06133 ) *
       9) comorbidity.score > 68.5 1086  935.6 0 ( 0.84530 0.15470 ) *
     5) severity.score > 31.5 661  745.1 0 ( 0.74887 0.25113 ) *
   3) comorbidity.score > 121.5 1298 1786.0 0 ( 0.55162 0.44838 )
     6) comorbidity.score < 167.5 770 1001.0 0 ( 0.64545 0.35455 ) *
     7) comorbidity.score > 167.5 528  716.5 1 ( 0.41477 0.58523 ) *
>
```



(b)

```
> cv.health
$size
[1] 5 4 3 2 1

$dev
[1] 4056.280 4131.010 4167.384 4234.041 4704.065

$k
[1]      -Inf  56.62428  67.67316  85.95055 476.79161

$method
[1] "deviance"

attr(,"class")
[1] "prune"         "tree.sequence"
```

(c)

```
> healthdata$caretrack = 1
> healthdata[comorbidity.score<68.5 & severity.score<31.5,]$caretrack = 0
> as.double(sum(healthdata$caretrack==1) / nrow(healthdata))
[1] 0.6948882
```

(d) The unawareness difference between male and female is about 0.0395, while the accuracy parity difference is about 0.0745

```
> # unawareness
> nrow(female_care)/nrow(female) - nrow(care)/nrow(healthdata)
[1] 0.03954762
>
> # Accuracy Parity
> nrow(female_care)/nrow(female) - nrow(male_care)/nrow(male)
[1] 0.07450459
```

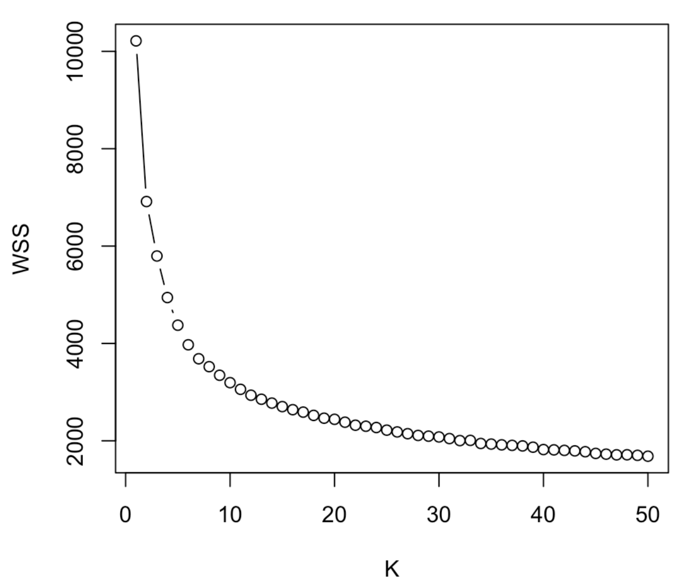
2.

(a)

```
> linReg = lm(SAT_AVG~.-INSTNM, data=clgdata)
> summary(linReg)

Call:
lm(formula = SAT_AVG ~ . - INSTNM, data = clgdata)

Residuals:
    Min      1Q  Median      3Q     Max 
-254.21  -44.63    3.83   45.58  326.06 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     6.515e+02  1.165e+01  55.941  < 2e-16 ***
UGDS            2.019e-04  3.755e-04   0.538  0.59089    
COSTT4_A        3.288e-05  5.068e-04   0.065  0.94828    
TUITIONFEE_OUT  1.798e-03  6.596e-04   2.725  0.00653 ** 
TUITFTE        -7.138e-04  6.609e-04  -1.080  0.28034    
AVGFACSAL       1.626e-02  1.453e-03  11.197  < 2e-16 ***
PFTFAC          4.090e+01  9.069e+00   4.510 7.16e-06 ***
C150_4          4.226e+02  1.962e+01  21.533  < 2e-16 ***
PFTFTUG1_EF    -1.672e+01  1.375e+01  -1.216  0.22419    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 70.62 on 1127 degrees of freedom
Multiple R-squared:  0.6839,	Adjusted R-squared:  0.6817 
F-statistic: 304.8 on 8 and 1127 DF,  p-value: < 2.2e-16
```

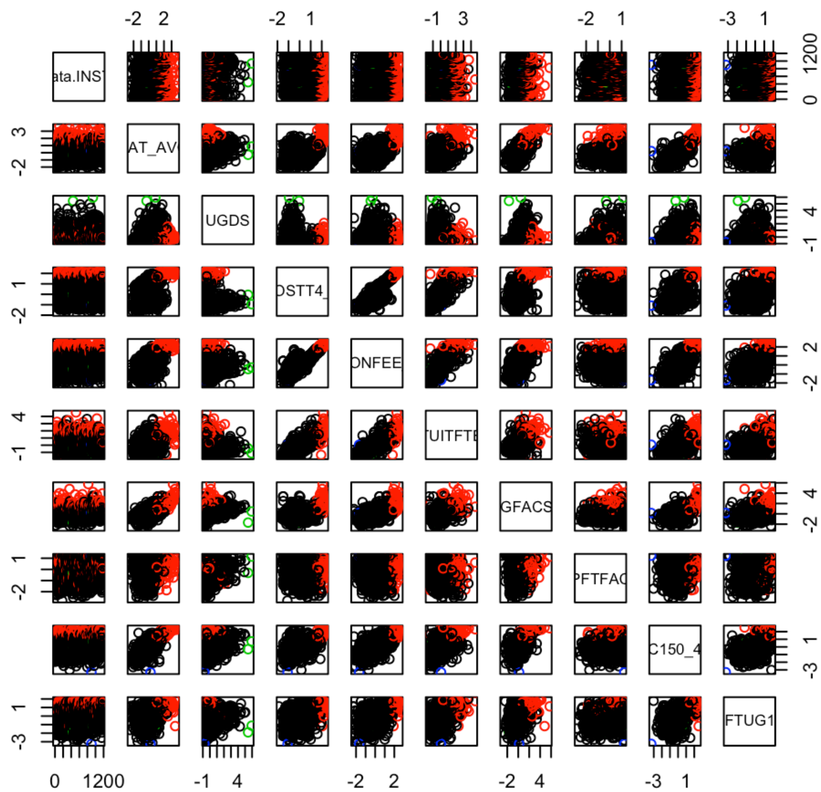
(b) I will choose k=10, because WSS[10] is quite small and k=10 is not large.

```
> wss
 [1] 10215.000  6914.788  5795.651  4942.431  4373.990  3974.122  3684.776  3506.198  3353.539  3200.516  3068.700
[12]  2936.936  2854.959  2774.736  2709.847  2644.923  2590.017  2537.823  2470.561  2429.626  2399.490  2344.324
[23]  2282.184  2256.681  2205.501  2183.340  2133.925  2118.222  2086.069  2071.296  2026.466  2019.438  1977.567
[34]  1963.573  1934.162  1926.669  1892.010  1882.702  1853.873  1832.858  1826.580  1796.540  1790.871  1771.559
[45]  1762.516  1731.407  1718.260  1711.130  1685.122  1672.938
> which.min(wss)
[1] 50
```



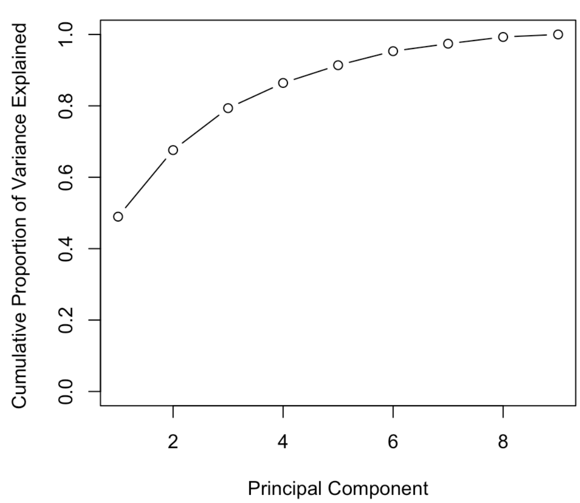
(c) Four centroids are as below

```
                      [,1]        [,2]       [,3]        [,4]
SAT_AVG       -0.163300802  2.22905202  0.2895200  0.22960182
UGDS          -0.005827803 -0.05769499  5.6884622 -0.78024995
COSTT4_A      -0.144857457  2.01363472 -0.4975825 -1.08523374
TUITIONFEE_OUT -0.151068563  2.10317592 -0.3725193 -1.67110411
TUITFTE       -0.145128740  2.00994194 -0.7855692  0.06155869
AVGFACSAL     -0.144104164  1.99035081 -0.6197098  0.15640380
PFTFAC        -0.015128468  0.18069646  0.4173429  1.22734896
C150_4        -0.136317674  1.90702499  0.2130569 -3.31557448
PFTFTUG1_EF   -0.086476336  1.26756405 -1.5151287 -3.25316393
```



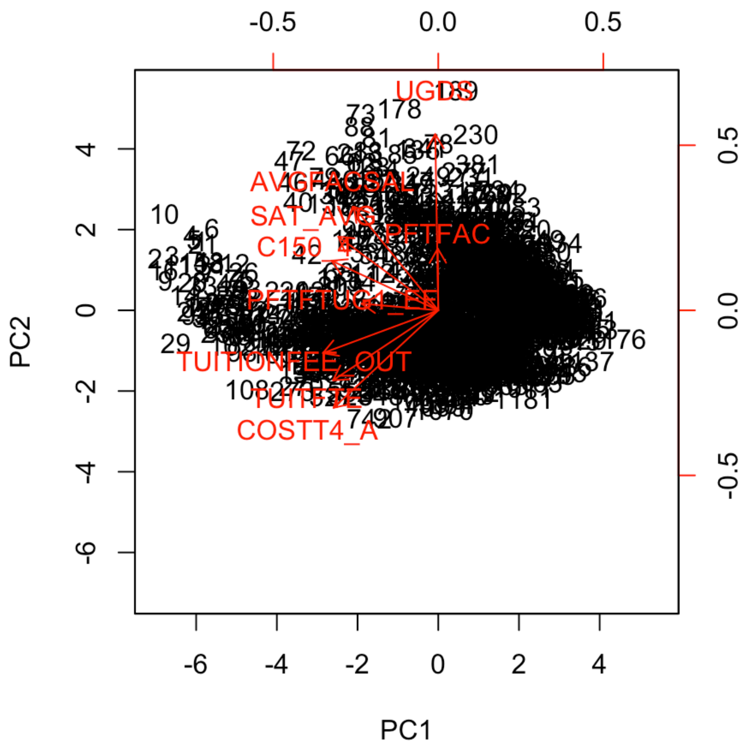
(d) PVE values and cumulative PVE plot is as below.

```
> pve
[1] 0.489610289 0.186469957 0.117469839 0.070554030 0.049739869 0.039294997 0.020962470 0.018814019 0.007084531
```



(e)

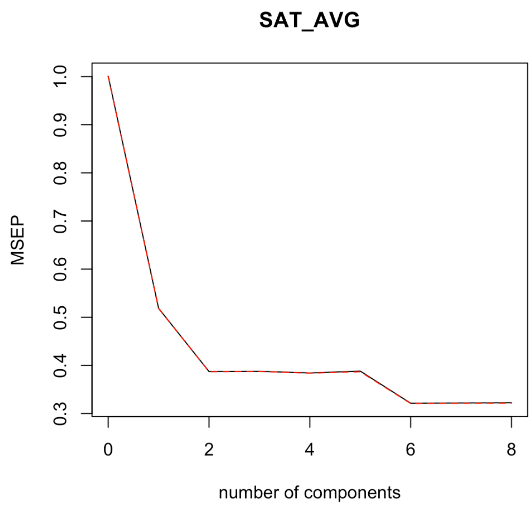
PC1 indicates a linear combination of COSTT4_A, TUITFTE, TUITIONFEE_OUT, PFTFTUG1_EF, C150_4, SAT_AVG and AVGFACSAL, and obviously that PFTFTUG1_EF plays an important part on it. While PC2 indicates another linear combination of UGDS, PFTFTUG1_EF, C150_4, SAT_AVG, AVGFACSAL, COSTT4_A, TUITFTE and TUITIONFEE_OUT, and obviously that UGDS, PFTFTUG1_EF play an important part on it. In addition, PC1 and PC2 are orthogonal.



(f)

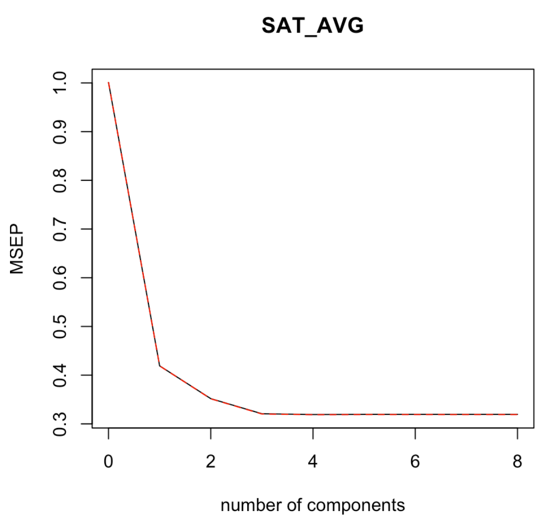
Below is the summary for PCR, I will choose M=5 because it explains 92% of the variance.

```
> summary(pcr.fit)
Data:   X dimension: 1136 8
        Y dimension: 1136 1
Fit method: svdpc
Number of components considered: 8

VALIDATION: RMSEP
Cross-validated using 5 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
CV               1   0.7201   0.6223   0.6227   0.6199   0.6230   0.5669   0.5674   0.5677
adjCV            1   0.7198   0.6221   0.6224   0.6196   0.6219   0.5663   0.5668   0.5671

TRAINING: % variance explained
         1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
X          48.09    67.41    80.62    88.49    92.92    97.07    99.20   100.00
SAT_AVG    48.38    61.45    61.45    61.79    62.11    68.35    68.38    68.39
```



Below is the summary for PLS, I will choose M=6 because it explains 94% of the variance.

```
> summary(pls.fit)
Data:    X dimension: 1136 8
         Y dimension: 1136 1
Fit method: kernelpls
Number of components considered: 8

VALIDATION: RMSEP
Cross-validated using 5 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
CV               1   0.6474   0.5931   0.5664   0.5649   0.5652   0.5652   0.5652   0.5652
adjCV            1   0.6473   0.5929   0.5658   0.5646   0.5649   0.5648   0.5648   0.5648

TRAINING: % variance explained
         1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
X          46.84    66.53    71.87    79.44    87.68    94.64    97.73   100.00
SAT_AVG    58.23    65.10    68.32    68.38    68.39    68.39    68.39    68.39
```



PCR works better because it only requires M=5 while PCR asks for M=6