Business Analytics

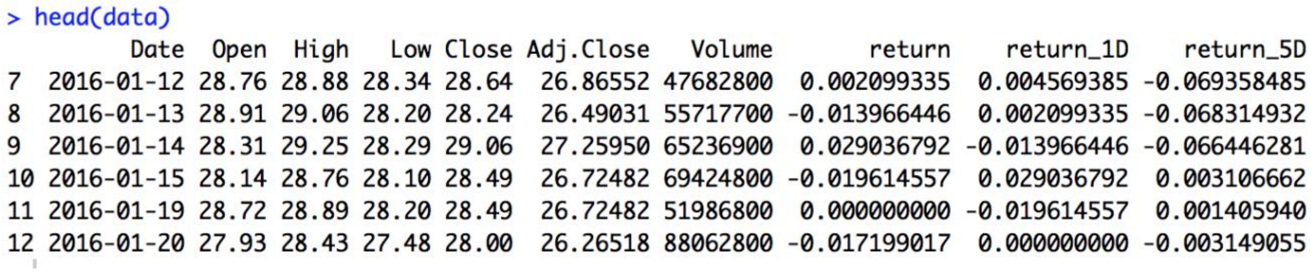
Assignment-2

Name-Subham Kedia

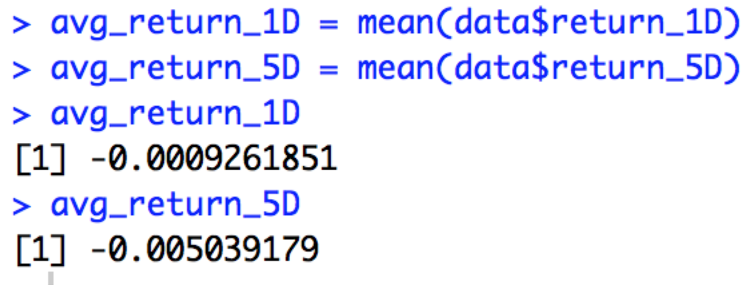
UNI-sk4355

Answer 1

Part a.

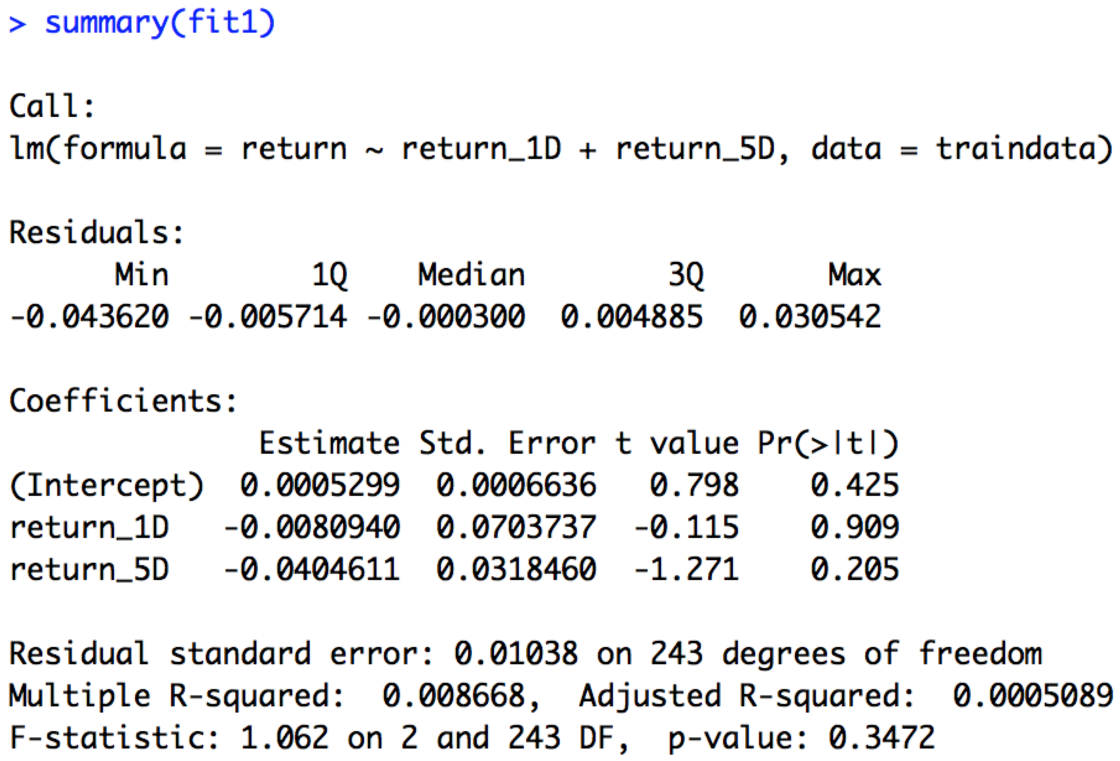


```
> head(data)
         Date  Open  High   Low Close Adj.Close   Volume       return    return_1D    return_5D
7  2016-01-12 28.76 28.88 28.34 28.64  26.86552 47682800  0.002099335  0.004569385 -0.069358485
8  2016-01-13 28.91 29.06 28.20 28.24  26.49031 55717700 -0.013966446  0.002099335 -0.068314932
9  2016-01-14 28.31 29.25 28.29 29.06  27.25950 65236900  0.029036792 -0.013966446 -0.066446281
10 2016-01-15 28.14 28.76 28.10 28.49  26.72482 69424800 -0.019614557  0.029036792  0.003106662
11 2016-01-19 28.72 28.89 28.20 28.49  26.72482 51986800  0.000000000 -0.019614557  0.001405940
12 2016-01-20 27.93 28.43 27.48 28.00  26.26518 88062800 -0.017199017  0.000000000 -0.003149055
```



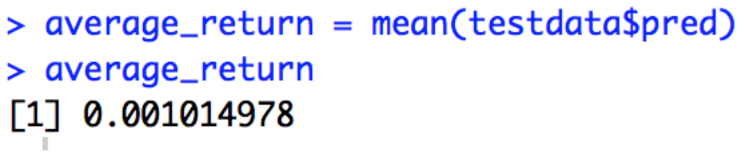
```
> avg_return_1D = mean(data$return_1D)
> avg_return_5D = mean(data$return_5D)
> avg_return_1D
[1] -0.0009261851
> avg_return_5D
[1] -0.005039179
```

Part b.

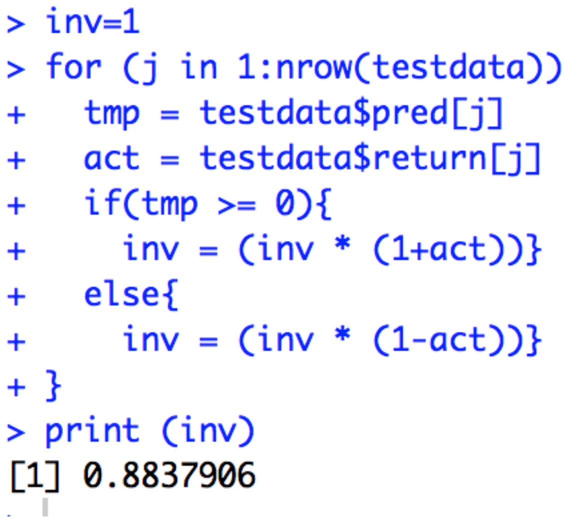


```
> summary(fit1)

Call:
lm(formula = return ~ return_1D + return_5D, data = traindata)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.043620 -0.005714 -0.000300  0.004885  0.030542 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0005299  0.0006636   0.798    0.425
return_1D   -0.0080940  0.0703737  -0.115    0.909
return_5D   -0.0404611  0.0318460  -1.271    0.205

Residual standard error: 0.01038 on 243 degrees of freedom
Multiple R-squared:  0.008668,  Adjusted R-squared:  0.0005089 
F-statistic: 1.062 on 2 and 243 DF,  p-value: 0.3472
```

Part c.



```
> average_return = mean(testdata$pred)
> average_return
[1] 0.001014978
```

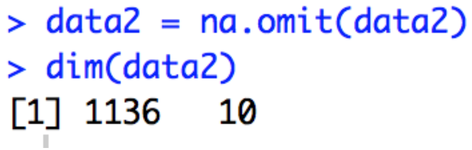


```
> inv=1
> for (j in 1:nrow(testdata)){
+   tmp = testdata$pred[j]
+   act = testdata$return[j]
+   if(tmp >= 0){
+     inv = (inv * (1+act))}
+   else{
+     inv = (inv * (1-act))}
+ }
> print (inv)
[1] 0.8837906
```

The long-short strategy that we implemented did not perform too well and we suffered a loss instead of a gain. The model could be improved by taking other features (factors) into consideration apart from just one-day returns and five-day returns. Firstly, we can also include the interaction term between one-day returns and five-day returns in our model to improve our model and get better predictions. Secondly, there are other factors which influence the returns that we get from a stock like weather, how the industry is doing, how the company is doing compared to the industry, what is the market share of the company, etc. So, it is very important to consider other factors while predicting returns in order to get a better estimate on how to invest, when to long and when to short.

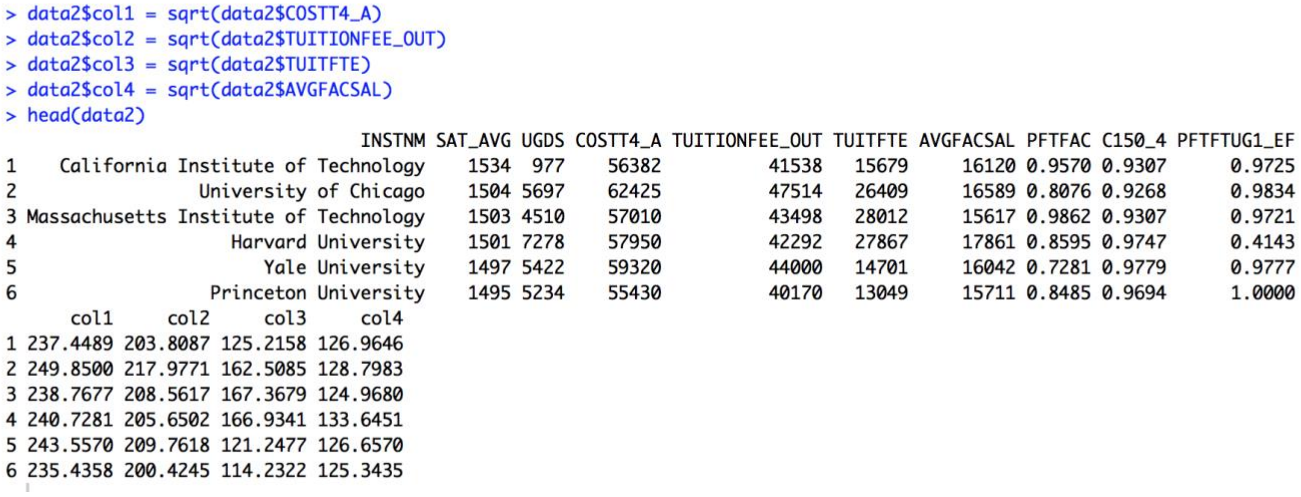
Answer 2.

Part a.

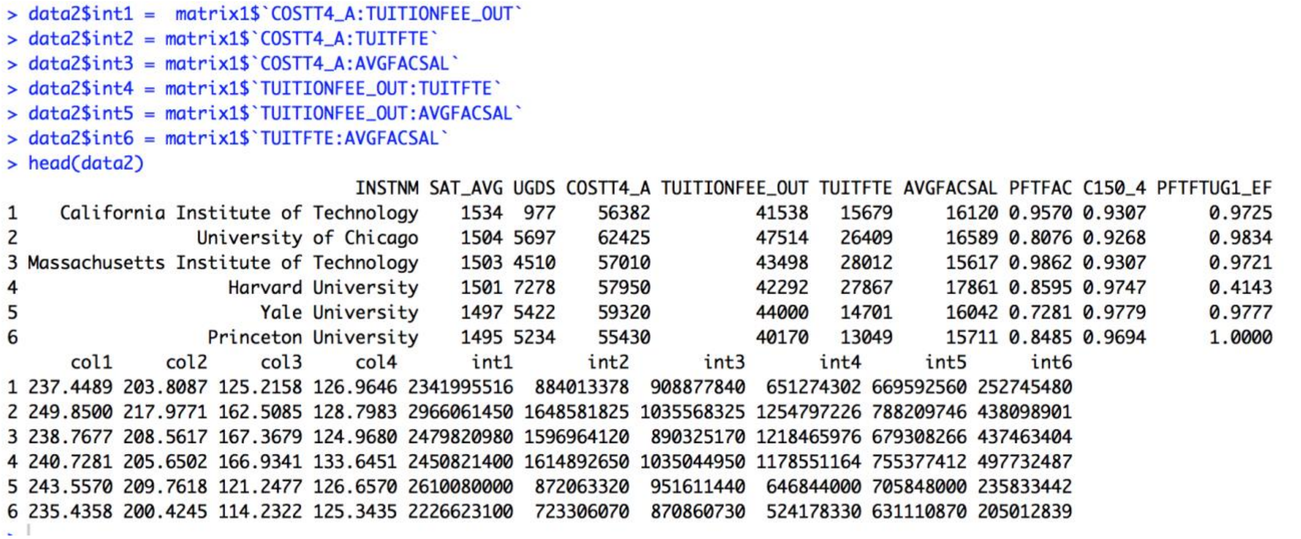


```
> data2 = na.omit(data2)
> dim(data2)
[1] 1136   10
```

Part b.

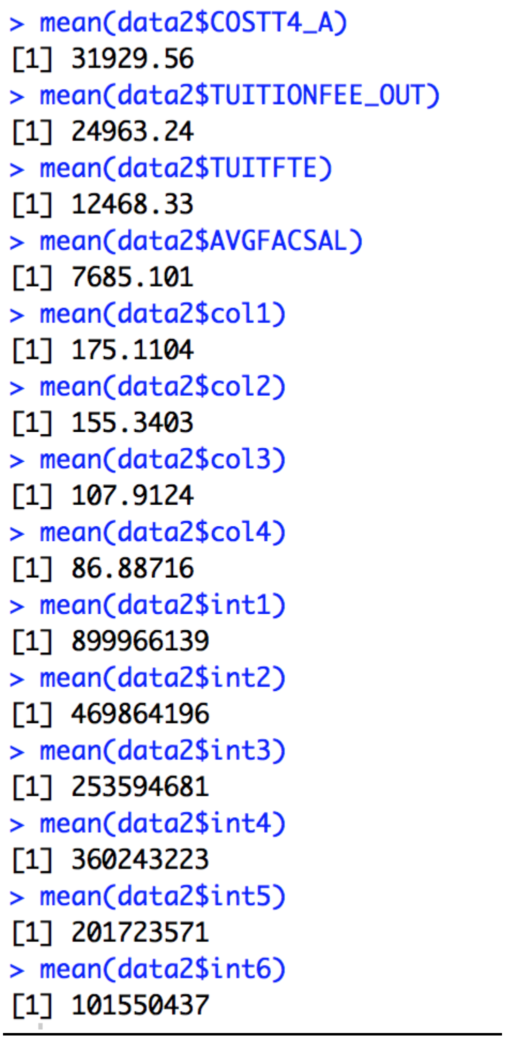


```
> data2$col1 = sqrt(data2$COSTT4_A)
> data2$col2 = sqrt(data2$TUITIONFEE_OUT)
> data2$col3 = sqrt(data2$TUITFTE)
> data2$col4 = sqrt(data2$AVGFACSAL)
> head(data2)
                                INSTNM SAT_AVG UGDS COSTT4_A TUITIONFEE_OUT TUITFTE AVGFACSAL PFTFAC C150_4 PFTFTUG1_EF
1    California Institute of Technology    1534  977    56382          41538   15679     16120 0.9570 0.9307      0.9725
2                 University of Chicago    1504 5697    62425          47514   26409     16589 0.8076 0.9268      0.9834
3 Massachusetts Institute of Technology    1503 4510    57010          43498   28012     15617 0.9862 0.9307      0.9721
4                    Harvard University    1501 7278    57950          42292   27867     17861 0.8595 0.9747      0.4143
5                       Yale University    1497 5422    59320          44000   14701     16042 0.7281 0.9779      0.9777
6                  Princeton University    1495 5234    55430          40170   13049     15711 0.8485 0.9694      1.0000
      col1     col2     col3     col4
1 237.4489 203.8087 125.2158 126.9646
2 249.8500 217.9771 162.5085 128.7983
3 238.7677 208.5617 167.3679 124.9680
4 240.7281 205.6502 166.9341 133.6451
5 243.5570 209.7618 121.2477 126.6570
6 235.4358 200.4245 114.2322 125.3435
```



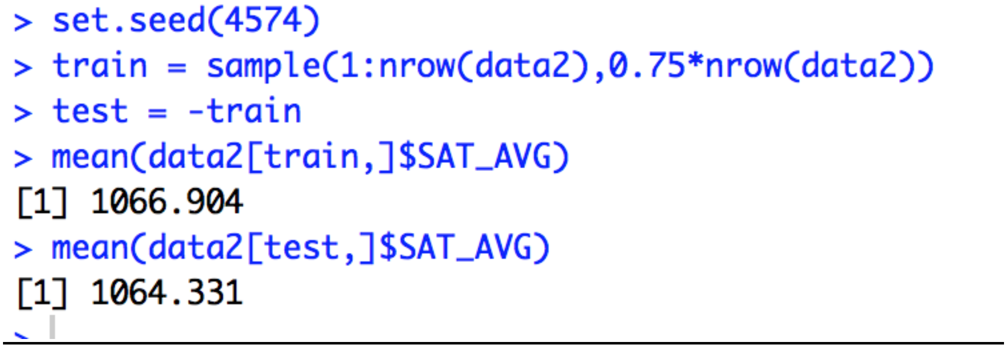
```
> data2$int1 =  matrix1$`COSTT4_A:TUITIONFEE_OUT`
> data2$int2 = matrix1$`COSTT4_A:TUITFTE`
> data2$int3 = matrix1$`COSTT4_A:AVGFACSAL`
> data2$int4 = matrix1$`TUITIONFEE_OUT:TUITFTE`
> data2$int5 = matrix1$`TUITIONFEE_OUT:AVGFACSAL`
> data2$int6 = matrix1$`TUITFTE:AVGFACSAL`
> head(data2)
                                INSTNM SAT_AVG UGDS COSTT4_A TUITIONFEE_OUT TUITFTE AVGFACSAL PFTFAC C150_4 PFTFTUG1_EF
1    California Institute of Technology    1534  977    56382          41538   15679     16120 0.9570 0.9307      0.9725
2                 University of Chicago    1504 5697    62425          47514   26409     16589 0.8076 0.9268      0.9834
3 Massachusetts Institute of Technology    1503 4510    57010          43498   28012     15617 0.9862 0.9307      0.9721
4                    Harvard University    1501 7278    57950          42292   27867     17861 0.8595 0.9747      0.4143
5                       Yale University    1497 5422    59320          44000   14701     16042 0.7281 0.9779      0.9777
6                  Princeton University    1495 5234    55430          40170   13049     15711 0.8485 0.9694      1.0000
      col1     col2     col3     col4       int1       int2       int3       int4      int5      int6
1 237.4489 203.8087 125.2158 126.9646 2341995516  884013378  908877840  651274302 669592560 252745480
2 249.8500 217.9771 162.5085 128.7983 2966061450 1648581825 1035568325 1254797226 788209746 438098901
3 238.7677 208.5617 167.3679 124.9680 2479820980 1596964120  890325170 1218465976 679308266 437463404
4 240.7281 205.6502 166.9341 133.6451 2450821400 1614892650 1035044950 1178551164 755377412 497732487
5 243.5570 209.7618 121.2477 126.6570 2610080000  872063320  951611440  646844000 705848000 235833442
6 235.4358 200.4245 114.2322 125.3435 2226623100  723306070  870860730  524178330 631110870 205012839
```

There are 19 covariates in the data set now including the institution name column. If we do not take into account the institution name as our covariate then we have 18 covariates in our data frame now. There are 20 columns in our data set one of which is SAT_AVG, which we will try to predict using different models later.



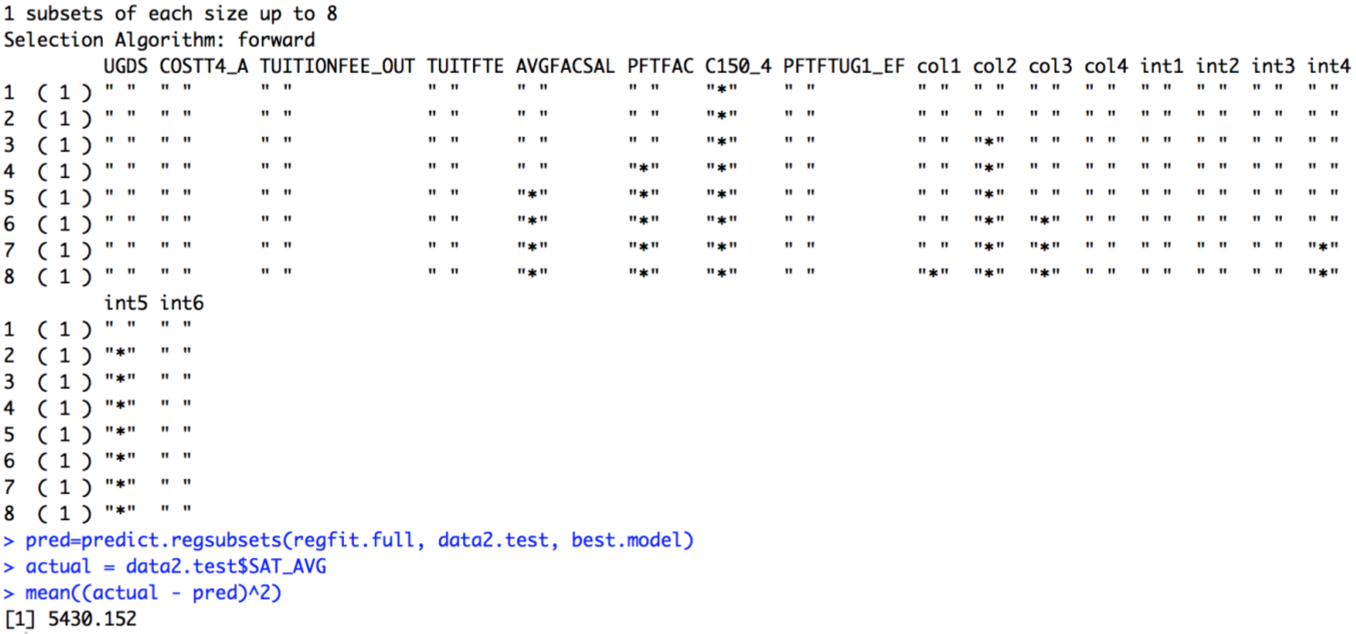
```
> mean(data2$COSTT4_A)
[1] 31929.56
> mean(data2$TUITIONFEE_OUT)
[1] 24963.24
> mean(data2$TUITFTE)
[1] 12468.33
> mean(data2$AVGFACSAL)
[1] 7685.101
> mean(data2$col1)
[1] 175.1104
> mean(data2$col2)
[1] 155.3403
> mean(data2$col3)
[1] 107.9124
> mean(data2$col4)
[1] 86.88716
> mean(data2$int1)
[1] 899966139
> mean(data2$int2)
[1] 469864196
> mean(data2$int3)
[1] 253594681
> mean(data2$int4)
[1] 360243223
> mean(data2$int5)
[1] 201723571
> mean(data2$int6)
[1] 101550437
```

Part c.

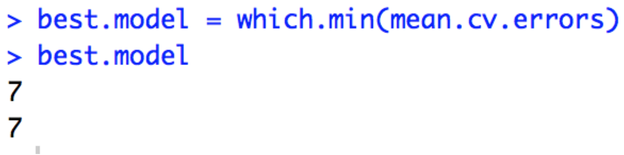


```
> set.seed(4574)
> train = sample(1:nrow(data2),0.75*nrow(data2))
> test = -train
> mean(data2[train,]$SAT_AVG)
[1] 1066.904
> mean(data2[test,]$SAT_AVG)
[1] 1064.331
```

Part d.

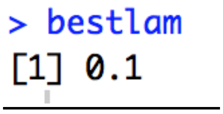


```
1 subsets of each size up to 8
Selection Algorithm: forward
         UGDS COSTT4_A TUITIONFEE_OUT TUITFTE AVGFACSAL PFTFAC C150_4 PFTFTUG1_EF col1 col2 col3 col4 int1 int2 int3 int4
1  ( 1 ) " "  " "      " "            " "     " "       " "    "*"    " "         " "  " "  " "  " "  " "  " "  " "  " " 
2  ( 1 ) " "  " "      " "            " "     " "       " "    "*"    " "         " "  " "  " "  " "  " "  " "  " "  " " 
3  ( 1 ) " "  " "      " "            " "     " "       " "    "*"    " "         " "  "*"  " "  " "  " "  " "  " "  " " 
4  ( 1 ) " "  " "      " "            " "     " "       "*"    "*"    " "         " "  "*"  " "  " "  " "  " "  " "  " " 
5  ( 1 ) " "  " "      " "            " "     "*"       "*"    "*"    " "         " "  "*"  " "  " "  " "  " "  " "  " " 
6  ( 1 ) " "  " "      " "            " "     "*"       "*"    "*"    " "         " "  "*"  "*"  " "  " "  " "  " "  " " 
7  ( 1 ) " "  " "      " "            " "     "*"       "*"    "*"    " "         " "  "*"  "*"  " "  " "  " "  " "  "*" 
8  ( 1 ) " "  " "      " "            " "     "*"       "*"    "*"    " "         "*"  "*"  "*"  " "  " "  " "  " "  "*" 
         int5 int6
1  ( 1 ) " "  " " 
2  ( 1 ) "*"  " " 
3  ( 1 ) "*"  " " 
4  ( 1 ) "*"  " " 
5  ( 1 ) "*"  " " 
6  ( 1 ) "*"  " " 
7  ( 1 ) "*"  " " 
8  ( 1 ) "*"  " " 
> pred=predict.regsubsets(regfit.full, data2.test, best.model)
> actual = data2.test$SAT_AVG
> mean((actual - pred)^2)
[1] 5430.152
```

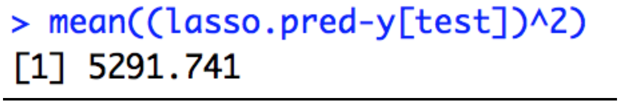


```
> best.model = which.min(mean.cv.errors)
> best.model
7
7
```

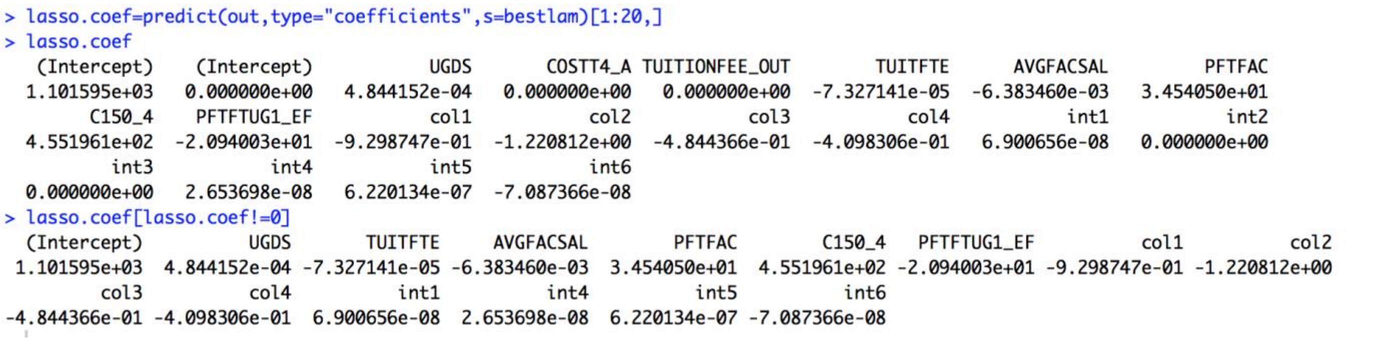
Part e.



```
> bestlam
[1] 0.1
```

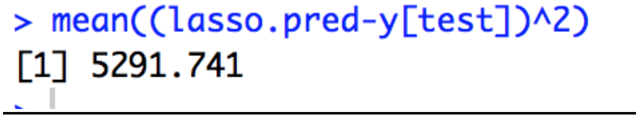


```
> mean((lasso.pred-y[test])^2)
[1] 5291.741
```



```
> lasso.coef=predict(out,type="coefficients",s=bestlam)[1:20,]
> lasso.coef
  (Intercept)   (Intercept)          UGDS      COSTT4_A TUITIONFEE_OUT       TUITFTE     AVGFACSAL        PFTFAC 
 1.101595e+03  0.000000e+00  4.844152e-04  0.000000e+00   0.000000e+00 -7.327141e-05 -6.383460e-03  3.454050e+01 
       C150_4   PFTFTUG1_EF          col1          col2           col3          col4          int1          int2 
 4.551961e+02 -2.094003e+01 -9.298747e-01 -1.220812e+00  -4.844366e-01 -4.098306e-01  6.900656e-08  0.000000e+00 
         int3          int4          int5          int6 
 0.000000e+00  2.653698e-08  6.220134e-07 -7.087366e-08 
> lasso.coef[lasso.coef!=0]
  (Intercept)          UGDS       TUITFTE     AVGFACSAL        PFTFAC        C150_4   PFTFTUG1_EF          col1          col2 
 1.101595e+03  4.844152e-04 -7.327141e-05 -6.383460e-03  3.454050e+01  4.551961e+02 -2.094003e+01 -9.298747e-01 -1.220812e+00 
         col3          col4          int1          int4          int5          int6 
-4.844366e-01 -4.098306e-01  6.900656e-08  2.653698e-08  6.220134e-07 -7.087366e-08 
```

Part f.



```
> mean((lasso.pred-y[test])^2)
[1] 5291.741
```

The above is the MSE of our LASSO model on our data set.

Part g.

The insights that we gained from our model and from our predictions is that what is the average SAT score required to get admitted to a particular university and how does it vary based on factors like tuition fees, faculty salary, etc.

The quality of the institutions is influenced mainly by features which include C150_4 (Completion rate for first-time, full-time students at four-year institutions) and PFTFAC (Proportion of faculty that is Full-Time). The higher the number of students completing full-time degree from a 4-year institution and higher the proportion of faculty that is full-time, the better is the quality of the institutions. The feature PFTFTUG1_EF (share of undergraduate students who are first-time, full-time degree-/certificate-seeking undergraduate students) has a negative effect on the quality of institutions. The higher the number of Full-time faculty (PFTFAC) at an institution, the better is its quality. Basically, higher the faculty to student ratio, better is the quality of education at the institution.

Hence, as a future parent, student, taxpayer, and/or secretary of education I would look at the total number of full-time faculty, the number of full-time degree students who completed their degree from that institution and also the average SAT score required to get admitted to the institution.