NAME: Juan Villegas

Project 2B: Tuition

Please explain all steps and results clearly and cogently, so that a reasonably intelligent manager could understand it. Include your Rcode as part of this assignment. The data, tuition.csv is contained in the Data folder on Blackboard.

1. The dataset is tuition, which contains 1283 records. Partition the data into 80%-20% training dataset (name it tutiontrain) and testing dataset (name it tuitiontest). (Use seed 117 to create this partition.)

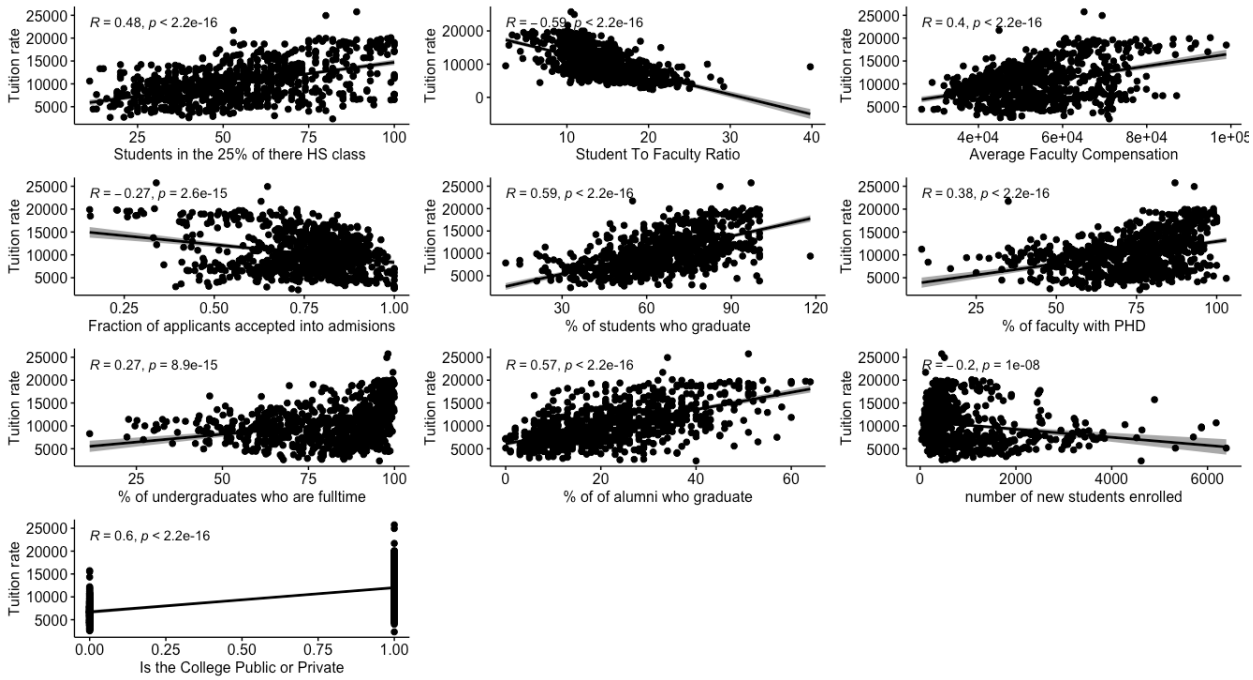
All_Data	804 obs. of 12 variables
InTrain	int [1:644, 1] 1 4 5 7 8 9 10 11 12 13 ...
natuiton	804 obs. of 11 variables
tuition	1284 obs. of 11 variables
tuitiontest	160 obs. of 12 variables
tuitiontrain	644 obs. of 12 variables

In the figure above, you can see that I have the tuition dataframe, which is the default dataset. I then created natuition to get rid of all the N/A values and only have complete values. This reduces the number of observations from 1284 to 804. Next, I partitioned the data into a testing and training dataframes. 80% of the data is in tuitiontrain (644 obs.) and 20% of my data is in tuitiontest (160 obs.)

The variables, for 1283 (total) US Colleges, are as follows:

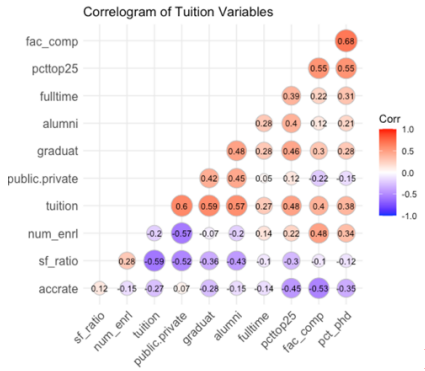
- tuition: College tuition ("out-of-state" rate for those with in-state discount).
- pcttop25: Percent of new students from the top 25% of high school class.
- sf_ratio: Student to faculty ratio.
- fac_comp: Average faculty compensation.
- accrate: Fraction of applicants accepted for admission.
- graduat: Percent of students who graduate.
- pct_phd: Percent of faculty with Ph.D.'s.
- fulltime: Percent of undergraduates who are full time students.
- alumni: Percent of alumni who donate.
- num_enrl: Number of new students enrolled.
- public: Is the college a public or private institution? public=0, private=1

2. Provide a table describing the relationship of each explanatory variable with tuition (scatter plots optional). If the relationship is not linear, make it so by transforming the X variable. [Extra credit to students who can show me how to make scatterplots matrices in R (as shown in textbook but not covered in class).] Otherwise just make scatterplots but that will take a while. Most of the scatterplots here look linear. One possible exception is tuition rate and % of faculty with PHD. It seems that as the % of faculty with PHD goes up, the data values go more above and below the regression model.



I made this matrix using the gridExtra package. To do this you have to create individual scatterplots. It can be using basic R commands, ggplot, ggscatter, or whatever your preference is. You give each scatterplot a name and then put each scatterplot you made in this code: grid.arrange() ex = grid.arrange(a,b,c,d,e,f,g,h,i).

3. Investigate the correlation amongst the explanatory variables. Suggest a creative course of action (rather than simply omitting a variable) for dealing with any medium or strong correlations encountered. Describe any danger from leaving correlated variables in the model. Describe any danger from simply omitting variables.



Looking at the table below, the variable strongest correlated to tuition is whether a school public or private (.6). Excluding or target variable, the other moderately strong correlations are "% of professors with PHD" and "average faculty compensation". These variables can potentially weaken my model because I'm not adding incremental information. Instead, I'm adding 'noise' to my model. My adjusted R squared would be lowered.

A creative way of dealing with variables like this is to use forward/backward or stepwise selection. I can let my computer go through variables and pick which variables to use for my model. If 2 variables are strongly correlated, my model would most likely only choose one of them. This is better than me randomly/using bias to get rid of predictor variables.

```
                tuition   pcttop25   sf_ratio   fac_comp     accrate     graduat    pct_phd    fulltime     alumni    num_enrl public.private
tuition       1.0000000  0.4830180 -0.5856490  0.3978552 -0.27404612  0.59447510  0.3777172  0.26884191  0.5705928 -0.20024956     0.60075799
pcttop25      0.4830180  1.0000000 -0.3002371  0.5485081 -0.45109961  0.46130070  0.5464996  0.39024403  0.3977497  0.21593178     0.12055549
sf_ratio     -0.5856490 -0.3002371  1.0000000 -0.1009158  0.12140749 -0.35866597 -0.1198611 -0.10359836 -0.4257300  0.27642643    -0.52260215
fac_comp      0.3978552  0.5485081 -0.1009158  1.0000000 -0.52531029  0.29885150  0.6829224  0.22488487  0.1232102  0.47739373    -0.21596288
accrate      -0.2740461 -0.4510996  0.1214075 -0.5253103  1.00000000 -0.28385892 -0.3518935 -0.14021978 -0.1471670 -0.14631640     0.07217343
graduat       0.5944751  0.4613007 -0.3586660  0.2988515 -0.28385892  1.00000000  0.2780175  0.27864568  0.4842101 -0.07061861     0.41923004
pct_phd       0.3777172  0.5464996 -0.1198611  0.6829224 -0.35189355  0.27801750  1.0000000  0.31437178  0.2096451  0.33524516    -0.14668469
fulltime      0.2688419  0.3902440 -0.1035984  0.2248849 -0.14021978  0.27864568  0.3143718  1.00000000  0.2815248  0.14462519     0.05039286
alumni        0.5705928  0.3977497 -0.4257300  0.1232102 -0.14716702  0.48421014  0.2096451  0.28152483  1.0000000 -0.20197662     0.45409907
num_enrl     -0.2002496  0.2159318  0.2764264  0.4773937 -0.14631640 -0.07061861  0.3352452  0.14462519 -0.2019766  1.00000000    -0.57057389
public.private  0.6007580  0.1205555 -0.5226022 -0.2159629  0.07217343  0.41923004 -0.1466847  0.05039286  0.4540991 -0.57057389     1.00000000
```

4. Make a new data frame, modtuitiontrain that contains only the records in tuitiontrain that have complete cases. (How many records are in this data frame?) You will use these data for questions 5 and 6. As mentioned in question 1, I have 804 observations (11 variables including tuition) in my natution data frame compared to 1284 observations when I have the default data frame.

5. Create a model using the forward selection based on the partial F test to select the variables. Show how the model changes each step along the way. What is your final model? Interpret the coefficients of this model in the context of tuition

```
#Forward Selection step by step
tuition_empty <- lm(tuition ~ 1,data = tuitiontrain)

tuition_empty
(StepF1 <- add1(tuition_empty, scope = tuitiontrain[,1:11], test = "F", trace = TRUE))
tuition_empty1 <- lm(tuition ~ public.private,data = tuitiontrain)
summary(tuition_empty1)
(StepF2 <- add1(tuition_empty1, scope = tuitiontrain[,1:11], test = "F", trace = TRUE))
tuition_empty2 <- lm(tuition ~ public.private+ fac_comp,data = tuitiontrain)
summary(tuition_empty2)
(StepF3 <- add1(tuition_empty2, scope = tuitiontrain[,1:11], test = "F", trace = TRUE))

tuition_empty3 <- lm(tuition ~ public.private+ fac_comp + alumni ,data = tuitiontrain)

summary(tuition_empty3)

(StepF4 <- add1(tuition_empty3, scope = tuitiontrain[,1:11], test = "F", trace = TRUE))

tuition_empty4 <- lm(tuition ~ public.private+ fac_comp + alumni +sf_ratio ,data = tuitiontrain)

summary(tuition_empty4)

(StepF5 <- add1(tuition_empty4, scope = tuitiontrain[,1:11], test = "F", trace = TRUE))

tuition_empty5 <- lm(tuition ~ public.private+ fac_comp + alumni +sf_ratio +graduat ,data = tuitiontrain)

summary(tuition_empty5)

(StepF6 <- add1(tuition_empty5, scope = tuitiontrain[,1:11], test = "F", trace = TRUE))

tuition_empty6 <- lm(tuition ~ public.private+ fac_comp + alumni +sf_ratio +graduat + pct_phd ,data = tuitiontrain)

summary(tuition_empty6)

(StepF7 <- add1(tuition_empty6, scope = tuitiontrain[,1:11], test = "F", trace = TRUE))

tuition_empty7 <- lm(tuition ~ public.private+ fac_comp + alumni +sf_ratio +graduat+pct_phd + num_enrl ,data = tuitiontrain)

summary(tuition_empty7)

(StepF8 <- add1(tuition_empty7, scope = tuitiontrain[,1:11], test = "F", trace = TRUE))

tuition_empty8 <- lm(tuition ~ public.private+ fac_comp + alumni +sf_ratio +graduat +pct_phd + num_enrl + fulltime ,data = tuitiontrain)

summary(tuition_empty8)

(StepF9 <- add1(tuition_empty8, scope = tuitiontrain[,1:11], test = "F", trace = TRUE))
summary(tuition_empty8)
```

```
> (StepF9 <- add1(tuition_empty8, scope = tuitiontrain[,1:11], test = "F", trace = TRUE))
Single term additions

Model:
tuition ~ public.private + fac_comp + alumni + sf_ratio + graduat +
    pct_phd + num_enrl + fulltime
         Df Sum of Sq        RSS    AIC F value Pr(>F)
<none>                2975750400 9900.9
pcttop25  1   1422145 2974328256 9902.6  0.3031 0.5821
accrate   1   1787309 2973963091 9902.5  0.3810 0.5373
> summary(tuition_empty8)

Call:
lm(formula = tuition ~ public.private + fac_comp + alumni + sf_ratio +
    graduat + pct_phd + num_enrl + fulltime, data = tuitiontrain)

Residuals:
   Min     1Q Median     3Q    Max
 -9183  -1309    -59   1271  11081

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -3.093e+03  8.326e+02  -3.715 0.000221 ***
public.private  4.071e+03  2.763e+02  14.736  < 2e-16 ***
fac_comp        1.303e-01  1.088e-02  11.972  < 2e-16 ***
alumni          4.531e+01  8.490e+00   5.338 1.31e-07 ***
sf_ratio       -1.619e+02  2.575e+01  -6.289 5.95e-10 ***
graduat         3.048e+01  6.272e+00   4.860 1.48e-06 ***
pct_phd         3.149e+01  7.751e+00   4.063 5.45e-05 ***
num_enrl       -3.811e-01  1.188e-01  -3.207 0.001411 **
fulltime        1.203e+01  5.841e+00   2.059 0.039886 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2165 on 635 degrees of freedom
Multiple R-squared:  0.7339,    Adjusted R-squared:  0.7305
F-statistic: 218.9 on 8 and 635 DF,  p-value: < 2.2e-16
```

My final model includes the significant predictor variables: public.private+ fac_comp + alumni +sf_ratio +graduat +pct_phd + num_enrl + fulltime on the target variable tuition.

When a university is private, tuition is expected to increase by $4000

For every 1 increase in faculty compensation, we expect tuition to go up by $13

For every 1% increase of alumni that donate we expect tuition to go up by $45

For every 1 increase of student to faculty ratio, we expect tuition to go down by $160

For every 1% of students that graduate, we expect tuition to go up by $30

For every 1% increase in faculty with PHD, we expect tuition to go up by $31

For every new student enrolled, we expect tuition to go down by $38

For every 1 % increase of students who are full time, we expect tuition to go up by $12

6. Use the stepAIC command – forward, backward, and both – to create 3 models. Investigate the differences in the models, if any, among the three different methods, stepwise, backwards, and forwards. Construct a table showing method, variables included, *AIC*, and the standard error of the estimate. Which model do you prefer and why?

```
> summary(M_Both)

Call:
lm(formula = tuition ~ public.private + fac_comp + alumni + sf_ratio +
    graduat + pct_phd + num_enrl + fulltime, data = tuitiontrain)

Residuals:
   Min     1Q Median     3Q    Max
 -9183  -1309    -59   1271  11081

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -3.093e+03  8.326e+02  -3.715 0.000221 ***
public.private  4.071e+03  2.763e+02  14.736  < 2e-16 ***
fac_comp        1.303e-01  1.088e-02  11.972  < 2e-16 ***
alumni          4.531e+01  8.490e+00   5.338 1.31e-07 ***
sf_ratio       -1.619e+02  2.575e+01  -6.289 5.95e-10 ***
graduat         3.048e+01  6.272e+00   4.860 1.48e-06 ***
pct_phd         3.149e+01  7.751e+00   4.063 5.45e-05 ***
num_enrl       -3.811e-01  1.188e-01  -3.207 0.001411 **
fulltime        1.203e+01  5.841e+00   2.059 0.039886 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2165 on 635 degrees of freedom
Multiple R-squared:  0.7339,    Adjusted R-squared:  0.7305
F-statistic: 218.9 on 8 and 635 DF,  p-value: < 2.2e-16
```

```
> summary(M_Backwards)

Call:
lm(formula = tuition ~ sf_ratio + fac_comp + graduat + pct_phd +
    fulltime + alumni + num_enrl + public.private, data = tuitiontrain)

Residuals:
   Min     1Q Median     3Q    Max
 -9183  -1309    -59   1271  11081

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -3.093e+03  8.326e+02  -3.715 0.000221 ***
sf_ratio       -1.619e+02  2.575e+01  -6.289 5.95e-10 ***
fac_comp        1.303e-01  1.088e-02  11.972  < 2e-16 ***
graduat         3.048e+01  6.272e+00   4.860 1.48e-06 ***
pct_phd         3.149e+01  7.751e+00   4.063 5.45e-05 ***
fulltime        1.203e+01  5.841e+00   2.059 0.039886 *
alumni          4.531e+01  8.490e+00   5.338 1.31e-07 ***
num_enrl       -3.811e-01  1.188e-01  -3.207 0.001411 **
public.private  4.071e+03  2.763e+02  14.736  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2165 on 635 degrees of freedom
Multiple R-squared:  0.7339,    Adjusted R-squared:  0.7305
F-statistic: 218.9 on 8 and 635 DF,  p-value: < 2.2e-16
```

```
> summary(M_Forward)

Call:
lm(formula = tuition ~ public.private + fac_comp + alumni + sf_ratio +
    graduat + pct_phd + num_enrl + fulltime, data = tuitiontrain)

Residuals:
   Min     1Q Median     3Q    Max
 -9183  -1309    -59   1271  11081

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -3.093e+03  8.326e+02  -3.715 0.000221 ***
public.private  4.071e+03  2.763e+02  14.736  < 2e-16 ***
fac_comp        1.303e-01  1.088e-02  11.972  < 2e-16 ***
alumni          4.531e+01  8.490e+00   5.338 1.31e-07 ***
sf_ratio       -1.619e+02  2.575e+01  -6.289 5.95e-10 ***
graduat         3.048e+01  6.272e+00   4.860 1.48e-06 ***
pct_phd         3.149e+01  7.751e+00   4.063 5.45e-05 ***
num_enrl       -3.811e-01  1.188e-01  -3.207 0.001411 **
fulltime        1.203e+01  5.841e+00   2.059 0.039886 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2165 on 635 degrees of freedom
Multiple R-squared:  0.7339,    Adjusted R-squared:  0.7305
F-statistic: 218.9 on 8 and 635 DF,  p-value: < 2.2e-16
```

Looking at the 3 models created by the AIC command, I can tell that all 3 models have the same exact predictor variables. The only difference is that they're added in different orders. The standard error for all the predictor variables & the target variable is the same. All the models have the same R^2 of .7339 and adjusted R^2 of .7305. The residual standard error is also the same on all 3 models with a residual standard error of 2165 on 635 degrees of freedom. I don't have a preference of any model because they're all displaying the same information. I like Forward selection the most, it makes the most sense to me.

7. a. Missing data appear to be a problem with this data set. Prepare a new data frame which is a copy of the original dataset but where the missing values are each replaced with their field means. (Name this data frame mod2). Report on how this substitution has affected the fields (summary stats, etc), if at all. What do you think of this method of dealing with missing

values? (Can you suggest a better method, which does not rely on complicated programming?)

```
> summary(tuition)
    tuition        pcttop25        sf_ratio        fac_comp        accrate          graduat         pct_phd          fulltime        alumni          num_enrl       public.private
 Min.   : 1044   Min.   :  6.00   Min.   : 2.30   Min.   : 26500   Min.   :0.1540   Min.   :  8.00   Min.   :  8.00   Min.   :11.43   Min.   : 0.00   Min.   :  18.0   Min.   :0.0000
 1st Qu.: 6114   1st Qu.: 37.00   1st Qu.:11.80   1st Qu.: 43600   1st Qu.:0.6837   1st Qu.: 47.00   1st Qu.: 57.00   1st Qu.:68.59   1st Qu.:11.00   1st Qu.: 234.5   1st Qu.:0.0000
 Median : 8670   Median : 50.00   Median :14.30   Median : 50900   Median :0.7840   Median : 60.00   Median : 71.00   Median :83.42   Median :19.00   Median : 446.5   Median :1.0000
 Mean   : 9284   Mean   : 52.28   Mean   :14.87   Mean   : 52680   Mean   :0.7581   Mean   : 60.42   Mean   : 68.72   Mean   :78.79   Mean   :20.92   Mean   : 782.4   Mean   :0.6438
 3rd Qu.:11675   3rd Qu.: 65.00   3rd Qu.:17.60   3rd Qu.: 60100   3rd Qu.:0.8610   3rd Qu.: 74.00   3rd Qu.: 82.00   3rd Qu.:91.88   3rd Qu.:29.00   3rd Qu.: 984.2   3rd Qu.:1.0000
 Max.   :25750   Max.   :100.00   Max.   :91.80   Max.   :107500   Max.   :1.0000   Max.   :118.00   Max.   :103.00   Max.   :99.94   Max.   :81.00   Max.   :7425.0   Max.   :1.0000
 NA's   :1       NA's   :197      NA's   :3       NA's   :163      NA's   :12       NA's   :96       NA's   :32       NA's   :28      NA's   :214     NA's   :4        NA's   :1
> summary(mod2)
    tuition        pcttop25        sf_ratio        fac_comp        accrate          graduat         pct_phd          fulltime        alumni          num_enrl       public.private
 Min.   : 1044   Min.   :  6.00   Min.   : 2.30   Min.   : 26500   Min.   :0.1540   Min.   :  8.00   Min.   :  8.00   Min.   :11.43   Min.   : 0.00   Min.   :  18.0   Min.   :0.0000
 1st Qu.: 6117   1st Qu.: 39.00   1st Qu.:11.80   1st Qu.: 44700   1st Qu.:0.6850   1st Qu.: 48.00   1st Qu.: 57.00   1st Qu.:68.91   1st Qu.:12.00   1st Qu.: 235.8   1st Qu.:0.0000
 Median : 8673   Median : 52.28   Median :14.30   Median : 52680   Median :0.7820   Median : 60.42   Median : 70.00   Median :82.89   Median :20.92   Median : 449.5   Median :1.0000
 Mean   : 9284   Mean   : 52.28   Mean   :14.87   Mean   : 52680   Mean   :0.7581   Mean   : 60.42   Mean   : 68.72   Mean   :78.79   Mean   :20.92   Mean   : 782.4   Mean   :0.6438
 3rd Qu.:11668   3rd Qu.: 63.00   3rd Qu.:17.52   3rd Qu.: 58500   3rd Qu.:0.8602   3rd Qu.: 72.00   3rd Qu.: 82.00   3rd Qu.:91.59   3rd Qu.:26.00   3rd Qu.: 981.0   3rd Qu.:1.0000
 Max.   :25750   Max.   :100.00   Max.   :91.80   Max.   :107500   Max.   :1.0000   Max.   :118.00   Max.   :103.00   Max.   :99.94   Max.   :81.00   Max.   :7425.0   Max.   :1.0000
```

I can see that by replacing the N/A values with mean values, I am reducing the variance in the data. I can see that what's mostly being affected in Q1 & Q3. They're getting closer together. This makes me visualize a boxplot of both the original and new data and see that the box is tighter compared to the old one. Another method can be asking an organization if they can find the missing data so we can have a complete data set.

b. Now fit the model with the variables that you chose in question 6 as your preferred model to the data in mod2. Investigate the differences in the results, if any, between the two models. Construct a table showing method, variables included, *AIC*, and the standard error of the estimate. In this situation, which model do you prefer and why?

```
> summary(M_Forward2)

Call:
lm(formula = tuition ~ graduat + public.private + fac_comp +
    pct_phd + sf_ratio + alumni + pcttop25 + num_enrl + fulltime,
    data = mod2)

Residuals:
    Min      1Q  Median      3Q     Max
-9361.0 -1385.4    89.4  1404.2 12142.8

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -4.456e+03  5.535e+02  -8.051 1.87e-15 ***
graduat         4.196e+01  4.669e+00   8.987  < 2e-16 ***
public.private  3.892e+03  1.951e+02  19.950  < 2e-16 ***
fac_comp        1.081e-01  8.046e-03  13.436  < 2e-16 ***
pct_phd         3.930e+01  4.987e+00   7.881 6.90e-15 ***
sf_ratio       -1.007e+02  1.421e+01  -7.088 2.26e-12 ***
alumni          4.200e+01  6.834e+00   6.147 1.06e-09 ***
pcttop25        8.946e+00  4.637e+00   1.929   0.0539 .
num_enrl       -1.894e-01  9.923e-02  -1.909   0.0565 .
fulltime        7.662e+00  4.407e+00   1.738   0.0824 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2361 on 1274 degrees of freedom
Multiple R-squared:  0.6827,    Adjusted R-squared:  0.6804
F-statistic: 304.5 on 9 and 1274 DF,  p-value: < 2.2e-16
```

```
> summary(M_Forward)

Call:
lm(formula = tuition ~ public.private + fac_comp + alumni + sf_ratio +
    graduat + pct_phd + num_enrl + fulltime, data = tuitiontrain)

Residuals:
   Min     1Q Median     3Q    Max
 -9183  -1309    -59   1271  11081

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -3.093e+03  8.326e+02  -3.715 0.000221 ***
public.private  4.071e+03  2.763e+02  14.736  < 2e-16 ***
fac_comp        1.303e-01  1.088e-02  11.972  < 2e-16 ***
alumni          4.531e+01  8.490e+00   5.338 1.31e-07 ***
sf_ratio       -1.619e+02  2.575e+01  -6.289 5.95e-10 ***
graduat         3.048e+01  6.272e+00   4.860 1.48e-06 ***
pct_phd         3.149e+01  7.751e+00   4.063 5.45e-05 ***
num_enrl       -3.811e-01  1.188e-01  -3.207 0.001411 **
fulltime        1.203e+01  5.841e+00   2.059 0.039886 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2165 on 635 degrees of freedom
Multiple R-squared:  0.7339,    Adjusted R-squared:  0.7305
F-statistic: 218.9 on 8 and 635 DF,  p-value: < 2.2e-16
```

The first screenshot is the model with the data frame that replaces the N/A values with the mean of each column. the second screenshot contains the dataframe with the removed values using the training dataset. I prefer the model with the removed N/A values because it has one less predictor variable and the R^2 + the adjusted R^2 are both higher. The residual standard error is also a lot less in this model, and all the predictor variables chosen are considered significant while the model with mod 2 has insignificant predictor variables.

8. Now, select a model from those discussed above. Apply the model to predict tuition using the values of the explanatory variables in the tuitiontest data. (How did you handle the missing values?) Assess the accuracy of your model in terms of predicting tuitions. Give me a file that contains your predictions in addition to your assessment of accuracy.

The model I chose from was the forward selection model using AIC with these predictor variables.

```
M_Train <- lm(formula = tuition ~ public.private + fac_comp + alumni + sf_ratio +
        Graduat + pct_phd + num_enrl + fulltime, data = tuitiontrain)
```

I used the data frame that omitted the n/a values from my data frame because when I used the mean in place of N/A values, and I created a model using forward selection, I had more predictor variables in my model. That model had a bigger standard error and smaller r^2/adj r^2.

```
> TuitionPred <- predict(M_Train,newdata = tuitiontest)
> actuals_preds <- data.frame(cbind(actuals = tuitiontest$public.private + tuitiontest$pct_phd + tuitiontest$fulltime + tuitiontest$num_enrl+ tuit
iontest$graduat+ tuitiontest$fac_comp+tuitiontest$alumni +tuitiontest$sf_ratio,predicteds= TuitionPred))
> head(actuals_preds)
    actuals predicteds
6  57949.16   6828.961
8  56721.51   4686.590
11 36936.42   8558.478
38 48883.17   8582.880
53 56758.48  12949.134
56 44197.85   9886.736
```

This is the head of some of the actual vs predicted results.

```
> min_max_accuracy <- mean(apply(actuals_preds,1,min)/apply(actuals_preds,1,max))
> min_max_accuracy
[1] 0.192926
> mape <- mean(abs((actuals_preds$predicteds - actuals_preds$actuals))/actuals_preds$actuals)
> mape
[1] 0.807074
```

When I conducted an accuracy test on my model, I got that my model is 19.29% accurate and has a mean absolute percent error of 80.71%