STA141A: Analysis on Graduation Rate

Name	Expected Contribution	UC Davis Email
Keying Liu	Linear model and report writing	kyliu@ucdavis.edu
Xinhui Luo	Linear Regression and General Editing	xinluo@ucdavis.edu
Natalie Aceves	Data Visualization and Report Writing	ncacevesgarcia@ucdavis.edu
Leah Khan	Report Writing and Plot Formatting	lkhan@ucdavis.edu

Instructor: Emanuela Furfaro

STA 141A - Fundamentals of Statistical Data Science

University of California, Davis

September 7, 2022

Introduction & Research Question

Graduation refers to the successful completion of a course of study at a university, college, or school, for which one receives a degree or diploma. Graduation rate, on the other hand, is an important indicator to consider when students are choosing a school. High graduation rates could infer higher quality education, while low graduation rates can infer education of lower quality. So, it is important to determine the factors that might affect graduation rates the most. And further, we want to understand why they are important. In this project we are going to determine the relationships between "College" from the "ISLR" package. This dataset came from US Colleges from the 1995 issue of US News and World Report and from the StatLib library which is maintained at Carnegie Mellon University.

In this project, to better analyze the data, we will use the following research questions to guide us through the process of studying:
1.Which attribute has the most impact on the graduation rate? How does it impact the graduation rate?
2. Is linear regression model a good fit for this data ?
3. Is there any common result given by both methods?

Data Description

The dataset contains 777 observations from different universities consisting of 18 variables. For this project we are going to use graduation rate as a response, and mainly focus on predictors: "enrollment rate", "top10%", "top25%", "F.Undergrad", "P.Undergrad". "S.F.Ratio", and "Expend Instructional."

Variable	Description	Variable	Description
Private (Yes/No)	Private/Public University	Top 10 Perc	% of New students from top 10% high school
Apps	# of application received	Top 25 Perc	% of New students from top 25% high school
Accept	# of application accepted	F.Undergrad	# of full time undergraduates
Enroll	# of new students enrolled	P.Undergrad	# of part time undergraduates
OutState	Out of state tuition	Room & Board	Room and board costs
Books	Estimated book costs	phD	% of faculty with ph.D,'s
Personal	Estimated personal spending	Terminal	% of faculty with terminal degree
S.F.Ratio	student/faculty ratio	perc.alumni	% of alumni who donate
Expend Instructional	Instructional expenditure per student	Grad. Rate	Graduation rate

Data Visualization

The bivariate plots indicate that our predictors and response are all quantitative variables, while the plots are scatter plots showing the density of data distribution.

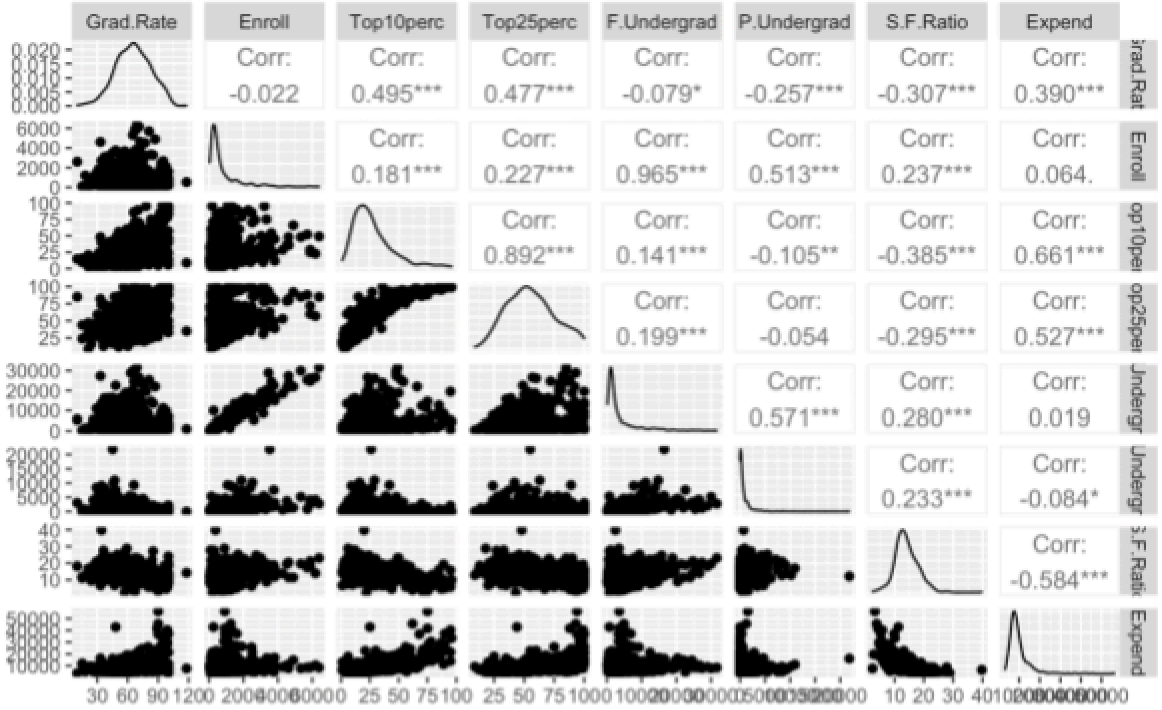
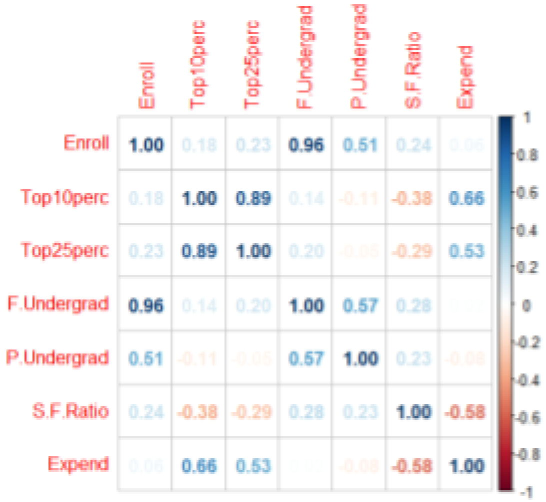


Figure 1. distribution and correlation plot for predictors and response(total of 8 variables). Univariate plots along the diagonal and bivariate plots on bottom left, and correlation at the top left with p value indicating their significance.



	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	S.F.Ratio	Expend
Enroll	1.00	0.18	0.23	0.96	0.51	0.24	0.06
Top10perc	0.18	1.00	0.89	0.14	-0.11	-0.38	0.66
Top25perc	0.23	0.89	1.00	0.20	-0.05	-0.29	0.53
F.Undergrad	0.96	0.14	0.20	1.00	0.57	0.28	0.02
P.Undergrad	0.51	-0.11	-0.05	0.57	1.00	0.23	-0.08
S.F.Ratio	0.24	-0.38	-0.29	0.28	0.23	1.00	-0.58
Expend	0.06	0.66	0.53	0.02	-0.08	-0.58	1.00

Figure 2. Correlation matrix highlighting the correlation relationship between predictors

In this plot, we can see that the “*Top10perc*” and “*Top25Perc*” are very highly correlated. The same can be said for “*F.Undergrad*” and “*Enroll*”. This seems to cause multicollinearity. To fix this we decide to replace “*Top10Perc*” and “*F.Undergrad*” with two new variables: “*Outstate*” and “*Room.Board*.

Supervised Learning Analysis: Linear Regression

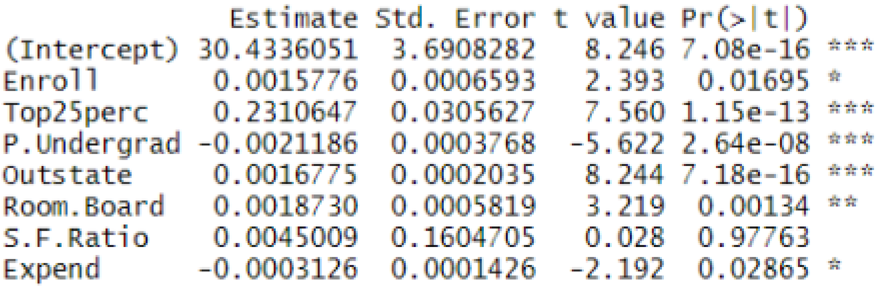
We use a regression model to explain the graduation rate for our data titled “Statistics for a large number of US Colleges from 1995 Issue of US News and World Report.” To build a supervised model which fits the data, we first consider the linear model in following format:

$$GraduationRate = \beta_0 * Enroll + \beta_1 * Top10perc + \beta_2 * Top25perc + \beta_3 * F.Undergrad + \beta_4 * P.Undergrad + \beta_5 * S.F.Ratio + \beta_6 * Expend$$

However, from Figure 2. We discovered that because of the high correlation between *Top10perc* and *Top25Perc*, and *F.Undergrad* and *Enroll,* it may have caused multicollinearity. Thus we then consider a new linear model in following format:

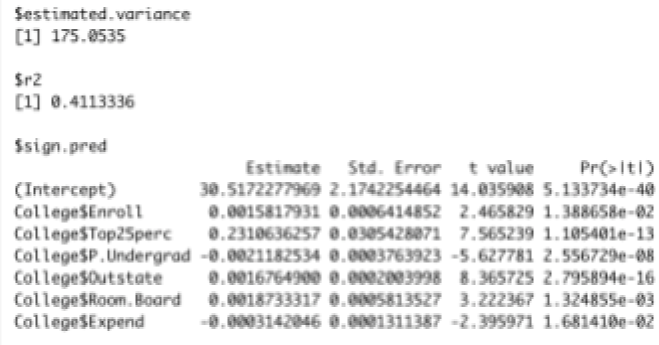
$$GraduationRate = \beta_0 * Enroll + \beta_1 * Outstate + \beta_2 * Top25perc + \beta_3 * Room.Board + \beta_4 * P.Undergrad + \beta_5 * S.F.Ratio + \beta_6 * Expend$$

Table 1. Summary for new linear model



```
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 30.4336051  3.6908282   8.246 7.08e-16 ***
Enroll       0.0015776  0.0006593   2.393  0.01695 *  
Top25perc    0.2310647  0.0305627   7.560 1.15e-13 ***
P.Undergrad -0.0021186  0.0003768  -5.622 2.64e-08 ***
Outstate     0.0016775  0.0002035   8.244 7.18e-16 ***
Room.Board   0.0018730  0.0005819   3.219  0.00134 ** 
S.F.Ratio    0.0045009  0.1604705   0.028  0.97763    
Expend      -0.0003126  0.0001426  -2.192  0.02865 *  
```

The summary suggests that S.F.Ratio is not significant to the linear model, with p-value be 0.97763.(Table 1) We remove the S.F.Ratio and continue to calculate different sets of possible linear regression model. For example, we test for only the significant predictors and compare their estimated variance and multiple R-squared to determine which model fits better.



```
$estimated.variance
[1] 175.0535

$r2
[1] 0.4113336

$sign.pred
                       Estimate   Std. Error  t value     Pr(>|t|)
(Intercept)         30.5172277969 2.1742254464 14.035908 5.133734e-40
College$Enroll       0.0015817931 0.0006414852  2.465829 1.388658e-02
College$Top25perc    0.2310636257 0.0305428071  7.565239 1.105401e-13
College$P.Undergrad -0.0021182534 0.0003763923 -5.627781 2.556729e-08
College$Outstate     0.0016764900 0.0002003998  8.365725 2.795894e-16
College$Room.Board   0.0018733317 0.0005813527  3.222367 1.324855e-03
College$Expend      -0.0003142046 0.0001311387 -2.395971 1.681410e-02
```

Figure 3. Final linear model with the unbiased estimator of variance and R2 value as well as the coefficient of those predictors.

The final linear model we obtained that best fit the data is the model in which we remove the S.F.Ratio predictors(Figure 3). And we obtain the estimated variance is 175.0535, with about 41.13% of the data can be explained by our final model, and all of those predictors are significant at level of 0.05. Thus we have our final linear regression function of :

$$GraduationRate = 30.5172 + 0.0016 * Enroll + 0.2311 * Top25perc + 0.0017 * Outstate - 0.0021 * P.Undergrad + 0.0019 * Room.Board - 0.0003 * Expend$$

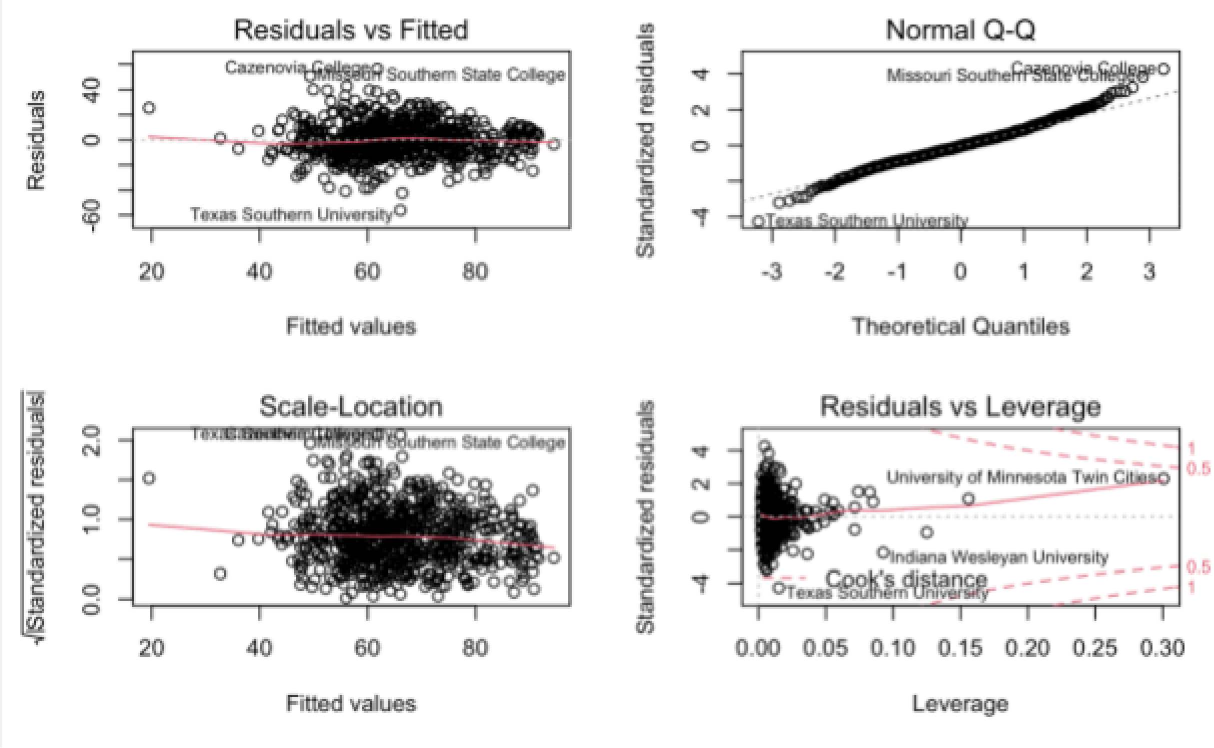
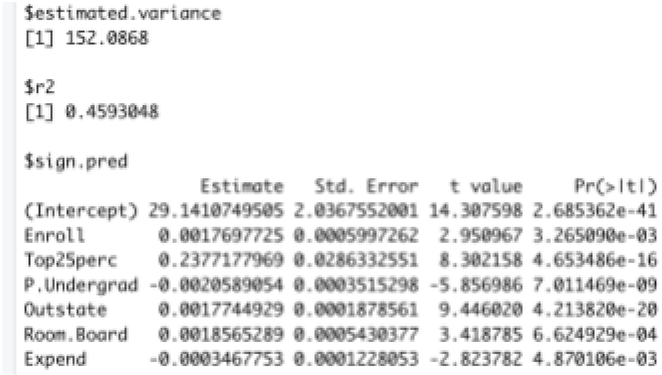


Figure 4. Residual Plot for final reduced linear regression model, removed S.F. Ratio from predictors.

From the Normal QQ plot(Figure 4), we can see that the linear regression model is a good fit for this data. Most of the data are on the straight line except the two tails. However, we are able to spot many outliers from Residual vs Leverage plot. And from Residual vs Fitted plot, we can discover many data are not distributed evenly around zero.



```
$estimated.variance
[1] 152.0868

$r2
[1] 0.4593048

$sign.pred
                 Estimate   Std. Error   t value     Pr(>|t|)
(Intercept) 29.1410749505 2.0367552001 14.307598 2.685362e-41
Enroll       0.0017697725 0.0005997262  2.950967 3.265090e-03
Top25perc    0.2377177969 0.0286332551  8.302158 4.653486e-16
P.Undergrad -0.0020589054 0.0003515298 -5.856986 7.011469e-09
Outstate     0.0017744929 0.0001878561  9.446020 4.213820e-20
Room.Board   0.0018565289 0.0005430377  3.418785 6.624929e-04
Expend      -0.0003467753 0.0001228053 -2.823782 4.870106e-03
```

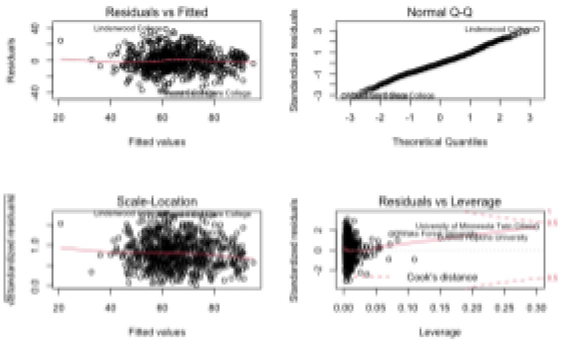


Figure 5. summary and residual plot comparison after removing outliers from final reduced model.

In addition to the best fit linear regression model we find, removing outliers from the final model also helps explain more of the data from 41.13% to 45.93% and reduce the estimated variance from 175 to 152(Figure 5)

Unsupervised Learning Analysis

since classification method might not work with our data since our response is numeric but not categorical, we decided to use unsupervised learning analysis to better examine our data.

Hierarchical clustering Analysis

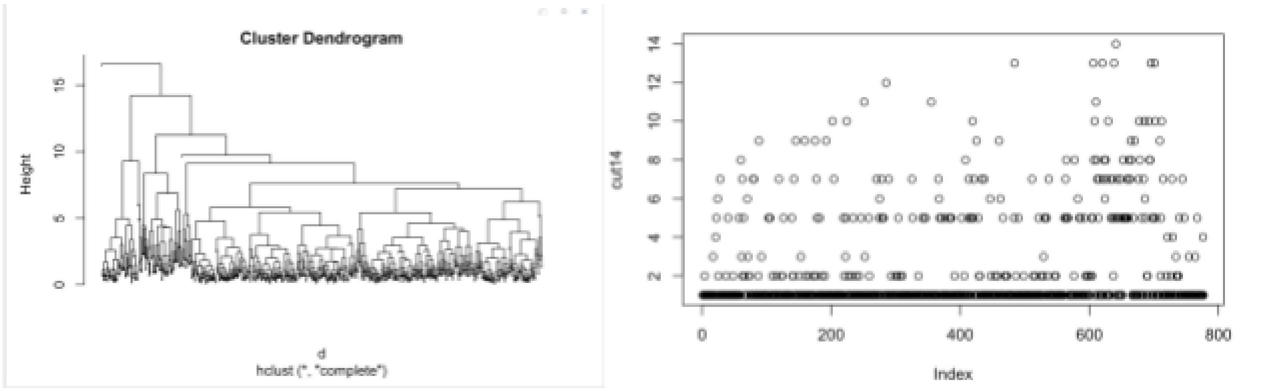
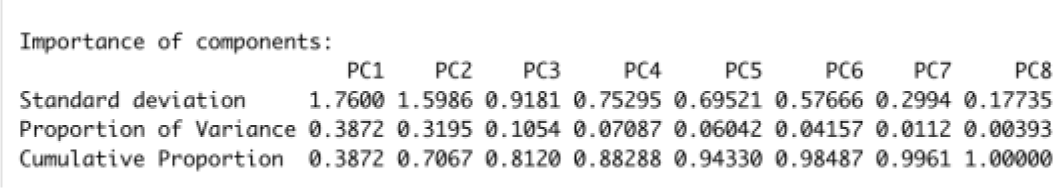


Figure 6. Dendrogram for College data

Looking at the cluster dendrogram(Figure 6), it is hard to interpret anything due to its tremendous amount of ending branches. We decided to use "*cutress*" formula and divided our data into 14 groups(Figure 6). And this returns us to the right side picture. It appears that the majority of data falls within groups 1, 2, 5, and 7. There is an approximate trend: the higher the cut number is, the fewer data in that cut.

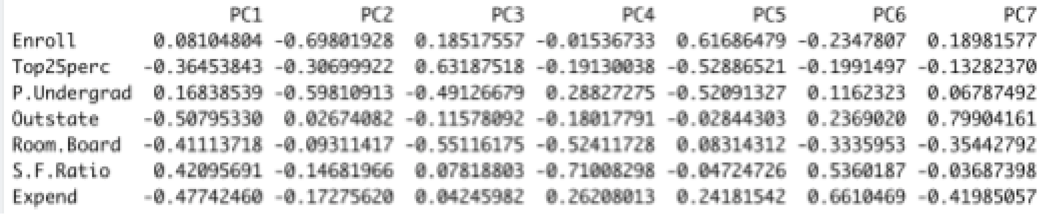
Given a large amount of dataset, it is hard to analysis it as it will produce massive plots which are difficult to make conclusions, therefore beside hierarchical clustering analysis, we decided to use principal component analysis to help us gain better understanding in the dataset.

Principal Component Analysis



```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6    PC7     PC8
Standard deviation     1.7600 1.5986 0.9181 0.75295 0.69521 0.57666 0.2994 0.17735
Proportion of Variance 0.3872 0.3195 0.1054 0.07087 0.06042 0.04157 0.0112 0.00393
Cumulative Proportion  0.3872 0.7067 0.8120 0.88288 0.94330 0.98487 0.9961 1.00000
```

Figure 7. Summary of PCA showing sd and proportion of variance for each principle component



```
                   PC1         PC2         PC3         PC4         PC5        PC6         PC7
Enroll      0.08104804 -0.69801928  0.18517557 -0.01536733  0.61686479 -0.2347807  0.18981577
Top25perc  -0.36453843 -0.30699922  0.63187518 -0.19130038 -0.52886521 -0.1991497 -0.13282370
P.Undergrad 0.16838539 -0.59810913 -0.49126679  0.28827275 -0.52091327  0.1162323  0.06787492
Outstate   -0.50795330  0.02674082 -0.11578092 -0.18017791 -0.02844303  0.2369020  0.79904161
Room.Board -0.41113718 -0.09311417 -0.55116175 -0.52411728  0.08314312 -0.3335953 -0.35442792
S.F.Ratio   0.42095691 -0.14681966  0.07818803 -0.71008298 -0.04724726  0.5360187 -0.03687398
Expend     -0.47742460 -0.17275620  0.04245982  0.26208013  0.24181542  0.6610469 -0.41985057
```

Figure 8. Loading for PCA using data after removing outlier

The summary of principal component analysis is computed from function prcomp() which data is scaled to have unit variance.(Figure 7), as we can see the first two components together only explain 70% of the total variance, thus in order to explain more than 90% of the total variance we need to consider up to the fifth principal component analysis.

While the loading of the principal component analysis(Figure8.) suggested that *outstate* and *expend* have the largest absolute value of PC1. The two largest loadings in absolute value of the first principal component are between variable *Outstate* and *expend*. This pair has the largest correlation.

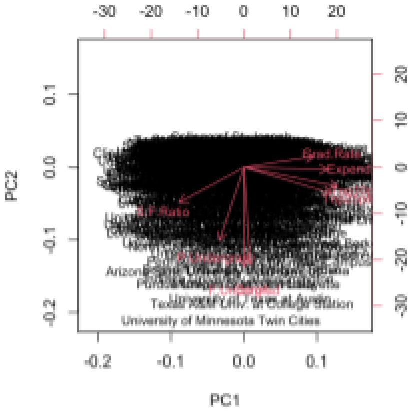


Figure 9. Biplot for PCA showing the significant predictors for pc1 and pc2.

One can see in the bi-plot (Figure 9)the pca that when the *expenditure* is high, the *graduation rates* also high. On the other hand, when S.F. Ratio is high, the *graduation rate* will be low. The plot also suggest that the *undergrate student population* is associated with pc2 while *expend*, top 25 percent and *outstae* have more saying in pc1.

Overall, the result of principal component analysis does seems to support the result from linear regression model where outstate is the most significant predictor to the dataset.

Interpretation of Results

1.Which attribute has the most impact on the graduation rate? How does it impact the graduation rate?

According to Table 1. The attribute that impacts the graduation rate the most will be the predictor *Outstate* given it has the largest absolute value of t-value and smallest p-value, meaning it is the most significant predictors for our model. The coefficient for outstate tuition is 0.0017, at which each unit of increase for *outstate tuition* will cause 0.17% increase in *graduation rate*, assuming other predictors are fixed.

2. Is linear regression model a good fit for this data ?

Linear regression model does seem like a good fit for this data assuming we are using the chosen predictors, given the circumstance that there are many different possible regression models. The final model after removing the outlier can explain up to 45.93% of the data with variance be 152.03(Figure 5). Also according to Figure 5, the normal Q-Q plot also shows that the linear regression model is a good fit for the data by reading the figure.

3. Is there any common result given by both methods?

We discover, from the bi-plot(Figure 9), that the *S.F. Ratio* is pointing in the opposite direction from the graduation rate. This implies that the graduation rates are lower when this variable is high. Also, in the linear regression model we find that the *S.F Ratio* also has the lowest significance which led us to remove it in the final model(Table 1). It can be seen that in the bi-plot we also found that there is a positive relationship between *expenditure* and *graduation rates*. Meanwhile, the linear regression shows a high significance between *graduation rate* and *outstate* (Figure 3).

Conclusion

We used linear regression models and unsupervised learning analysis to analyze our data. Both of them provided us information that was similar yet also different. We are interested in learning what affects graduation rates. Through supervision, we found that *outstate* had a high significance while *U.F Ratio* did not. Through unsupervised, we found that *expenditure* had a high positive relationship while *U.F. Ratio* did not here as well. Due to only 70% of the data being represented, we chose to conclude that the supervised method was the more accurate method. Also, we can see that the QQ plot (Figure 5)resulted in the data making a straight line. This further shows that the data is linearly distributed as it follows the dashed line. Further concluding that the linear regression model fits this dataset best.

Appendix: R Code

```
library(ISLR)
library("corrplot")
data("College")
summary(College$Grad.Rate)
#full model
fmodel = lm(Grad.Rate~ ., data = College)
summary(fmodel)
#subset of the variables
College.sub <- College[,c(4, 5, 6, 7, 8, 15, 17)]
#Corration plot
corrplot(cor(College.sub), method = "number")

# replace new predictor outstate and room.board
College.new <- College[,c(4, 6, 8, 9, 10, 15, 17)]

#Create function to better try out different linear models
# input is a linear model using lm(), this function takes its summary and output a dataframe that
contains the unbaised estimated variance and R2 and coefficient entry that are significant at level 0.05,
as well as plot redisal plots
modelFactory<-function(x){
 a<-summary(x)
output<-list(estimated.variance=a$sigma^2,r2=a$r.squared,sign.pred=a$coefficients[a$coefficients[,4]<
0.05,],4)
 # find predictors with significant level > 0.05
 par(mfrow=c(2,2))
 plot(x)
 return(output)
}

#initial linear model
model1 = lm(College$Grad.Rate ~ ., data = College.new)
modelFactory(model1)
#remove SF ratio
modelFactory(lm(College$Grad.Rate ~
College$Enroll+College$Top25perc+College$P.Undergrad+College$Outstate+College$Room.Board+
College$Expend, data = College.new))
# remove enroll & sf.ratio
modelFactory(lm(College$Grad.Rate
~College$Top25perc+College$P.Undergrad+College$Outstate+College$Room.Board+College$Expend
, data = College.new))
# remove expand and enroll and sf.ratio
modelFactory(lm(College$Grad.Rate
~College$Top25perc+College$P.Undergrad+College$Outstate+College$Room.Board, data =
College.new))
# remove expand and sf.ratio
modelFactory(lm(College$Grad.Rate ~
College$Enroll+College$Top25perc+College$P.Undergrad+College$Outstate+College$Room.Board,
data = College.new))

#Remove outlier
College_red = College.new[abs(rstandard(model1)) <= 3,]
Grad_red = College$Grad.Rate[abs(rstandard(model1)) <= 3]

#initial linear model
modelFactory(lm(Grad_red ~ ., data = College_red))
#remove SF ratio
modelFactory(lm(Grad_red ~ Enroll+Top25perc+P.Undergrad+Outstate+Room.Board+Expend, data =
College_red))
# remove enroll & sf.ratio
modelFactory(lm(Grad_red ~Top25perc+P.Undergrad+Outstate+Room.Board+Expend, data =
College_red))
# remove expand and enroll and sf.ratio
modelFactory(lm(Grad_red ~Top25perc+P.Undergrad+Outstate+Room.Board, data = College_red))
# remove expand and sf.ratio
modelFactory(lm(Grad_red ~ Enroll+Top25perc+P.Undergrad+Outstate+Room.Board, data =
College_red))
```

```
library(ISLR)
data("College")
summary(College)

hca <- prcomp(College.new, scale. = TRUE)
summary(hca)

hca <- prcomp(college, scale. = TRUE)
summary(hca)

print(hca$rotation)
print(hca$sdev)


scaled_data <- scale(College_pca)
correlation_matrix <- cor(scaled_data)
print(correlation_matrix)
correlation_matrix_values <- unique(sort(abs(correlation_matrix),
decreasing = TRUE))
colnames(scaled_data)[which(abs(correlation_matrix) ==
correlation_matrix_values[2], arr.ind = TRUE)[1, ]]

colnames(scaled_data)[which(abs(correlation_matrix) ==
correlation_matrix_values[3], arr.ind = TRUE)[1, ]]
hc = hclust(dist(College_pca), method = "average")
plot(hc, cex =0.05, hang = 0.009)

pca <- prcomp(scaled_data, scale. = TRUE)
summary(pca)
biplot(pca, cex = 0.8)

cut11 = cutree(hc, k = 11)
plot(cut11)


hc_scaled_data <- hclust(dist(scaled_data), method = "complete")
plot(hc_scaled_data, cex = 0.005, hang = 0.009)

cut14 = cutree(hc, k = 14)
plot(cut14)
```

Reference

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) An Introduction to Statistical Learning with applications in R, https://www.statlearning.com, Springer-Verlag, New York