Analysis

Exploratory Data Analysis (EDA)

The first step of conducting analysis is to understand the data by conducting exploratory data analysis. To conduct the EDA, we obtained descriptive statistics and summaries of all variables. For the quantitative variables, we wrote a function called output_quantitative_stats() to get minimum, maximum, range, median, first and third quartiles, IQR, Mean and Sd of all the quantitative variables including "UGDS_BLACK", "UGDS_HISP", "UGDS_ASIAN", "UGDS_ASIAN", "UGDS_AIAN", "UGDS_NHPI", "UGDS_2MOR","UGDS_NRA", "UGDS_UNKN","UGDS_WHITE", "UGDS", "ADM_RATE", "COSTT4_A", "MD_EARN_WNE_P10", "C100_4", "PCTFLOAN", "CIP_SUM", "MINORATIO", and "STU_APPLIED".

Similarly, we wrote a function called output_qualitative_stats() to generate a table with both the frequency and the relative frequency of the qualitative variables including "WEST", "MIDWEST", "NORTHEAST", "SOUTH", "MAJOR_CITY", "MINOQ1", "MINOQ2", "MINOQ3",and "MINOQ4". To understand the data better, we also want to generate some plots to visualize the data. We wrote the functions histogram_generator() and boxplot_generator() to generate histograms and boxplots of the quantitative variables and condition_boxplot_generator() to generate conditional boxplots between "STU_APPLIED" and the qualitative variables. To study the association between "STU_APPLIED" and the rest of predictors, we also obtained the correlation matrix of all quantitative variables using function cor(), the scatterplot matrix using function pairs(), the anova between "STU_APPLIED" and all the qualitative variables using function aov().

Then, we divide our data set into 8 separate clusters according to region and its proximity to major cities. In order to develop strategies to improve competitiveness for each cluster, we first tabulate some key statistics for each cluster.

	WM	WN	MM	MN	NM	NN	SM	SN
Size	117.00	95.00	149.00	261.00	127.00	292.00	230.00	284.00
STU_APP_avg	16505.36	11347.99	8033.18	6553.14	13623.85	7271.64	11538.00	7786.49
STU_APP_sd	28415.84	20056.26	13990.95	9733.35	22181.73	11061.65	16745.93	16166.58
STU_APP_min	112.01	122.68	106.29	105.00	172.26	138.13	170.00	280.88
STU_APP_max	169325.84	137878.19	87521.71	66552.25	121409.40	100035.24	105381.98	221057.35
MD_EARN_avg	43744.44	43400.00	40934.23	40521.84	47893.70	45403.42	40241.30	37694.37
MD_EARN_sd	10596.73	11062.87	8391.54	7613.28	15545.28	11779.51	8518.42	7305.08
MD_EARN_min	22100.00	22000.00	22800.00	18500.00	19300.00	12000.00	23500.00	19900.00
MD_EARN_max	79400.00	86000.00	69300.00	79200.00	113400.00	118800.00	76700.00	64000.00
C100_4_avg	0.38	0.35	0.36	0.39	0.49	0.46	0.32	0.32
C100_4_sd	0.24	0.22	0.20	0.20	0.23	0.22	0.21	0.19
C100_4_min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
C100_4_max	1.00	0.90	1.00	1.00	0.91	0.90	1.00	1.00
PCTFLOAN_avg	0.56	0.52	0.65	0.62	0.57	0.61	0.60	0.62
PCTFLOAN_sd	0.17	0.18	0.15	0.15	0.19	0.19	0.17	0.16
PCTFLOAN_min	0.19	0.00	0.19	0.00	0.04	0.00	0.14	0.15
PCTFLOAN_max	0.92	0.93	0.94	0.93	0.92	0.92	0.97	0.98
MINORITY_avg	0.54	0.46	0.33	0.27	0.45	0.35	0.53	0.42
MINORITY_sd	0.20	0.19	0.16	0.14	0.19	0.19	0.26	0.23
MINORITY_min	0.13	0.15	0.05	0.00	0.06	0.00	0.11	0.03
MINORITY_max	0.92	0.95	0.98	1.00	0.99	0.99	1.00	1.00

Table 1: Cluster Statistics

The table contains the number of institutions in each cluster, with Northeast region not located in major city cluster having the most instituions. And since we are interested in predicting students applied by variables

such as earnings, graduation rate, minority ratio and percentage of students with loans, we look at the mean for this variable in each cluster first.

ANOVA is designed to test whether there are any statistically significant differences between the means of independent groups. Since our measure of school competitiveness is the number of students applied, we will test whether the means of students applied is different among clusters to gauge whether our clustering criteria makes sense.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cluster	1	1923277409.64	1923277409.64	6.90	0.0087
Residuals	1553	432980783904.29	278802822.86		

Table 2: ANOVA Test Result

The test result shows that we have a p-value smaller than 0.01, which means that we can reject the null hypothesis that the means are the same across all 8 clusters.

Then, after the premilinary EDA, we start to run regression and explore the relationship between variables.

To start with, we run an OLS regression for all variables that we believe have an impact on number of students applied.

Students applied = Median_Earning + Completion_rate + Percentage_with_Student_Loans + Major_City + Minority_Ratio + West + Midwest + Northeast

The dependent variable is number of students applied in the fall term. It is calculated by dividing people admitted during fall term by the admission rate.

We use 6 main variables as regressors in the regression:

1. Median_Earning: Student's median earning 10 years after graduation
2. Completion_rate: Percentage of students graduated within 4 years.
3. Percentage_with_Student_Loans: Percentage of students with Student Loans.
4. Major_City: Whether the institution is located near a major city or countryside. This dummy variable equals 1 if it is located in a major city, 0 if countryside.
5. Minority_Ratio: THe ratio of non-white students to the total population.
6. West, Midwest, East: Region dummy variables. We divided all schools into 4 regions and if an institution belongs to a certain region, the corresponding dummy varible will be 1, and others be 0. We drop one dummy variable Northeast in the multiple linear regression to avoid perfect collinearity.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7646.3936	2374.4776	3.22	0.0013
MD_EARN_WNE_P10	0.4786	0.0406	11.78	0.0000
PCTFLOAN	-28506.4649	2208.7849	-12.91	0.0000
C100_4	20369.2500	2471.9502	8.24	0.0000
COSTT4_A	-0.4679	0.0383	-12.21	0.0000
MAJOR_CITY	2874.5782	739.9924	3.88	0.0001
MINORATIO	14298.1355	1799.6127	7.95	0.0000
WEST	998.5487	1143.0763	0.87	0.3825
MIDWEST	451.4619	964.1410	0.47	0.6397
NORTHEAST	-994.0266	973.1039	-1.02	0.3072

Table 3: OLS Regression Output for the Full Data Set

We noticed that the coefficients are large and each variable comes with different units. In order to make the regression result more intepretable, we take logs of all the quantitative data in the regression. Because we

can take log of 0, we replace 0 with 0.0001 in our data set.

In addition, since our data set is large enough, we will first split the set into train set and test set in order to gauge our the performance of our estimated coefficients. We divide the data set according to 3:1 train and test ratio. The train set is used to build the model and test set calculate the SE.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-8.9840	1.3240	-6.79	0.0000
ln_MD_EARN_WNE_P10	2.9014	0.1211	23.96	0.0000
ln_PCTFLOAN	-0.3611	0.0372	-9.71	0.0000
ln_C100_4	0.2781	0.0198	14.03	0.0000
ln_COSTT4_A	-1.3325	0.0704	-18.92	0.0000
MAJOR_CITY	0.1624	0.0541	3.00	0.0027
MINORATIO	0.9794	0.1288	7.61	0.0000
WEST	-0.1255	0.0834	-1.50	0.1326
MIDWEST	-0.0500	0.0703	-0.71	0.4775
NORTHEAST	-0.1012	0.0710	-1.43	0.1540

Table 4: OLS Regression Output After Taking Log

Based on the regression output, some findings match with our expectations:

1. For every 1% increase in median earning 10 years after graduation, we expect the number of students applied to increase by 2.9%, holding other varaibles constant.
2. Every 1% increase in percentage of students with student loans is associated with a 0.36% decrease in number of students applied, holding other variables constant.
3. If the 4-year completion rate goes up by 1%, the number of students applied is predicted to increase by 0.28%, holding other varaibles constant.
4. If cost of attendence increase by 1%, on average, we expect the number of students applied decrease by 1.33%, holding other variables constant.
5. If an institution is located near a major city, the number of students applied will be 0.16% higher than schools in countryside, holding other variables constant.
6. If the minority ratio increases by 1%, we predict the number of students applied will increase by 0.98%.

Those 6 coefficients are all very significant at 1% significance level with p-value close to 1.

Since OLS is the most common and versatile method, it is our first choice. However, in order to decide which method fits our data better, we also apply Ridge regression (RR), Lasso regression (LR), Principal Components regression (PCR) and Partial Least Squares regression (PLSR) method.

In order to improve our accuracy on predicting MSE, we use cross validation. We used sample() function to get a 3:1 train test split of our original data and for reproducibility purpose, we set.seed() before running the simulation.

For ridge and lasso regression method, we used cv.glmnet() in R package "glmnet" to conduct the ten-fold cross-validation on the train set. We then used the best fitted lambda we found from the train set to build a model and calculate MSE from the test set in order to gauge our performance.

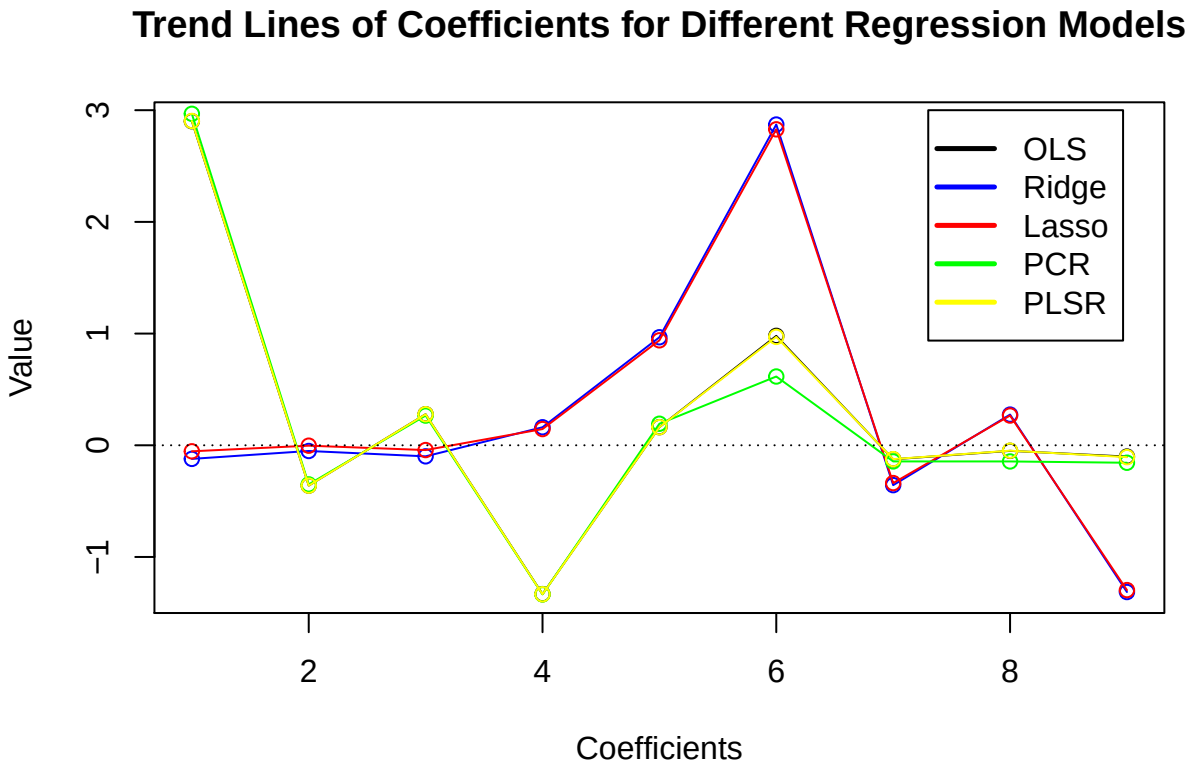
Similarly, for pcr and plsr regression method, we used function pcr() and plsr() in "pls" package to perform the 10-fold cross-validation. We also use the best fitted m from the train set to build the model and obtain the MSE from the test set.

	ols	ridge	lasso	pcr	plsr
ln_MD_EARN_WNE_P10	2.90140	-0.12192	-0.05400	2.96763	2.90339
ln_PCTFLOAN	-0.36113	-0.05108	-0.00359	-0.35031	-0.36169
ln_C100_4	0.27808	-0.09914	-0.04290	0.26582	0.27818
ln_COSTT4_A	-1.33248	0.16132	0.14510	-1.33504	-1.33405
MAJOR_CITY	0.16237	0.96714	0.93968	0.19294	0.16156
MINORATIO	0.97941	2.87174	2.82871	0.61573	0.97012
WEST	-0.12554	-0.35703	-0.33850	-0.14452	-0.12407
MIDWEST	-0.04997	0.27480	0.26615	-0.14481	-0.04819
NORTHEAST	-0.10121	-1.31394	-1.29657	-0.15651	-0.10642

Table 5: Regression Coefficients for 5 Regression Methods

	MSE
ols	0.97751
ridge	0.97494
lasso	0.96921
pcr	0.97788
plsr	0.97624

Table 6: MSE of 5 Regression Methods



We notice that all regerssion methods have similar MSE with lasso resulting the smallest. However, the coefficients are rather different between OLS, pcr, plsr and ridge, lasso. After intepretating the coefficients, we decide to use ols on the regression that we will run for each cluster for two main reasons:

1. lm() provides us with a p-value associated with each coefficient. This gives us information regarding to whether the impact of a certain varaible is significant which plays a vital role in determining our advice to schools in a certain cluster. In contrast, due to the way Ridge and Lasso regression are designed, although they give out better prediction, but the SE associated with each coefficient is unreliable, so we will lose this information if we go with those regression methods.

2. We notice that some of our regression variables are correlated. Therefore, ridge and lasso regression, aiming to reduce the dimension by eliminating unnecessary variables, can cause a problem. If both varaibles can affect the school competitiveness through the same channel, we don't want the regression randomly drop one since they explain the same portion of change in Y. Instead, we want to make the decision on our own based on the budget required for each change or whether it is practical to execute for a certain institution.

Therefore, weighing all the pros and cons for each regression method, we decide to use OLS to run the regression for each cluster.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-7.66	4.79	-1.60	0.11
ln_MD_EARN_WNE_P10	2.66	0.46	5.79	0.00
ln_PCTFLOAN	-2.16	0.39	-5.51	0.00
ln_C100_4	0.19	0.05	3.65	0.00
ln_COSTT4_A	-1.29	0.27	-4.78	0.00
MINORATIO	0.19	0.50	0.38	0.71

Table 7: Regression Coefficients WM Cluster

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-9.49	5.21	-1.82	0.07
ln_MD_EARN_WNE_P10	3.31	0.49	6.73	0.00
ln_PCTFLOAN	-0.16	0.12	-1.35	0.18
ln_C100_4	0.31	0.07	4.45	0.00
ln_COSTT4_A	-1.68	0.24	-6.95	0.00
MINORATIO	0.50	0.56	0.90	0.37

Table 8: Regression Coefficients WN Cluster

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-6.28	5.40	-1.16	0.25
ln_MD_EARN_WNE_P10	2.13	0.49	4.33	0.00
ln_PCTFLOAN	-1.51	0.36	-4.19	0.00
ln_C100_4	0.15	0.06	2.52	0.01
ln_COSTT4_A	-0.89	0.31	-2.85	0.00
MINORATIO	1.82	0.58	3.12	0.00

Table 9: Regression Coefficients MM Cluster

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-12.51	3.39	-3.69	0.00
ln_MD_EARN_WNE_P10	3.51	0.32	11.08	0.00
ln_PCTFLOAN	-0.35	0.08	-4.19	0.00
ln_C100_4	0.27	0.05	5.84	0.00
ln_COSTT4_A	-1.62	0.16	-10.31	0.00
MINORATIO	1.00	0.39	2.56	0.01

Table 10: Regression Coefficients MN Cluster

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-10.51	3.46	-3.04	0.00
ln_MD_EARN_WNE_P10	2.14	0.29	7.40	0.00
ln_PCTFLOAN	-1.00	0.19	-5.23	0.00
ln_C100_4	0.06	0.07	0.87	0.39
ln_COSTT4_A	-0.46	0.21	-2.20	0.03
MINORATIO	1.11	0.45	2.44	0.02

Table 11: Regression Coefficients NM Cluster

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.17	2.89	-0.06	0.95
ln_MD_EARN_WNE_P10	1.79	0.24	7.45	0.00
ln_PCTFLOAN	-0.48	0.07	-6.78	0.00
ln_C100_4	0.67	0.09	7.40	0.00
ln_COSTT4_A	-1.05	0.14	-7.39	0.00
MINORATIO	1.80	0.27	6.72	0.00

Table 12: Regression Coefficients NN Cluster

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.71	4.56	-0.16	0.88
ln_MD_EARN_WNE_P10	2.11	0.46	4.61	0.00
ln_PCTFLOAN	-1.62	0.29	-5.59	0.00
ln_C100_4	0.21	0.05	4.03	0.00
ln_COSTT4_A	-1.38	0.22	-6.34	0.00
MINORATIO	0.74	0.31	2.39	0.02

Table 13: Regression Coefficients SM Cluster

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.60	3.35	-0.48	0.63
ln_MD_EARN_WNE_P10	2.45	0.36	6.88	0.00
ln_PCTFLOAN	-0.85	0.24	-3.53	0.00
ln_C100_4	0.26	0.05	5.12	0.00
ln_COSTT4_A	-1.61	0.17	-9.27	0.00
MINORATIO	0.83	0.27	3.05	0.00

Table 14: Regression Coefficients SN Cluster