Analysis

```r
library(xtable)
library(Matrix)
load("../../data/data-outputs/anova.RData")
```

After dividing our data set into 8 separate clusters, our first step is to perform an analysis of variance. 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 1: 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, in order to develop strategies to improve competitiveness for each cluster, we first tabulate some key statistics for each cluster.

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

$Studentsapplied = Median_Earning + Completion_rate + Percentage_with_Student_Loans + Major_City + Minority_Ratio + West + Midwest + Northeast$

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.

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.

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

	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 2: OLS Regression Output for the Full Data Set

	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 3: OLS Regression Output After Taking Log

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 and lasso regression method.

For Ridge and Lasso, because of the way the regression is set up, SE are not reliable measures. But comparing all the coefficients, we can see that

	ols	ridge	lasso
(Intercept)	-8.98	-8.86	-8.60
ln__MD__EARN__WNE__P10	2.90	-0.12	-0.05
ln__PCTFLOAN	-0.36	-0.05	-0.00
ln__C100__4	0.28	-0.10	-0.04
ln__COSTT4__A	-1.33	0.16	0.15
MAJOR__CITY	0.16	0.97	0.94
MINORATIO	0.98	2.87	2.83
WEST	-0.13	-0.36	-0.34
MIDWEST	-0.05	0.27	0.27
NORTHEAST	-0.10	-1.31	-1.30

Table 4: Regression Coefficients for 3 Regression Methods