
University Characterization

Undergraduate Graduation Rates

Peter Glantzis

Kevin Jones

Edward Mengel

Purpose

- To determine how various factors explain undergraduate graduation rates for universities within the US → Profiling
 - How may a university influence the competitive standing of their overall program with respect to average graduation rates
 - Situation: Assumed that we were consulting to a university, and attempting to help them determine what actions to take in order to raise their graduation rate to an above average level with respect to all US undergraduate programs



Initial Dataset

- Obtained dataset (circa 1996) from US News and World Report regarding undergraduate programs within the United States
- Initial dataset included 1,302 records and contained 17 distinct variables

College Name	State	Type	Math SAT	Verbal SAT	ACT	# Apps	# Accept	# Enroll	% Top 10	% Top 25	# Full Time	# Part Time	State Tuition	Tuition	% PHD	Faculty Ratio	Grad Rate
University of Illinois	IL	Public	617	522	27	14939	11652	5705	52	88	25422	911	2760	7560	87	17.4	81
Indiana University	IN	Public	530	466	24	16587	13243	5873	25	72	24763	2717	2984	9766	77	21.3	68
University of Iowa	IA	Public				9224	8025	3262	23	52	15412	2878	2291	8149		13.1	62
University of Michigan	MI	Public	634	543	27	19152	12940	4897	66	92	22045	1339	5040	15732	90	11.5	87
Michigan State University	MI	Public	524	461	23	18117	15777	5180	23	57	26640	4120	4103	10658	93	14	71
University of Minnesota	MN	Public	568	484	23	11057	6397	3524	26	55	16502	21836	3171	8949	88	12.2	45
Northwestern University	IL	Private	670	600	29	12289	5200	1902	85	98	7450	45	16404	16404	96	6.8	92
Ohio State University	OH	Public				15076	12860	5330	24	52	31039	6005	3087	9315		13.1	56
Penn State University	PA	Public	583	500		19315	10344	3450	48	93	28938	2025	4966	10645	77	18.1	63
Purdue University	IN	Public	545	453	24	21804	18744	5874	29	60	26213	4065	2884	9556	86	18.2	67
University of Wisconsin	WI	Public				14901	10932	4631	36	80	23945	2200	2737	9096	93	11.5	72

Dataset Manipulation

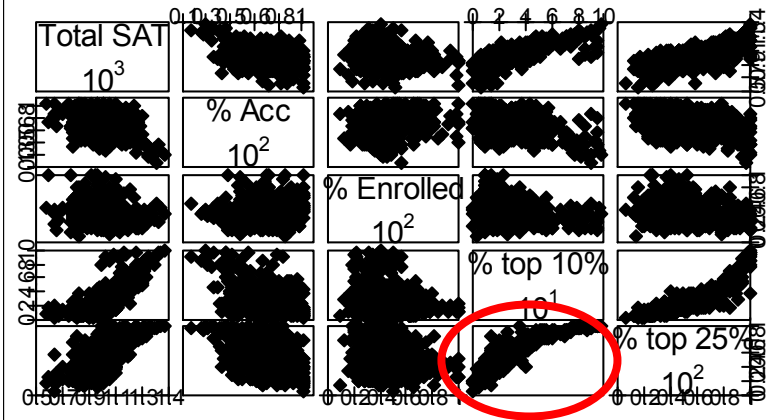
- Cleansed data (1,302 \longrightarrow 766 records)
 - Removed 298 records due to missing ACT and SAT data
 - Removed 8 records due to missing acceptance data
 - Removed 164 records due to missing top 10 or 25% of class data
 - Removed 10 records due to missing tuition data
 - Removed 15 records due to missing PhD data
 - Removed 39 records due to missing graduation rate data
 - Removed 1 record due to data entry error (%enrolled = 244%)
 - Removed 1 record due to outlier (std/fac ratio = 91.8, avg = 15, std = 5)
 - Created other variables
 - Created categorical response variable for graduation rate (above/below average – 62%)
 - Created total SAT score
 - Summed Math and Verbal scores
 - Converted ACT scores (“saved” 157 records)
 - Transformed absolute variables
 - Percent of applications accepted by the university
 - Percent of accepted applications that resulted in enrolled students
 - Percent of full time students
 - Removed Variables
 - Removed variables deemed uncontrollable by the university
 - Private vs. Public
 - State
- 41% of total dataset

Data Summary

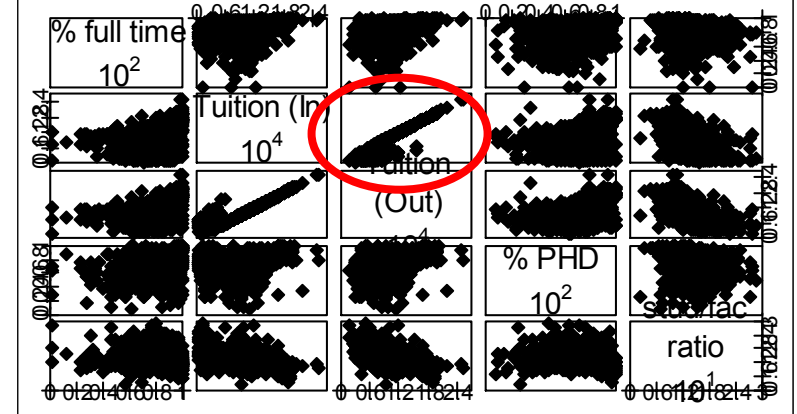
Category	Minimum	Maximum	Mean	Median	Standard Deviation
SAT	600	1,410	989	990	116
Accept	15	100	75	78	15
Enroll	10	244	43	40	16
Top 10%	1	98	25	21	18
Top 25%	6	100	53	51	20
Full Time	1	100	80	84	17
In Tuition	480	25,180	8,476	8,846	5,238
Out Tuition	1,672	25,180	9,786	9,213	4,011
PHD	8	100	71	74	17
Faculty Ratio	3	92	15	14	5
Graduation Rate	10	118	62	62	18

Exploratory Plots

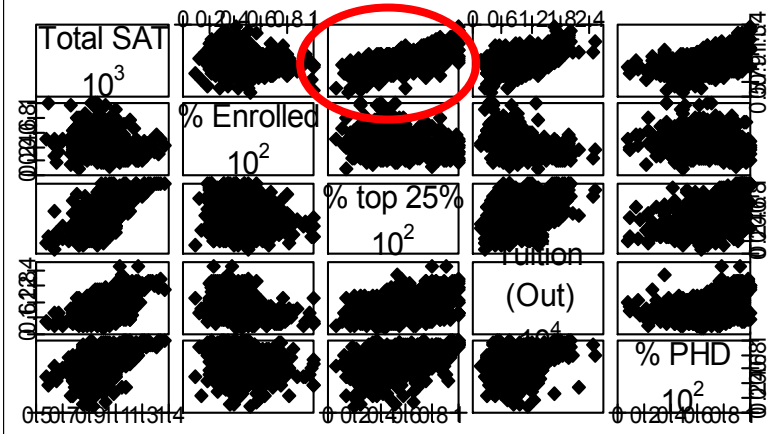
Matrix Plot 1



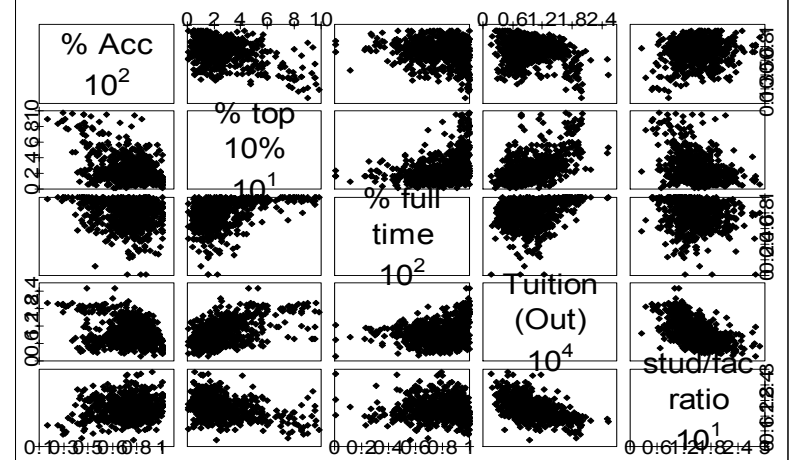
Matrix Plot 2



Matrix Plot 3



Matrix Plot 4



Correlation Table

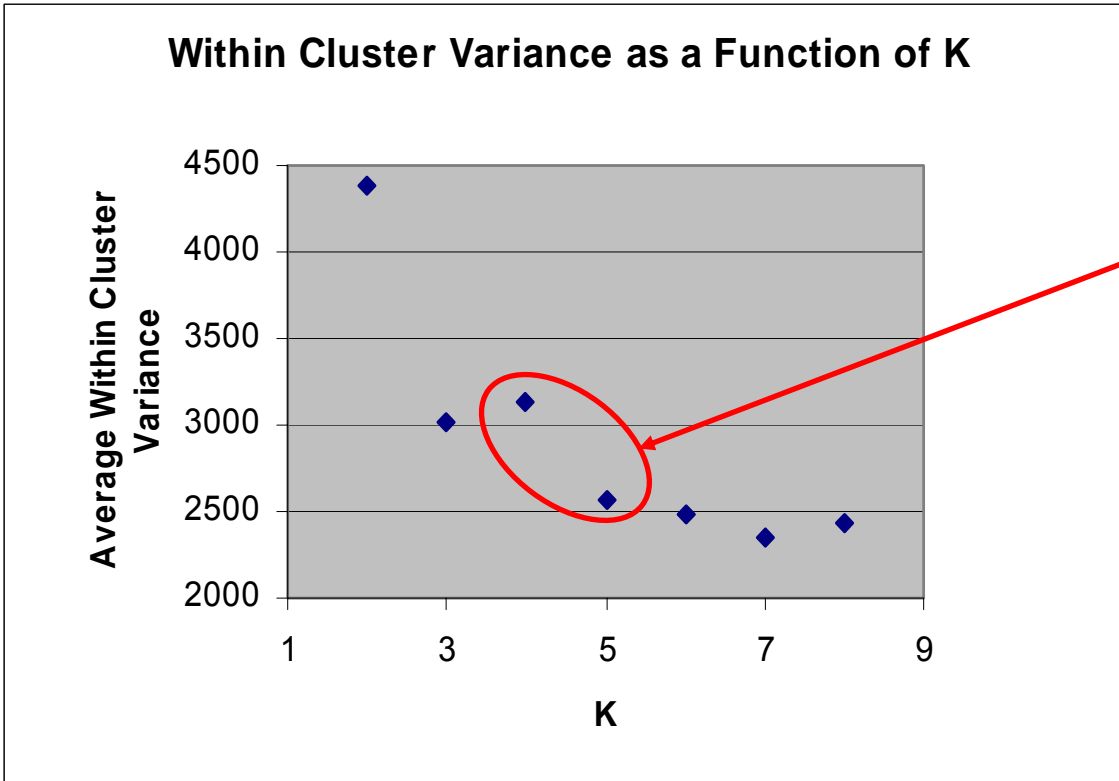
- Choose to develop a logistic regression model
 - Removed the *Top 25%* and *In-State Tuition* variables
 - As to be expected, high correlation between:
 - In-state Tuition and Out-of-state Tuition
 - Percent of students in top 10% of high school class and percent in top 25%
 - High possibility of confounding factor with in-state tuition (public vs private given that private institutions do not have in-state tuition so larger discrepancy)

	SAT	Accept	Enroll	Top 10%	Top 25%	Full Time	In Tuition	Out Tuition	PHD	Faculty Ratio
SAT	1.000									
Accept	-0.361	1.000								
Enroll	-0.164	0.056	1.000							
Top 10%	0.752	-0.428	-0.194	1.000						
Top 25%	0.741	-0.413	-0.212	0.889	1.000					
Full Time	0.414	-0.139	-0.126	0.374	0.348	1.000				
In Tuition	0.443	-0.115	-0.369	0.456	0.386	0.219	1.000			
Out Tuition	0.555	-0.218	-0.407	0.539	0.484	0.269	0.927	1.000		
PHD	0.519	-0.307	-0.211	0.516	0.522	0.269	0.159	0.340	1.000	
Faculty Ratio	-0.263	0.091	0.161	-0.289	-0.234	-0.136	-0.456	-0.442	-0.114	1.000

Final List of Variables

- Predictors:
 - Total SAT Score
 - Percent of Applications Accepted
 - Percent of Accepted Applications Enrolled
 - Percent of Enrolled Students In Top 10% of High School Class
 - Percent of Enrolled Students Attending Full-Time
 - Out-of-State Tuition
 - Percent of Faculty with PHD
 - Student to Faculty Ratio
- Response:
 - Graduation Rate (dummy – above/below average)

K-Cluster Analysis: Selection of K



After K=4 the graph “flattens out”, so we selected K to be 4

K	Avg Variance Within Cluster
2	4377.481234
3	3019.39247
4	3134.114229
5	2571.277858
6	2486.068013
7	2347.193057
8	2439.606356

K-Cluster Analysis: Results

Sample Sets From Each Cluster

1
Adelphi University
Briar Cliff College
Central Connecticut State University
Embry Riddle Aeronautical University
Fairleigh Dickinson University
Northeastern Illinois University
Northeastern Illinois University
Sacred Heart University
Trinity College
Webster University

Count: 224

2
Beaver College
Clarkson University
DePaul University
Fordham University
Hofstra University
Loyola University
Marist College
Pace University
Roanoke College
Villanova University

Count: 240

3
Ball State University
Clemson University
Colorado State University
CUNY - City College
Hampton University
Mississippi State University
Ohio University
Texas Tech University
University of Nebraska at Lincoln
Virginia Military Institute

Count: 229

4
Boston University
College of William and Mary
Duke University
Hamilton College
Massachusetts Institute of Technology
Pomona College
Princeton University
Sarah Lawrence College
University of California at Berkeley
University of North Carolina at Chapel Hill

Count: 73

K-Cluster Analysis: Results (cont)

Intuition: 1,3,2,4, Lowest → Highest

Cluster	Total SAT	% top 10%	% top 25%	% full time
Cluster-1	890.776348	13.86403	36.1886	67.111713
Cluster-2	1030.50417	29.14583	59.950006	85.739966
Cluster-3	974.415998	19.867258	49.535401	80.365807
Cluster-4	1196.513503	66.702681	90.418985	95.605413
Low est	1	1	1	1
	3	3	3	3
	2	2	2	2
Highest	4	4	4	4

Cluster	Tuition (In)	Tuition (Out)	% PHD	Graduation rate
Cluster-1	7910.96585	8482.928306	54.70175	54.078947
Cluster-2	11934.43827	12241.01347	76.162505	72.129161
Cluster-3	3094.110038	6410.749312	74.008854	52.50443
Cluster-4	15440.26463	16151.10224	90.527013	83.783862
Low est	3	3	1	1
	1	1	3	3
	2	2	2	2
Highest	4	4	4	4

Intuition: N/A

Cluster	% Enrolled
Cluster-1	44.076842
Cluster-2	37.083777
Cluster-3	51.383859
Cluster-4	34.127988
Highest	1
	3
	2
Low est	4

Intuition: 1,3,2,4, Highest → Lowest

Cluster	% Acc	stud/fac ratio
Cluster-1	80.635907	14.104823
Cluster-2	79.061152	13.085418
Cluster-3	72.571482	17.972568
Cluster-4	53.964578	11.086488
Highest	1	1
	2	3
	3	2
Low est	4	4

K-Cluster: Conclusions

- ❑ Most predictors, including graduation rate (our response variable in other analyses) followed the intuitive 1,3,2,4 order
- ❑ Tuition (both in-state and out-of-state) was an exception: Cluster 1 schools were more expensive on average than cluster 3 schools
- ❑ Percent Accepted was also an exception: Cluster 2 schools accepted a slightly higher percentage of applicants on average than cluster 3 schools
- ❑ **General Conclusion:**
 - The average graduation rate for all schools in the dataset was 62%
 - In order to achieve that average graduation rate, schools should aim to be in either clusters 2 or 4
 - However, a school may be able to “slip” a little on percent accepted as indicated by cluster 2’s position in that category

Logistic Regression Results

- Used standardized data to eliminate scale problems
- Cutoff = 0.5

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	0.26339805	0.09540495	0.00576525	*
Total SAT	0.42079526	0.15494211	0.00661114	1.52317238
% Acc	-0.08005286	0.10814465	0.45915514	0.92306757
% Enrolled	-0.21686934	0.10450006	0.0379584	0.80503517
% top 10%	0.49277952	0.16952206	0.00365059	1.63685954
% full time	0.07812132	0.093146	0.40163901	1.08125389
Tuition (Out)	1.18198669	0.15759324	0	3.26084614
% PHD	-0.05172557	0.10567362	0.6244989	0.94958943
stud/fac ratio	0.09741648	0.1056126	0.35632285	1.10231936

Residual df	756
Residual Dev.	763.273682
% Success in training data	51.372549
# Iterations used	7
Multiple R-squared	0.27988878

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	289	104
0	80	292

Error Report			
Class	# Cases	# Errors	% Error
1	393	104	26.46
0	372	80	21.51
Overall	765	184	24.05

Evaluation of Goodness of Fit

Residual df	756
Residual Dev.	763.273682
% Success in training data	51.372549
# Iterations used	7
Multiple R-squared	0.27988878

Deviance of the Naive Model:

$$D_0 = D / (1 - R^2)$$

$$D_0 = 763.273682 / (1 - 0.27988878^2)$$

$$D_0 = 1049.837829$$

Check of Statistical Significance:

$$d = D_0 - D$$

$$d = 1049.837829 - 763.273682$$

$$d = 268.5641476$$

$k = 8$ (number of predictors in the model)

In Excel: `chidist(d,k)`

$$p\text{-value} = 2.97122 \times 10^{-57} \approx 0$$

Reduction in deviance is statistically significant

Model provides a good overall fit – better than the naïve model

Model Interpretation

$$P = \frac{1}{1 + e^{-(\alpha + \beta_1 \text{ SAT} + \beta_2 \% \text{Acc} + \beta_3 \% \text{Enr} + \beta_4 \% \text{Top10} + \beta_5 \% \text{Full} + \beta_6 \text{ Tuition(out)} + \beta_7 \% \text{PHD} + \beta_8 \text{ Std/fac} + \epsilon)}}$$

Predictor Name	Coefficient/ Constant	Raw Data	Normalized Data
Constant term	0.26339805	1	1
Total SAT	0.42079526	1,211	1.920069097
% Acc	-0.0800529	12	-4.276230779
% Enrolled	-0.2168693	75	1.999841136
% top 10%	0.49277952	59	1.880931893
% full time	0.07812132	100	1.200032243
Tuition (Out)	1.18198669	19,840	2.506215091
% PHD	-0.0517256	37	-1.930161581
stud/fac ratio	0.09741648	9	-1.192472112

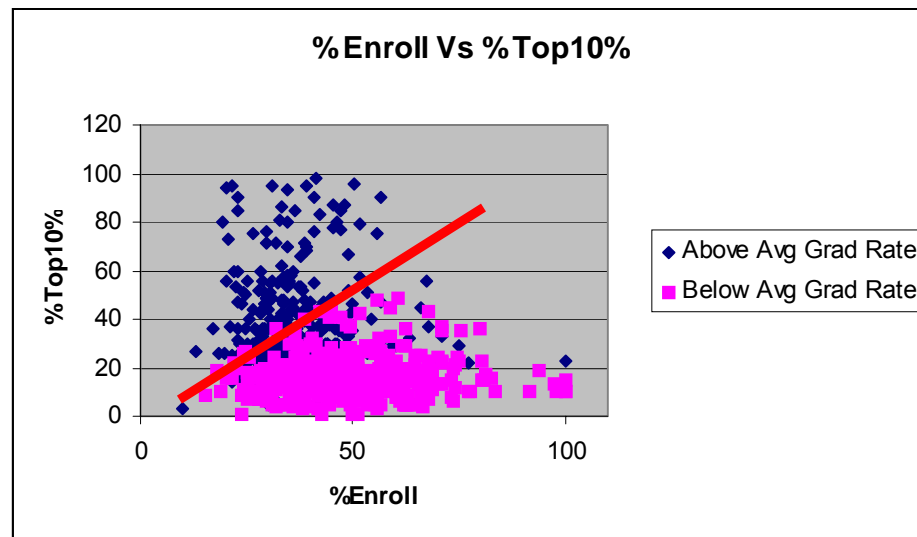
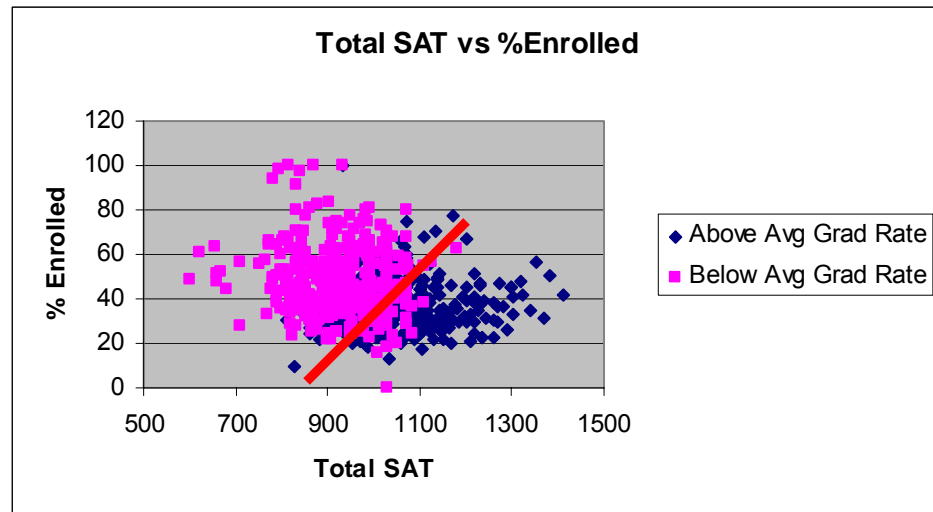
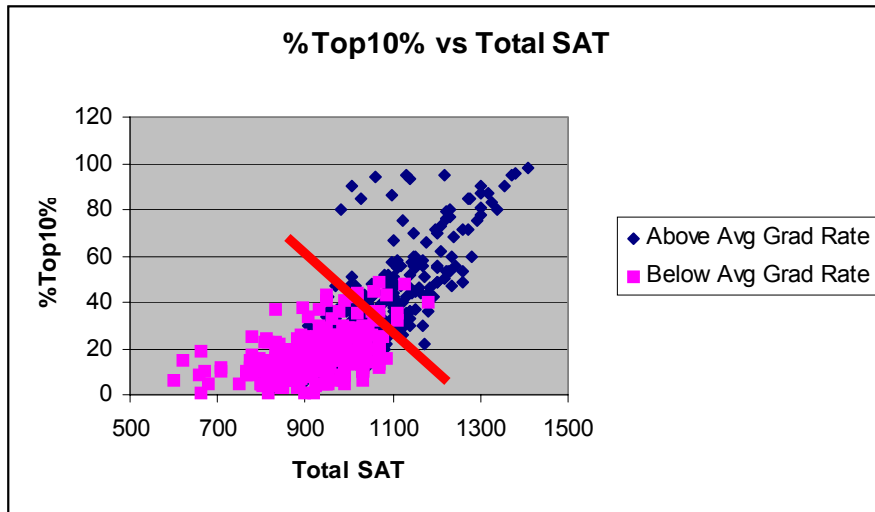
Probability Y = 1

0.9929426









Above Average Grad Rate = 1

Below Average Grad Rate = 0

Model Interpretation – Visual Analysis



Model Interpretation (continued)

Predictor	Effect On Probability Of Above Average Graduation Rate When Predictor Increases	Comments
Total SAT Score	Increase 	Expected
Percent of Applications Accepted	Decrease 	Expected, but not significant
Percent of Accepted Applications Enrolled	Decrease 	Somewhat expected based on K-cluster results
Percent of Enrolled Students In Top 10% of High School Class	Increase 	Expected
Percent of Enrolled Students Attending Full-Time	Increase 	Expected, but not statistically significant
Out-of-State Tuition	Increase 	Expected, but not statistically significant
Percent of Faculty with PHD	Decrease 	Unexpected, but not statistically significant
Student to Faculty Ratio	Increase 	Unexpected, but not statistically significant

Conclusions

- Increasing the quality of the student will increase the graduation rate
 - Increase the number of students from the top 10% of their high school class
 - Increase the total SAT score of the students accepted for admission
- The percentage of accepted applicants that enroll is an interesting predictor
 - Our results indicate that schools should attempt to decrease the percentage of accepted applicants that enroll
 - However, we do not recommend this approach