COLLEGIATE DATASET ANALYSIS

Matt McManus
May 12th, 2011
INFO 47470 – John Abowd & Lars Vilhuber

Cornell University

The Questions

1. Is it possible to accurately predict graduation rate by fitting a linear regression model?
2. Can schools be clustered together to see if and how they relate to one another?
3. Can regression models be fit to each cluster to spot these relationships?

Sources

- Data
  - Statlib datasets
  - http://lib.stat.cmu.edu/datasets/colleges/
  - usnews.data
- Statistical methods
  - “Data Mining for Business Intelligence”
    - Galit Shmueli, Nitin R. Patel, Peter C Bruce
  - Wikipedia
- Software
  - R
  - Freely availible
  - http://www.r-project.org/

Motivation – Regression Models

- Useful for Universities
  - Identify which factors positively or negatively are affecting graduation rates
- Prospective students
  - Students could use model to predict academic rigor of universities they are looking at
  - If financial predictors are included, students families would have a better idea if their money is going to good use

Method – Regression Models

- Linear Regression
  - Fit a predictive model to an observed data set of y and X values
  - If given a new set of X values, able to predict the y value based on model
  - $y = \beta 0 + \beta 1X1 + \beta 2x2 \ldots \beta iXi$

- Backwards Elimination Method
  - Create model with all possible predictor variables
  - Remove least statistically significant, re-create model, and repeat

Procedure – Regression Models

- Step 1 – Grad Rate distribution
  - Check histogram of graduation rates
  - If smooth – run model on graduation rate values
  - If not – run model of logarithmic values of graduation rates

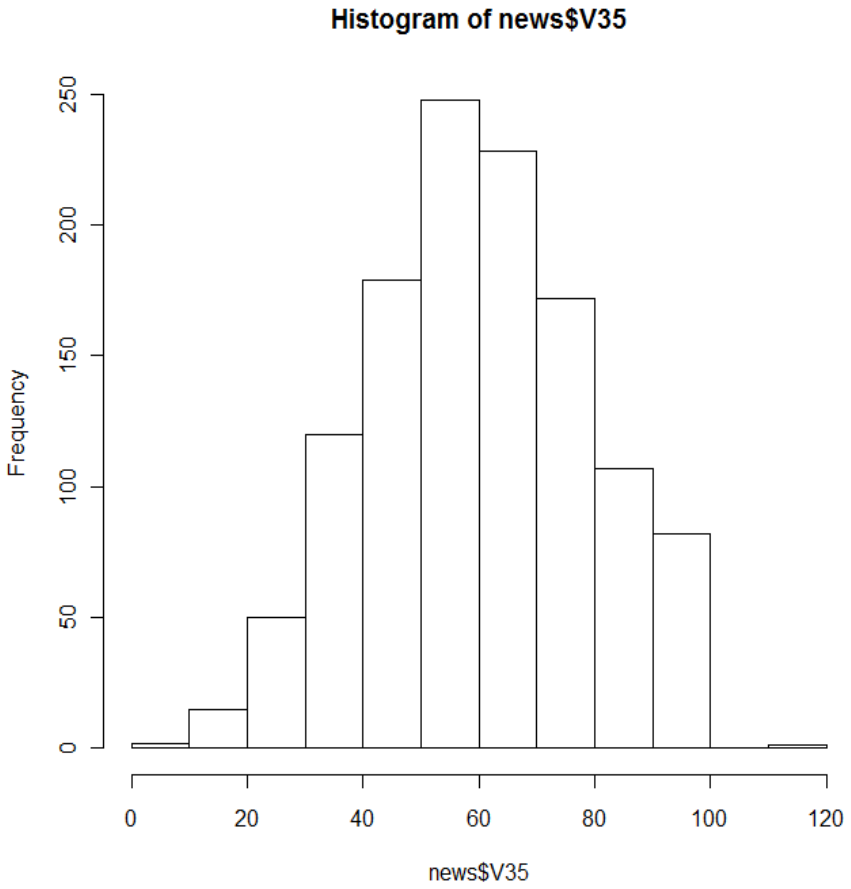
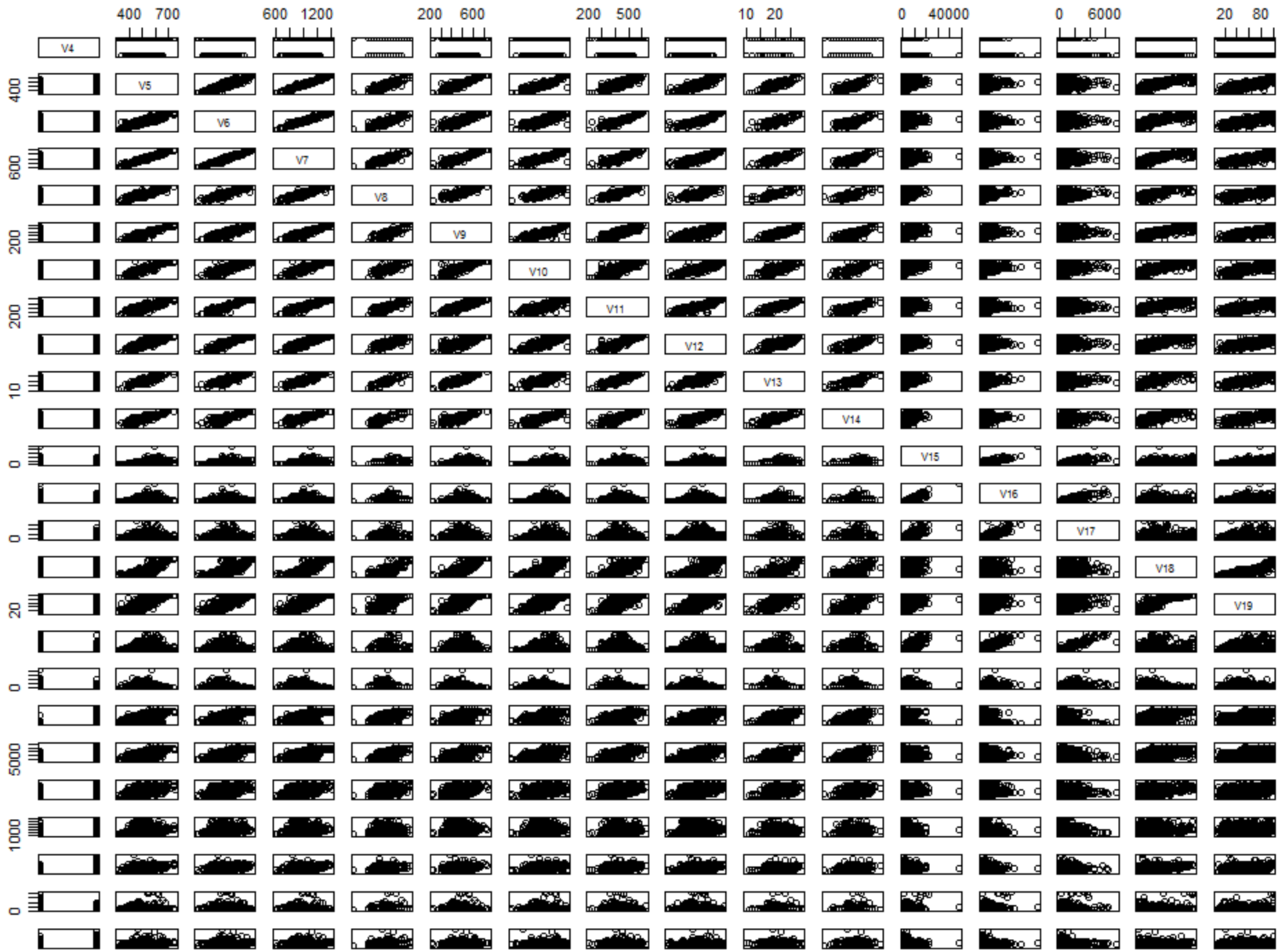


Figure 1: Grad Rate Histogram

Procedure – Regression Models

- Step 2 – Remove non-predictive variables
  - Remove data columns (variables) representing:
    - Federal ID number
    - College name
    - State
- Step 3 – Find correlated predictor variables
  - Create piecewise scatterplots to see if there are any significantly correlated predictors
  - No significantly correlated predictors were found, so first model was run on all variables
  - See Figure 2: Piecewise scatterplot


V4
V5
V6
V7
V8
V9
V10
V11
V12
V13
V14
V15
V16
V17
V18
V19
400 700
600 1200
200 600
200 500
10 20
0 40000
0 6000
20 80
400
600
200
200
10
0
0
20
0
5000
1000
0

Procedure – Regression Models

- Step 4 – Handle missing data
  - Some colleges had missing data for certain variables
    - Quartile test scores (SAT)
    - Average test scores (ACT)
  - Replaced missing values with the median values of the colleges that did have the variable
- Step 5 – Training/Testing sets
  - Randomly chose half of the data set to train the model
  - Remaining half is what training model was tested on
- Step 6 – Calculate error statistics/ choose model
  - Calculate RMSE and Cp statistic for regression model
  - Chose best fit

Error Statistics – Regression Models

- Root Mean Square Error (RMSE)
  - Measures variance between values predicted by model and actually observed variables
  - Calculation: Take the mean of the differences between actual and predicted values -> raise it to the ½ power
- Cp statistic
  - Stopping statistic for backwards elimination method
    - Helps pick best model with subset of predictors
      - Adresses overfitting
  - Cp = (n – Pfull – 1)(SSE/SSEfull) – n + 2(p+1)
    - N = # data items
    - Pfull = # predictors full model, p = # predictors of submodel
    - SSEfull = sum of squares error full model, SSE = sum of squares error submodel
      - SSE = SUM(predicted y – mean y)^2)
      - Sum of the squares of the deviations of the predicted values from the mean value

Results – Regression Models

- Summary
  - Performed Backwards Elimination on 3 different training models:
    - All predictor variables
      - Cp = 7.6
      - RMSE = 11.6
    - Only academic variables
      - Cp = 124.33
      - RMSE = 16.2
    - Only financial variables
      - Cp = 231.74
      - RMSE = 18.7
  - Model with all variables yielded best results
  - Also fitted a model on entire data set for use on future releases
    - Cp = 14.26
    - RMSE = unable to test until similar future dataset's are released

Results – Regression Models

- Model for all predictor variables

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	20.8592427	2.7925064	7.470	2.90e-13	***
`Public or Private`	6.3899627	1.6479664	3.877	0.000117	***
`No Of Apps Recieved`	0.0010084	0.0001738	5.801	1.07e-08	***
`Pct new students from top 25% of H.S Class`	0.1574173	0.0361503	4.355	1.57e-05	***
`No of parrtime undergraduates`	-0.0013729	0.0003580	-3.835	0.000139	***
`Out-of-state tuition`	0.0013175	0.0002057	6.405	3.07e-10	***
`Pct. alumni who donate`	0.3127244	0.0593144	5.272	1.89e-07	***

Results – Regression Models

- Model for only academic variables

```
Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                -2.280e+01  7.390e+00  -3.086 0.002079 **
`Public or Private`                         1.269e+01  1.106e+00  11.471  < 2e-16 ***
`Average Verbal SAT`                        5.540e-02  1.358e-02   4.079 4.82e-05 ***
`First Quartile Verbal SAT`                 1.125e-01  2.092e-02   5.379 8.99e-08 ***
`Third Quartile Verbal Sat`                -7.772e-02  2.026e-02  -3.837 0.000131 ***
`Third Quartile  ACT`                       8.640e-01  2.463e-01   3.508 0.000469 ***
`No Of Apps Recieved`                       1.013e-03  1.418e-04   7.143 1.59e-12 ***
`Pct new students from top 25% of H.S Class` 9.381e-02  3.149e-02   2.979 0.002950 **
`No of parrtime undergraduates`            -1.799e-03  2.913e-04  -6.178 8.90e-10 ***
`Pct. Faculty with Ph.D`                    9.348e-02  2.909e-02   3.213 0.001347 **
`Student faculty ratio`                    -1.904e-01  9.732e-02  -1.957 0.050622 .
---
```

Results – Regression Models

- Model for only financial variables

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.843e+01  1.820e+00  15.620  < 2e-16 ***
`In state tuition`       4.636e-04  2.098e-04   2.210 0.027292 *
`Out of State tuition`   1.541e-03  2.805e-04   5.493 4.81e-08 ***
`Room and Board costs`   1.401e-03  5.090e-04   2.752 0.006016 **
`Additional fees`        3.872e-03  1.006e-03   3.851 0.000124 ***
`Pct. Alumni who Donate` 3.054e-01  4.296e-02   7.110 1.99e-12 ***
---
```

Results – Regression Models

- Model on all of the data

```
Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                 -7.7611472  6.9270228  -1.120 0.262763
`Public or Private`                          5.9533604  1.2110201   4.916 1.01e-06 ***
`Average Verbal SAT`                         0.0449407  0.0130300   3.449 0.000582 ***
`3rd Quartile Math SAT`                      0.0358600  0.0162293   2.210 0.027324 *
`First Quartile Verbal SAT`                  0.0681859  0.0208932   3.264 0.001132 **
`Third Quartile Verbal Sat`                 -0.0853839  0.0229571  -3.719 0.000209 ***
`First Quartile  ACT`                        0.5424137  0.2332125   2.326 0.020195 *
`No Of Apps Recieved`                        0.0008919  0.0001444   6.176 9.02e-10 ***
`Pct new students from top 25% of H.S Class` 0.0767564  0.0305307   2.514 0.012066 *
`No of parrtime undergraduates`             -0.0018696  0.0002785  -6.713 2.94e-11 ***
`Out-of-state tuition`                       0.0010923  0.0001788   6.109 1.36e-09 ***
`Room and board costs`                       0.0016427  0.0004809   3.416 0.000657 ***
`Pct alumni who donate`                      0.1874660  0.0419070   4.473 8.44e-06 ***
`Instructional expenditure per student`     -0.0002778  0.0001033  -2.689 0.007270 **
---
```

Results – Regression Models

- Explanation of statistics
  - t value
    - Ratios of βk to the standard error of βk for k = 0,1….
  - Pr(>|t|)
    - p-values for the t-statistic
    - Probability of seeing a t-statistic at least this far from zero if βk is actually = 0
- Explanation of model choice
  - Only chose predictors that were significant at the .05 level
    - p values <= .05
  - Chose model that was started with all variables
    - Lowest Cp statistic
    - Cp statistic similar to predictor #, safe choice – not a lot of bias

Motivation – Clustering

- Useful for Universities
  - Identify what other colleges they are competing with
  - Identify areas of improvement
- Prospective students
  - Students going through the application process can use the information to pool schools based on different factors
    - Likelihood of being accepted
    - Affordability

Method – Clustering

- K-means algorithm
  - Partitions n observations into k clusters
  - Steps:
    - Randomly assign cluster centroids (means)
    - Associate each observation with the nearest initial mean
    - Make the centroid of each cluster the new mean
    - Re-assign and continue process until assignments don’t change anymore
  - Choose k clusters to minimize the within-cluster sum of squares
    - W(c) value

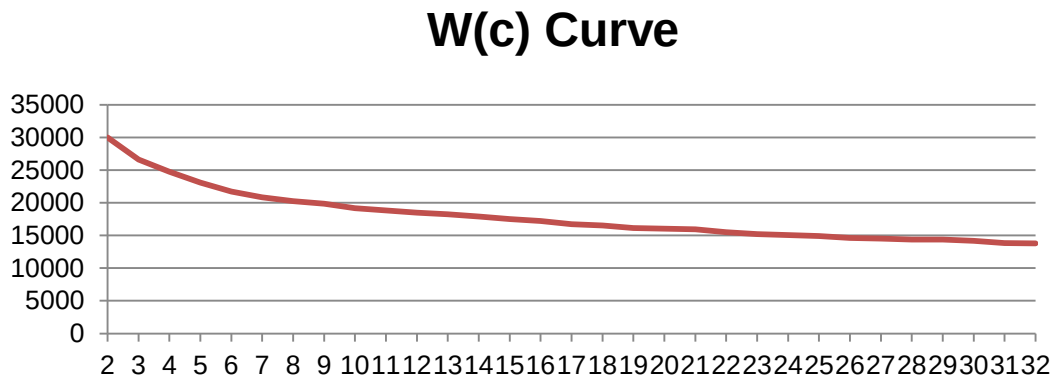
Procedure – Clustering

- Initial Steps
  - The initial steps were the same as in the regression method
  - Missing data was handled the same way

- Step 1 – Standardize data
  - Data must be standardized to prevent one measurement being dominant
  - Changes the definition of the distance measurement
  - Calculation used:
    - (Value – Mean)/ Standard deviation

Procedure – Clustering

- Step 2 – Plot k vs. W(c)
  - Determine optimal number of clusters k
  - Missing data was handled the same way



Results – Clustering

- 28 clusters
  - W(c) value of 14353.69
  - See appendix A for cluster assignments
    - Average of 47 colleges assigned to each of the 28 clusters
    - Hard to spot all trends due to large number of data
    - Some trends are obvious
      - Ex: Cluster 25 – all art schools
      - Ex: Cluster 26 – Harvard, Yale, Georgetown, Cornell, Air Force
  - Fitted backwards elimination regression model to each of the 28 clusters
    - Not every cluster yielded a good model, but some did
    - See appendix B for list of significant predictors per cluster

Results – Clustering

- Variable mean analysis
  - Analyzed the variable means of each cluster relative to the entire dataset
  - Gave insight into how schools are divided and what factors play a significant role in each cluster
  - See Appendix C for results

Conclusion

- Regression Models
  - Relatively accurate in presence of missing data
  - More useful for future releases of dataset
- Clustering
  - Very accurate in grouping colleges
  - Prestigious schools, art schools, technical schools, rural colleges, etc…all clustered together
  - Regression models on certain clusters were more accurate than the model on all observations
  - Variable means analysis was most use for interpreting the significance of each cluster

Appendix A

- clusters.csv

Appendix B – Sig Preds. Per Cluster

Cluster 1
Estimated Book Cost, Pct faculty w/ terminal degree
Cluster 2
Number parttime undergrads, Instructional expenditure per student
Cluster 5
Third Quartile Math SAT, In-state tuition, Room costs
Cluster 6
First quartile ACT, Third quartile ACT, In-state tuition, out-of-state tuition, room costs, addtil fees, pct faculty w/ phD, pct faculty w/ terminal degree, pct alumni who donate
Cluster 8
Avg Math SAT, Avg Combined SAT, Avg ACT, 1st quartile Verbal SAT, first quartile ACT, # applications received, # applicants accepted, # undergrad, in state tuition, board costs, pct faculty w/ terminal degree, instructional expenditure per student

Appendix B - Sig Preds. Per Cluster

Cluster 11
Estimated personal spending, pct alumni who donate
Cluster 13
Avg verbal SAT, avg combined SAT, 3rd quartile Math SAT, 1st quartile ACT, board costs, estimated personal spending, student/faculty ratio
Cluster 15
applications, 3rd quartile ACT, # applications accepted, # new students enrolled, # undergraduates, out-of-state tuition, room costs, estimated personal spending
Cluster 17
1st quartile verbal sat, # applications received & accepted, book cost, personal spending, % faculty with phD or terminal degree, student faculty ratio, % alumni who donate
Cluster 20
Addtl fees
Cluster 21
Instructional expenditure per student

Appendix B - Sig Preds. Per Cluster

Cluster 22

3rd quartile act, % students from top 25% of class, # part time undergrads, in state tuition, room and board, book cost, personal spending

Cluster 23

Act score, in state tuition, out of state tuition, board cost, personal spending, % faculty w/ degree

Cluster 24

Avg math sat, avg verbal sat, avg combined sat, avg act, 3rd quartile verbal sat, # applications accepted, # new students enrolled, in state tuition, estimated book costs, instructional expenditure per student

Cluster 28

Avg verbal sat, 1st quartile act, pct students from 10% of clas, pct faculty with phd, student faculty ratio

Appendix C – Variable Mean Results

Cluster 1 Key Points:

SAT Scores 30-50 points lower than average
1700 fewer applications than average
1000 fewer applicants accepted than average
500 fewer students enrolled
7% fewer students from top 25% of class
2500 fewer undergraduates
3600 higher in-state tuition
2230 higher out-of-state tuition
1000 higher room and board costs
2700 higher room costs
3% fewer faculty with phd
7.2% higher grad rate

Appendix C – Variable Mean Results

Cluster 2:
1000 more applications received & accepted
800 more new students enrolled
5% fewer from 25 & 10% of HS
4300 more undergrads
5700 lower in state tuition
3800 lower out of state tuition
1200 lower room and board
5% higher student/faculty ratio
6% less alumni who donate
3000 lower instructional expenditure per student
13% lower graduation rate

Appendix C – Variable Mean Results

Cluster 3:
50-80 points lower on SAT
1700 fewer applications
1000 fewer apps accepted
300 fewer students enrolled
8% fewer from top 25% of class
900 fewer undergrads
Higher parttime undergrads
Lower instate and out of state tuitions
20% fewer faculty w/ phd
35% higher student faculty ratio
4200 lower instructional expenditure per student
4.5 lower grad rate

Cluster 4
7000 higher undergrad
2400 higher apps
5000 lower tuition
9% more phd
6% more who donate
18% lower grad rate

Appendix C – Variable Mean Results

Cluster 5:
1000 more apps
2000 more undergrad
5000 lower tuition
11% more phd
7% less donate
1800 less instruct expend
8% lower grad rate

Cluster 6:
Higher standardized test scores 50-80 points
800 fewer apps received
550 fewer apps accepted
300 fewer students enrolled
20% from top 10% of schools
1800 fewer undergrads
4000 higher in state tuition
3500 higher out of state tuition
15% more faculty w/ phd
12% higher alumni who donate
1800 higher instructional expenditure
14% higher grad rate

Appendix C – Variable Mean Results

Cluster 7:
Lower than average in all categories
100 points lower on math sat
1200 fewer applicants
700 fewer accepted
1700 fewer undergrads
3500 lower tuition
700 lower room and board
16% fewer faculty w/ degree
3000 lower instructional expenditure per student
12% lower grad rate

Cluster 8:
Slightly higher test scores
7000 more applications
5000 more accepted
2000 more students enrolled
10% more from top 25%
11500 more undergrads
5000 lower in state tuition
1500 lower out of state
16% more faculty w/ phD

Appendix C – Variable Mean Results

Cluster 9:
Much higher test scores
1600 higher applications
300 fewer accepted
50% higher from top 10%
1200 fewer undergrads
10000 higher tuition
2000 higher room and board
23% higher faculty w/ degree
5% lower student faculty ratio
22% higher alumni who donate
15000 higher instructional expenditure per student
32% higher grad rate

Cluster 10:
Avg test scores
27% from top 10%
1700 fewer undergrads
9000 higher tuition
18% faculty w/ phd
-4% student faculty
20% donate
6000 higher instructional expenditure
23% grad rate

Appendix C – Variable Mean Results

Cluster 11
Lower test scores
1400 fewer applications
12% fewer in top 25%
2500 fewer undergrads
3880 higher tuition
5% higher grad rate

Cluster 12
Much lower test scores
2000 fewer apps received
20% fewer in top 25%
2500 fewer undergrads
2500 lower tuition
15% fewer faculty w/ degrees
2000 fewer instructional expenditure per student
15% lower grad rate

Cluster 13
Lower test scores
1500 fewer apps
11% fewer from top 25%
1500 fewer undergrads
3000 lower in-state tuition
9% lower grad rate

Appendix C – Variable Mean Results

Cluster 14
Much higher test scores
16000 higher apps
11000 higher accepted
4000 higher enrolled
28% more from top 10%
2000 more undergrads
3000 lower instate
1000 higher addtl fees
18% higher w/ phD's
3000 higher instruct expend
11% higher grad rate

Cluster 15
Lower test scores
Fewer applications
11% less from top 25%
2500 fewer undergrad
1000 lower out of state tuition
30% fewer w/ degree
5% lower grad rate

Clust 16
lower test scores
1200 more apps
1400 more undergrads
4000 less in state tuition
4% lower grad rate

Appendix C – Variable Mean Results

Cluster 17
Higher test scores
1300 more apps
11% from top 25%
5000 larger instate tuition
1600 more room and board
13 more phds
14 % higher grad rate

Cluster 18
lower tests
1800 fewer apps
2000 fewer undergrads
1300 lower out of state
2000 higher personal spending
17% fewer phds

Cluster 19
100-200 point higher test scores
5000 higher apps
46 from top 10%
9700 higher instate
1600 higher room and board
20% faculty w/ phds
6% lower student faculty ratio
18% alumni donate
15000 higher instruct expend per student
31% grad rate

Appendix C – Variable Mean Results

Cluster 20
Slightly higher test scores
1600 fewer apps
8% from top 25%
2300 fewer ndergrads
3000 higher tuition5% w/ phd
2% lower ratio
8% donate
9% higher grad rate

Cluster 21
Higher test scores
12% from top 10%
2200 fewer undergrad
9000 higher tuition
17% phd
4% lower student faculty ratio
14% donate
600 instruc expend per student
14% grad rate

Appendix C – Variable Mean Results

Cluster 22
1500 higher apps
12% from top 10%
5000 higher tuition
1000 higher room and board
12% fewer phds
4000 instruc expend per student
14% lower grad rate

Cluster 23
1500 more apps
1600 more undergrads
4000 lower tuition
16% lower phds
2% student faculty ratio
5% less donate
3200 lower instruct expend
9% lower grad rate

Cluster 24
Lower test scores
1300 fewer apps
20% less from top 25%
1700 fewer undergrads
2000 lower tuition
6% fewer phds
7% fewer donate
2000 lower instruct expend
11% lower grad rate

Appendix C – Variable Mean Results

Cluster 25
1300 fewer apps
6% fewer top 10%
2000 fewer undergrads
3000 larger tuition
23% less phd
4% reduced student faculty ratio
6% fewer donations
2000 higher instruct expend

Cluster 26
Much higher test scores
7000 more apps
1600 more accepts
50% from top 10%
6000 higher tuition
1500 higher room and board
16% phds
6% lower student faculty ratio
9% donate
18000 instruc expend per student
28% grad rate

Appendix C – Variable Mean Results

Cluster 27
Much lower test scores
2800 higher apps
1100 more enrolled
10% fewer from top 25%
6000 more undergrad
5000 less tuition
9% more with phds
3% higher student faculty ratio
8% fewer donate

Cluster 28
Slightly lower test scores
2000 fewer apps
6% fewer from top 10%
2700 fewer undergrads
1300 higher in state
15% fewer phds
2% lower ratio
1300 lower instruct expend

Thank you!
Questions?