Stat 159 - Project 3

Kartik Gupta
Lauren Hanlon
Morgan Smart
Dakota Lim

November 28, 2016

Abstract

In this paper we will consider the data provided in collegescorecard.ed.gov to consult on the following issue:

> *"The CEO of a biotech startup is looking for candidates. She is interested in diversifying the workforce in regards to women in STEM (Science, Technology, Engineering, and Math). Where should the startup focus their recruitment and outreach for maximum impact?"*

To ensure we use the newest data on recent college graduates we will only consider the most recent datafiles, ignoring older data. Next we will create a scoring system to rank colleges based on where we believe our client will have the most success in recruiting top tier female STEM majors.

Intrduction

We begin by cleaning the data. Since our raw data file contains hundreds of features, we create a new data matrix with only the desired predictors (see the Data section for further details). We also create several new features by combining columns, which imroves the regression model's interpretability by creating more consise predictors. All this is done through the `code/scripts/data-processing-script.R` script.
Next we consider a cross-validated linear regression to predict the mean earnings of female students working and not enrolled 6 years after entry

(MN_EARN_WNE_MALE0_P6) using the new data matrix, which we will use to create a scoring system for all schools. We use the mean earnings as a metric to represent the overall quality of a college's graduates, the intuition being that smarter more qualified candidates will be paid more. After preforming the regression, we check that all features are significant and use the weights of all significant terms.
We then use the regression equation, with only the significant terms included, to create a score for each school. After giving each collge a score, we will rank them by said score and present the top institutions to our client.

Data

For our consultation we consider the following predictors from the scaled data set produced from the 'data-processing-script.R':

1. $MN\_EARN\_WNE\_MALE0\_P6$ = Mean earnings of female students working and not enrolled 6 years after entry
2. $SATMTMID$ = Median SAT math score
3. ADM_RATE = Admission rate
4. $STEM\_DEG\_WOMEN$ = Approximate number of female STEM majors
5. $WOMENOONLY$ = Indicator if schools that are only female
6. $HIGHDEG4$ = Indicator if the school offers Bachelors degrees
7. $COUNT\_WNE\_MALE0\_P6$ = Number of female students working and not enrolled 6 years after entry

These predictors will be used to create a regression model predicting and, ultimately, create a score for each school.

Cleaning the Data

We chose the variables that seemed the most predictable of technical job performance and the variables that were about number of females in technical fields at a school. The combination of these two types of features will yield the best schools to target (in terms of number of potential recruits as well as recruit's skill level)

In cleaning the data, we decided to remove entries that contained missing values. If there was a missing value of the features we used, we removed that row from the dataset.

Other notable operations we performed on the data to clean it included removing schools that were no longer in operation, removing schools with only male students, and removing schools that specialize in a non-stem field (i.e. associate colleges). We only kept schools that have a predominant degree type listed as well as schools that offer at least bachelors degrees.

We had to change the data type of most columns either to numeric or a factor.

We added in a few columns of our own based off of the information given to us. These included columns such as the percentage of engineering/math/science/tech degrees awarded. We also calculated the number of STEM degrees awarded to women based off of a series of conditions. We did this so that we could analyze the data in terms that would be valuable to our client.

In addition to cleaning the dataset, we also created dummy variables for categorical variables. We also chose to mean center and standardize our variables to allow for accurate predictions and more straight-forward analysis.

Descriptive Statistics

To initially explore the data, we first broke up the dataset into general data, including:

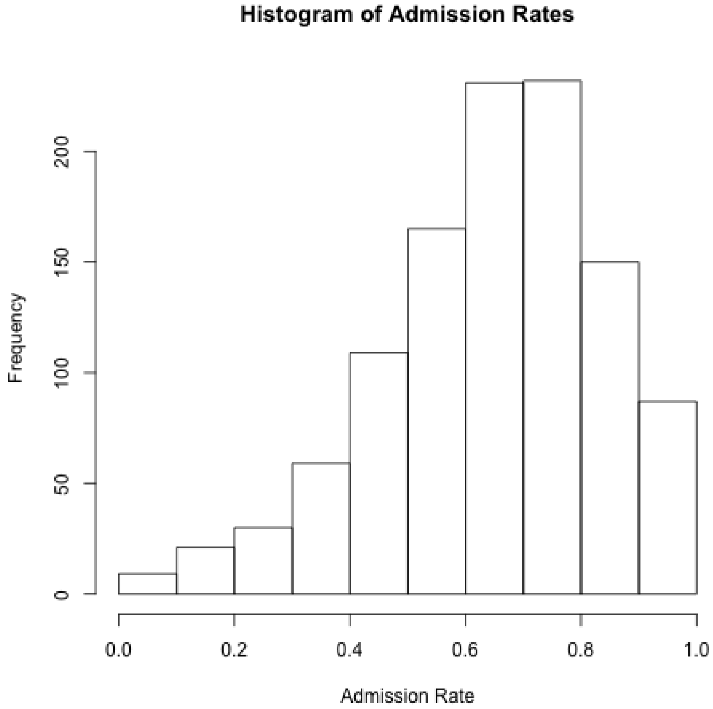
1. ADM_RATE
2. $SATMTMID$
3. $UGDS\_WOMEN$
4. $MN\_EARN\_WNE\_MALE0\_P6$
5. $COUNT\_WNE\_MALE0\_P6$

The summary statistics are shown below, but we would like to note some interesting facets of this table. The first being that the average admission rate is 0.647, which is for colleges in general. The mean SAT score for these schools is 531 out of the 770 maximum score, indicating that the mean SAT score for these schools is slightly below the median of 520. To turn our attention to focus on the data we have on women, we look at the number of undergrads enrolled in schools. The mean is actually above 0.50, indicating that more than half of the student body is made up of women. The mean earnings of women after 6 years of graduating hovers at around $34,000, while the maximum earnings reach $100,100. We note the high range of salaries, with a range of almost $80,000. The number of females working after 6 years is on average 1,000 and also has a relatively high range of 8,650.

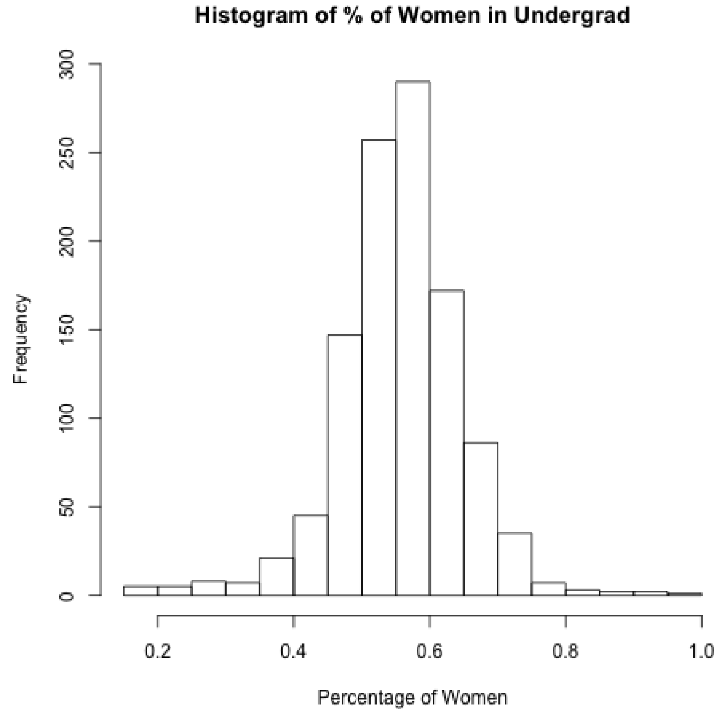
	ADM_RATE	SATMTMID	UGDS_WOMEN	MN_EARN_WNE_MALE0_P6	COUNT_WNE_MALE0_P6
1	Min. :0.0509	Min. :310.0	Min. :0.1505	Min. : 18300	Min. : 51
2	1st Qu.:0.5332	1st Qu.:485.0	1st Qu.:0.5094	1st Qu.: 29100	1st Qu.: 267
3	Median :0.6662	Median :520.0	Median :0.5580	Median : 33600	Median : 541
4	Mean :0.6469	Mean :531.5	Mean :0.5564	Mean : 34612	Mean :1038
5	3rd Qu.:0.7792	3rd Qu.:565.0	3rd Qu.:0.6084	3rd Qu.: 37900	3rd Qu.:1204
6	Max. :1.0000	Max. :770.0	Max. :0.9644	Max. :100100	Max. :8700

Table 1: General Summary Statistics

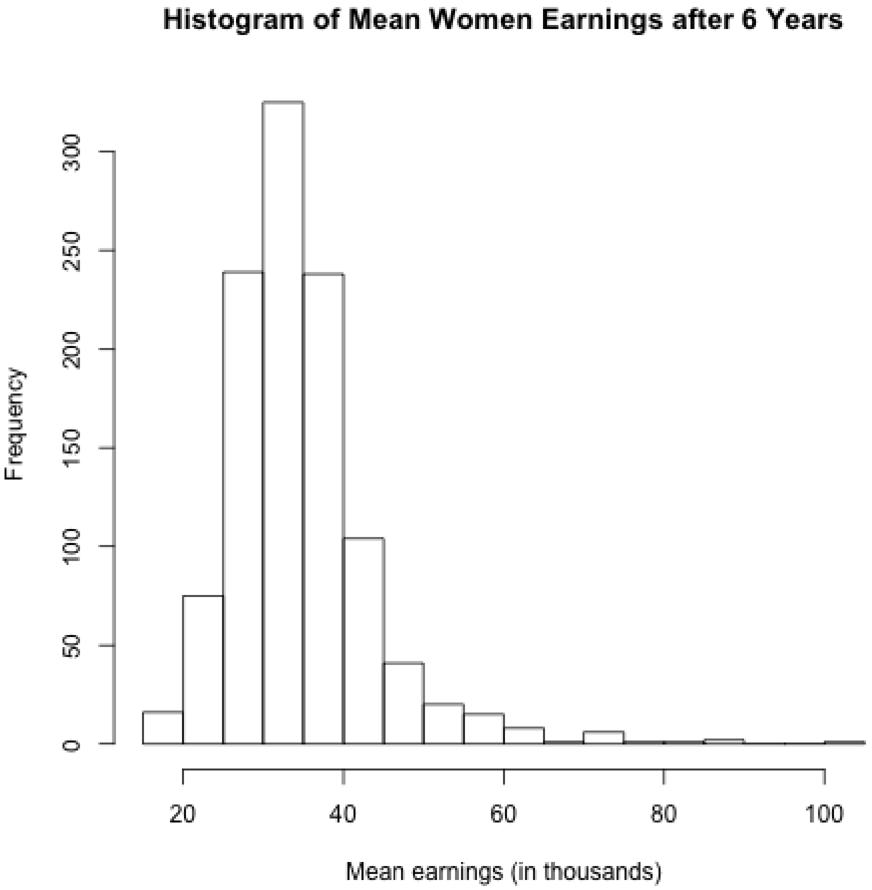
In the visualization below we note that admission rates center in the 60-80% range.



The histogram of the prearrange of women in undergrad is a visualization of the fact that there are over 0.50 women in most undergraduate student bodies.



The histogram of mean women earnings after 6 years displays the wide spread in incomes we found in the data. We see the majority of incomes below $50,000, with a very small frequency of incomes above that mark.



The next section of data we chose to explore were the degree types, namely the number of science, math and engineering type degrees. Shown in a table below, the means for these degrees range, with the highest average percentage in PCIP26, which is a science type of degree. The lowest average degree type is PCIP41, which is another type of science type degree.

	PCIP26	PCIP29	PCIP41	PCIP27	PCIP14	PCIP15
1	Min. :0.00000	Min. :0.0000000	Min. :0.000000	Min. :0.00000	Min. :0.00000	Min. :0.000000
2	1st Qu.:0.02650	1st Qu.:0.0000000	1st Qu.:0.000000	1st Qu.:0.00410	1st Qu.:0.00000	1st Qu.:0.000000
3	Median :0.04980	Median :0.0000000	Median :0.000000	Median :0.00910	Median :0.00000	Median :0.000000
4	Mean :0.05935	Mean :0.0001959	Mean :0.000163	Mean :0.01175	Mean :0.03536	Mean :0.007898
5	3rd Qu.:0.08430	3rd Qu.:0.0000000	3rd Qu.:0.000000	3rd Qu.:0.01590	3rd Qu.:0.03390	3rd Qu.:0.000000
6	Max. :0.36400	Max. :0.1233000	Max. :0.083900	Max. :0.08380	Max. :0.87400	Max. :0.289600

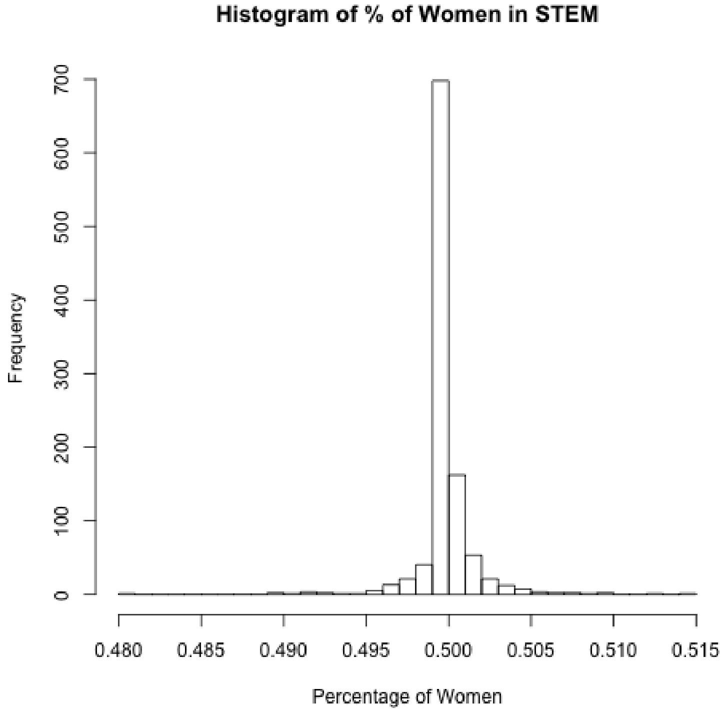
Table 2: Degree Type Summary Statistics

The third and final section of data we explored were the degree percentages per school. We looked at the percentage of engineering, math, science, tech, overall STEM then the percentage of STEM degrees awarded to women. The number for percentage of STEM degrees awarded to women was based off of the number of STEM degrees awarded to women divided by the total number of STEM degrees awarded.
The highest average degree type awarded were to science degrees, and the lowest average degree type awarded was math. The STEM degree type percentage average was just below 0.30, while the percentage of STEM degrees awarded to women were 0.50.

	ENG_DEG_PCNT	MATH_DEG_PCNT	SCIENCE_DEG_PCNT	TECH_DEG_PCNT	STEM_DEG_PCNT	STEM_DEG_PCNT_WOMEN
1	Min. :0.00000	Min. :0.00000	Min. :0.0000	Min. :0.00000	Min. :0.0000	Min. :0.4800
2	1st Qu.:0.00000	1st Qu.:0.00410	1st Qu.:0.1246	1st Qu.:0.00490	1st Qu.:0.1752	1st Qu.:0.5000
3	Median :0.00000	Median :0.00910	Median :0.1946	Median :0.01510	Median :0.2572	Median :0.5000
4	Mean :0.04326	Mean :0.01175	Mean :0.2085	Mean :0.02148	Mean :0.2850	Mean :0.5014
5	3rd Qu.:0.04970	3rd Qu.:0.01590	3rd Qu.:0.2732	3rd Qu.:0.02750	3rd Qu.:0.3680	3rd Qu.:0.5000
6	Max. :0.87400	Max. :0.08380	Max. :0.6925	Max. :0.31700	Max. :1.0000	Max. :1.0000
7						NA's :35

Table 3: Degree Percentages Summary Statistics

The visualization below shows a histogram of the total number of STEM degrees awarded, and we see a large percentage clustered below 5,000. The visualization directly following shows the percentage of these STEM degrees awarded to women. For this histogram we removed the schools with only women, and also only kept schools where there was some number of STEM degrees awarded to women. We see that the highest frequency is just below 0.50, and the spread is relatively small, between 0.48-0.52, meaning that most schools award a equal percentage of STEM degrees to women and men.



Correlation

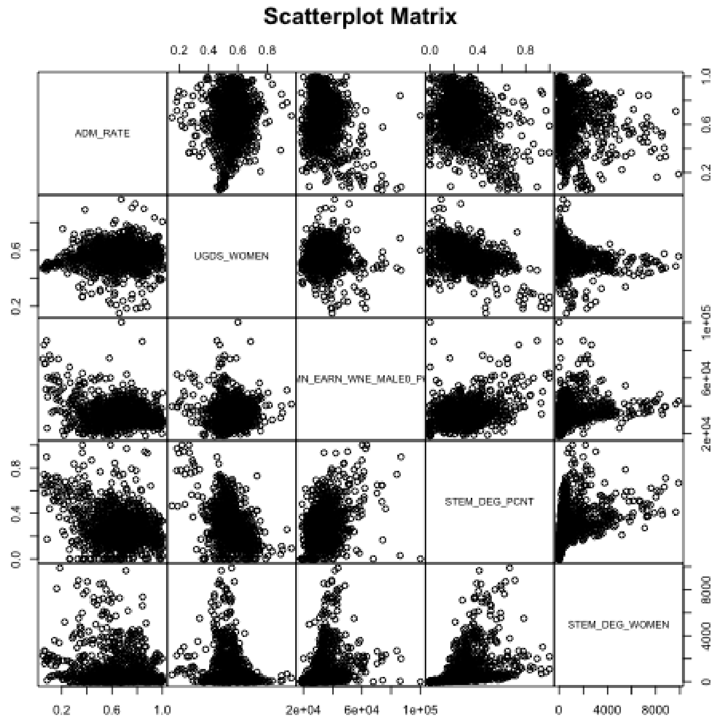
We wanted a way to see the correlations between variables in this dataset. To create our correlation matrix, we chose a few variables which we thought would have interesting correlations to study. For this we chose ADM_RATE, UGDS_WOMEN, MN_EARN_WNE_MALE0_P6, WOMEN_TOTAL, and STEM_DEG_WOMEN.

In the table below, we can see the correlations between these variables. There seems to be a very strong correlation between the mean earnings of women after 6 years and the number of stem degrees awarded to women. From this we can gather that women with degrees in STEM might have higher average earnings than other degrees. Other strong correlations to note are the admission rates and the number of undergraduate women at a school, as well as the mean earnings of women and the total percentage of STEM degrees awarded at a school.

We can see a visualization of these relationships below as well.

	ADM_RATE	UGDS_WOMEN	MN_EARN_WNE_MALE0_P6	STEM_DEG_PCNT	STEM_DEG_WOMEN
ADM_RATE	1.00	0.08	-0.25	-0.31	-0.18
UGDS_WOMEN	0.08	1.00	-0.07	-0.45	-0.24
MN_EARN_WNE_MALE0_P6	-0.25	-0.07	1.00	0.37	0.24
STEM_DEG_PCNT	-0.31	-0.45	0.37	1.00	0.42
STEM_DEG_WOMEN	-0.18	-0.24	0.24	0.42	1.00

Table 4: Correlation Matrix



Methods

In order to properly consult our client, we build score function and apply it to each school, recommending the top institutions to our client. To create the score function we use ordinary least squares regression (OLS) and cross validation.

OLS is an approach to predicting a quantitative response Y based on a multiple predictor variables X_1 through X_p, where Y and X_1 through X_p

are vectors and each value in each X_{ij} (x_{ij}) has a corresponding value in Y (y_i). The model assumes that the relationship between every X_i and Y is linear and that each X_j isn't correlated with any other X_j. OLS can be written as $Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$, where β_0 is the intercept and β_1 through β_p are the slopes of their corresponding predictor variable X_j. The beta values are all constants, unknowns, and together are the model coefficients. The interpretation of β_0 is the expected mean value of Y without a predictor variable and the interpretation of β_1 through β_p is the change in Y for a unit increase in the beta's corresponding X_j. Although the betas are unknown, we can estimate them using the OLS model: solving for the intercept and slopes that produce the plane closest to each point (x_{ij},y_i) in each X_j,Y, which is minimizing the residual sum of squares (RSS). Once we have an estimate for the betas, we can compute using OLS to determine the strength of the relationship between each X_j and Y, if the relationship is statistically significant, and asses how accurately the model predicts the relationship.

Cross-validation is a model validation technique used to assess how the results of an analysis will generalize to an independent data set. To cross-validate a model, we fit its required estimated parameters using k parts of a training dataset and then test the k created models on an unseen test dataset. We then chose the model that minimizes the predictive error. If the model performs well using the test dataset (meaning it has a low predictive error), we are more confident that this model is accurate (since it produces good prediction of data not used to fit the model). For our purposes, we will cross validate the resulting model from our OLS regression and choose the β vector that has the minimal cross-validated mean squared error.

Analysis

After running our regression script we observed the following values for our $\hat{\beta}_{OLS}$ estimator:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.00	0.02	-0.00	1.00
ADM_RATE	-0.07	0.03	-2.77	0.01
SATMTMID	0.61	0.03	21.75	0.00
STEM_DEG_WOMEN	-0.15	0.03	-5.05	0.00
COUNT_WNE_MALE0_P6	0.15	0.03	5.44	0.00
WOMENONLY	0.06	0.02	2.49	0.01
HIGHDEG4	0.21	0.02	8.55	0.00

Table 5: Predictor Significance

It's clear from the final column of the tabel that all the regressors are significant (with the exception of the intercept). Thus, we will utilize all values *except* the intercept in our score function.
We use OLS and only OLS because we wanted to use all of the features and there are very few features we're considering. If anything, we're lacking in features (due to the limitations of the initial dataset). The mean squared error of the prediction was 0.589633015445381. The mean squared error tells us how accurate the prediction is. However, we're not trying to predict salary, we simply was to see the magnitude of each feature on salary and then use these magnitudes as weights to determine the score of a school: predictability doesn'tmatter, just the magnitude of each feature. We chose salary as the response because we made the assumption that salary was a good measure of how well an employee preformed.

Results

From our 10-fold cross validation we observe a mean squared error of: 0.6064

Next we build the score function, utilizing the above $\hat{\beta}_{OLS}$ values as the weights for our function. This yields the following expression:

$f(X_{\cdot,i})$:= (-0.0715) $X_{1,i}$ + 0.6094 $X_{2,i}$ + (-0.1537) $X_{3,i}$ + 0.1469 $X_{4,i}$ + 0.0582 $X_{5,i}$ + 0.2076 $X_{6,i}$

Where $X_{j,i}$ represents the j-th regressor value from our OLS regression of the i-th observation in the data matrix. Using this equation, we apply it to all schools in our data matrix and sort by their relative score to yield the top 20 colleges we recomend the client recruits from:

SCORE	INSTNM	CITY
2.14	Massachusetts Institute of Technology	Cambridge
2.06	University of Chicago	Chicago
2.04	Washington University in St Louis	Saint Louis
2.04	Yale University	New Haven
2.03	Vanderbilt University	Nashville
2.02	Harvard University	Cambridge
2.01	Princeton University	Princeton
1.99	Rice University	Houston
1.98	Columbia University in the City of New York	New York
1.96	Northwestern University	Evanston
1.94	Carnegie Mellon University	Pittsburgh
1.93	Stanford University	Stanford
1.93	Duke University	Durham
1.92	Claremont McKenna College	Claremont
1.91	University of Pennsylvania	Philadelphia
1.85	Johns Hopkins University	Baltimore
1.80	Dartmouth College	Hanover
1.79	Williams College	Williamstown
1.78	Brown University	Providence
1.75	University of Notre Dame	Notre Dame

Table 6: Top 20 Colleges to Recruit From

Recommendations

After preforming a preliminary examination of the data, building the previously detailed OLS regression model, and building our score function, we recommend our client recruit from the top 20 schools given in Table 6.
It should be noted that our score function is derrived from our regression model, with the coefficients used were the same β values becoming the weights in the function. Using this data, we are sure our client will be able to recruit highly qualified female STEM majors for her company.

Considerations

One major shortcoming of our analysis for this project was that we didn't have exact statistics for the number of females graduating in each STEM major type, but instead had to assume that half of the STEM majors were female. In reality this might have been an overestimate, but this was the closest approximation given the data provided.
In analyzing the quality of students at a particular school, we would have liked to look at GPA as an indicator of one's intelligence rather than SAT Math scores.

Conclusions

After considering the request of our client, we filtered the data, built an OLS regression model, considered the significance of the coefficients, and built a comprehensive scoring function for any given college. Using this we were abel to recommend the following 20 universities to recruit from in Table 6.
It should be noted that the structure of this paper allows for it to be reproduced with each new data upload on collegescorecard.ed.gov