Paper 4054-2019

Using SAS® Enterprise Guide® and SAS® Visual Analytics to Model a Large Institutional Data Set

Pratik Patel, Maham Khan; Varsha Sharma, Shreyas Dalvi,
Advisor: Dr. Shabnam Mehra
University of South Florida

ABSTRACT

Insights and understanding of factors associated with student success is of great significance among higher education universities and the prospective student population including their parents. Prospective students are interested in learning about completion rates and their financial well-being after graduation in the form of expected earnings across universities. Various SAS® procedures in Enterprise Guide® software were used to import, merge, clean, and analyze a big dataset of universities and colleges collected by over a period of 10 years. The study uses publicly available College Scorecard data to study the factors associated with student success and financial well-being. SAS® Visual Analytics software was used in this study to visualize and present the findings of the study. In this study we find a few institutional characteristics to be strongly related to both student success and financial well-being.

INTRODUCTION

Institutional characteristics play a vital role in student success. Metrics for student success are a deciding factors for students to choose their school of higher education. Metrics that public institutions and students alike are most interested in are student's retention and completion rates. A student's financial well-being is another unique criteria for schools to measure student success, and of course this is of primary interest to prospective college students and their parents. This document will be used to explore if the same institutional characteristics that are related to student success are also related to financial well-being using SAS® Enterprise Guide® software and SAS® Visual Analytics software, we will also highlight how they are different.

Variables for our project were selected based on their relevance with the topic and their connectivity to our research questions.

1. Is student success at public four-year colleges associated with higher levels of financial well-being?
2. What Institutional Characteristics are associated with Student Success at four-year public colleges?
3. Are the same Institutional Characteristics that are associated with Student Success also associated with Financial Well-being? How are they different?

Our variables were thus segmented under three broad categories;

1. Student Success
    a. 4 Year Graduation Rate
    b. 4 Year Retention Rate

2. Institutional Characteristics
    a. Share of Undergraduate White students
    b. Share of Undergraduate Black students
    c. Share of Undergraduate Hispanic students
    d. Share of Undergraduate Asian students
    e. Tuition per full-time student
    f. Instructional expenditure per full-time student
    g. Average faculty salary
    h. Percent of full-time faculty
    i. Percent of financially independent students
    j. Percent of first-generation college students
    k. Median family income
    l. Share of undergraduates that are women
    m. Admission rate
    n. Median SAT verbal score
    o. Median SAT math score
    p. Percent of students who received a federal pell-grant
    q. Percentage of students who received a student loan
    r. Cost of attendance

3. Financial well-being
    a. Median earnings of graduated students after 6 years
    b. Median earnings of graduated students after 8 years
    c. Median earnings of graduated students after 10 years

We then set out to answer the above research questions using College Scorecard Data with the above variables.

DATA STRUCTURE AND PROCESSING

A variety of Base SAS® Procedures in SAS Enterprise Guide® was used and helped process multiple Microsoft Excel files to SAS® Datasets, merge the data, clean/filter our data, and of course help analyze it.

SCORECARD DATA

Selecting the most suitable as well as applicable database was an essential step in this project. Therefore, after reviewing a number of big databases like the Integrated Postsecondary Education Data System (IPEDS) & the National Student Loan Data System (NSLDS), we selected College scorecard data. College Scorecard is a database platform that compares the success rate of different institutions across the nation. It contains about 1900 different variables many of which are metrics that are used to assist the students and their

families to make a better decision about their lives and their future goals for secondary education. The source of this data is through federal reporting from the institutions. There are also several elements that provide information about the institutions itself. These include identifier, location, degree type, and profile, programs offered, and the academic profile of the students enrolled. Most of these elements in the College Scorecard are available from the IPEDS and the NSLDS databases.

The Scorecard dataset presents a unique combination of two large data sources. IPEDS contains a vast array of information on the institutions themselves. This allowed us to choose appropriate variables. It also provided as a platform to compare multiple institutions of United States. NSLDS is a data system connected with student federal aid data this large data system stores information related to student financial aid dispersion and repayment.

IMPORT, TRIMMING & MERGING

Since our data was abstracted into large multiple excel files from College Scorecard website our first SAS® task was to import the data using a repetitive Import Macro which utilized the PROC IMPORT procedure to do the heavy lifting:

```
%MACRO IMP(OUT, INPUT);
PROC IMPORT OUT = &OUT DATAFILE = &INPUT
            DBMS = XLSX REPLACE;
            GETNAMES = YES;
RUN;
%MEND IMP;
```

Our next step was to merge these goliath datasets into one. This one was done using PROC DATASETS with the FORCE Option. But prior to this a Macro which trimmed each imported SAS® Dataset was used where only 76 out of 1900 variables were selected; a DATA Step with a KEEP statement inside the Macro did the trick. Below is the code framework for our PROC DATASETS procedure:

```
PROC DATASETS NODETAILS FORCE NOLIST;
      /* First Append */< APPEND BASE = new-dataset-name DATA =
existing-dataset;  >
      …
      /* Last Append */ <APPEND BASE = new-dataset-name DATA =
existing-dataset; >
RUN;
```

CLEANING & TRANSFORMING

Data cleaning and reduction was performed in the next step. SAS® Arrays were used to remove values that were "NULL" and the ones that had "Privacy Suppressed" values. Additionally, we used the %DROPMISS Macro from a previous SAS® Global Forum

proceedings paper; "Dropping Automatically Variables with Only Missing Values" by Selvaratnum Sidharma. This Macro was used to eliminate any empty variables in our dataset following the use of the Arrays. We structured our data cleaning array as follows:

```
ARRAY WIPE[*] $ _CHARACTER_ ;

DO I = 1 TO DIM(WIPE);
     IF WIPE[I] = "NULL" THEN WIPE[I] = '';
     IF WIPE[I] = "PrivacySuppressed" THEN WIPE[I] = '';
END;
```

As the attentive reader may have noticed, the array above works on only character formats. Because our data for institutions are largely quantitative we split our data into two SAS® datasets so that the conversion of character to numeric data types in SAS® would be easier. This was done using a simple KEEP _NUMERIC_ and KEEP _CHARACTER_ option in a DATA step. Since our underlying data structure had many percentages in scientific notation we used another SAS® array with the FIND function to correct these values so that the conversion to numeric would work quickly and efficiently even while character type values were still present:

```
ARRAY CONV[*] (start-variable) -- (last-variable);

DO I = 1 TO DIM(CONV);
     IF FIND(CONV[I],'E-3') THEN CONV[I] = CONV[I]*1 ;
     IF FIND(CONV[I],'E-4') THEN CONV[I] = CONV[I]*1 ;
     IF FIND(CONV[I],'E-2') THEN CONV[I] = CONV[I]*1 ;
END;
```

Here the multiplier '1' helps re-adjust the variable value that is in scientific notation to one in decimal form. For example, 1.96E-2 becomes 0.0196. The advantage of using this array is that it snoops out values in all the specified variables and corrects them thanks to the conditional FIND function. This helps ensure the values are converted properly in a subsequent use of the INPUT function to convert the character variables into numeric. Upon correcting the form of the numbers, we used SAS® formats and the INPUT function to convert all the character variables that were supposed to be numeric into numeric type. Whereas for financial variables in order to format the data for our analysis we transformed these values by dividing the dollar amounts by 1000, using an Array.

After data cleaning and data trimming was performed we applied a filter for public institutions that granted at least a 4-year degree which were the focus of our analysis. A Data step was used to filter for these public institutions.

The final step for data cleaning and reduction was performed using a PROC SORT with the NO-DUP option. This procedure helped in removing any row that shows up more than one time in the data just as a precaution against an extra dataset append in the PROC DATASETS step.

DATA EXPLORATION & VISUALIZATION

The following statistical procedures were used to explore the date by checking for each variable's correlation and association of possibly interrelated variables. This type of statistical exploration was done to make decisions as to which variables should be used in our subsequent modeling analysis step.

1. Proc Freq
2. Proc Means
3. Proc Corr
4. Proc T-test
5. Proc ANOVA

PROC T-TEST and PROC ANOVA were used to assess the relationships between categorical and quantitative variables. The two categorical variables assessed using these procedures were primarily Main Campus designation and Region that the institution or school was located in.

To further look at the data we uploaded our dataset to SAS® Visual Analytics software where we were able to derive unique insights about the relationship between our variables of interest. Using SAS® Visual Analytics enabled us to look at many repeatedly measured data points in a few single charts. Figure 1 and 2 show some of our strongest correlated variables.

Figure 1

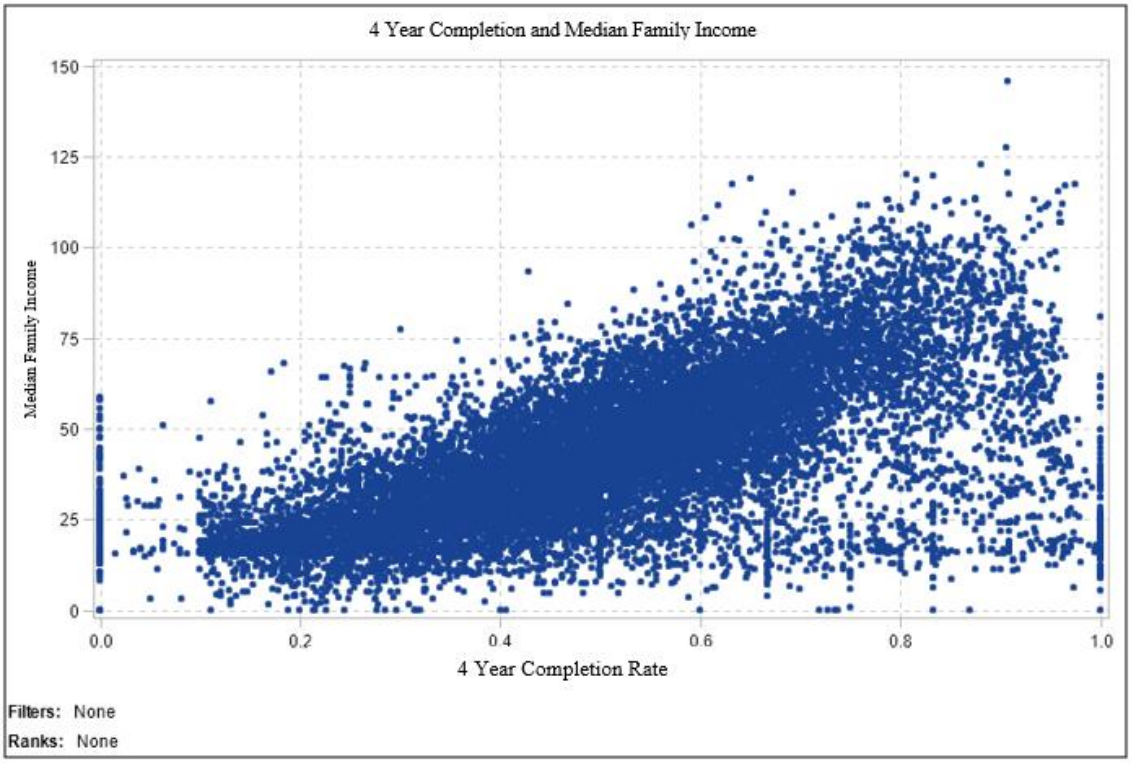

Figure 1. SAS® Visual Analytics Plot of 4 Year Completion and Median Family Income

Figure 2

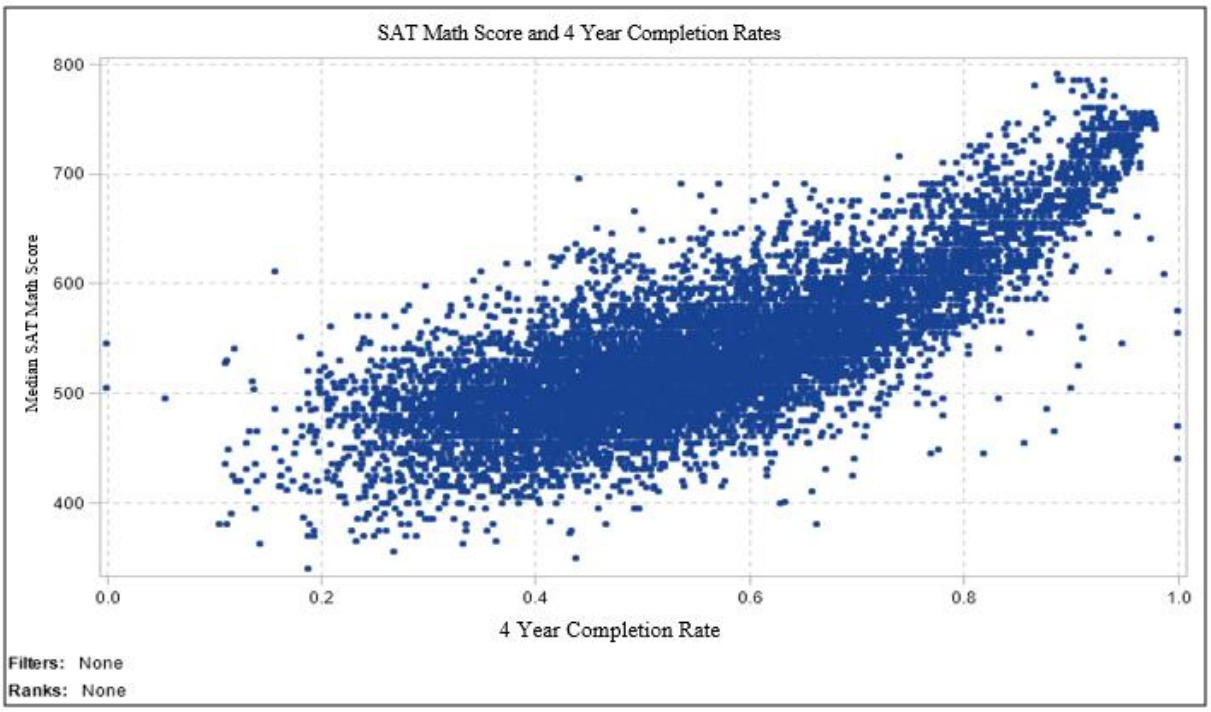

Figure 2. SAS® Visual Analytics Plot of 4 Year Completion and SAT Math scores

We can see strong linear relationships between the retention rates and midpoint of math SAT Scores, and the same with the annual family income.

PROC GENMOD: GEE ESTIMATION

Because our scorecard data is from a period of 2005 to 2015 where repeated measurements were made we needed a modeling method that could account for repeated measures of the institutions. Luckily for many statisticians and data analysts a very general but powerful functionality in the PROC GENMOD procedure is available. Using the REPEATED SUBJECT statements, we can analyze each institution and their dynamically changing characteristics over the years by treating each measurement of a specific college or university as a related but statistically independent observation. We used the WITHINSUBJECT option with a 'YEAR_TIME' variable to specify the proper sequence of measurements in order from the oldest to the most recent year which would allow SAS® to properly handle missing information for some of the institutions over the years in the dataset. Since our outcomes are continuous variables we modeled our data using the link function (the standard ordinary least squares regression function).

See below for its full specification as we look at Median earnings of each institution's student body 8 years after completion regressed on 4-year completion rates and 4-year retention rates of the institutions. Main and Region are categories added into the model to account for differences between institutions by main or non-main campus and geographic regions within the U.S:

```
PROC GENMOD DATA = SCORE.SASGF2019_GA_ANALYZE;
CLASS UNITID MAIN REGION YEAR_TIME;
MODEL MD_EARN_WNE_P8_N_1000 =
C100_4_N
RET_FT4_n
MAIN
REGION
/ LINK=IDENTITY;
REPEATED SUBJECT = UNITID / WITHINSUBJECT = YEAR_TIME;
WHERE CONTROL = 1;
RUN;
```

KEY FINDINGS

All our model results are presented in tables. We only include significant ($p < 0.05$) Beta estimates in each table.

Our findings from our model of Student Success and Financial Well-being indicate that public institutions with higher student retention and completion rates have a greater number of students that earn more money upon graduation (Table 1).

Table 1

Student Success and Financial Well-Being Models		
Student Success (Predictors)	**Financial Well-Being (Response)**	**Beta**
4 Year Retention Rate	Median Earnings	
	6 Years After Completion	26.15
	8 Years After Completion	31.23
	10 Years After Completion	35.54
4 Year Completion Rate		
	6 Years After Completion	6.6862
	8 Years After Completion	8.4218
	10 Years After Completion	9.613

Table 1. Student Success and Financial Well-Being Models with Significant Covariates only shown

Table 2 and 3 below show us that the institutional characteristics that are most associated with both higher 4-year retention and completion are higher average faculty salary, higher median family income of their student body, a higher share of women students, a higher median SAT Math score, and a lower percentage of First-generation college students.

Table 2

Institutional Characteristics and Student Success Model		
Institutional Characteristics (Predictors)	**Student Success (Response)**	**Beta**
Share of Undergraduate Whites	4 Year Retention Rate	-0.0245
Share of Undergraduate Black	4 Year Retention Rate	0.0299
Share of Undergraduate Hispanic	4 Year Retention Rate	0.0960
Instructional Expenditure per Full Time Student	4 Year Retention Rate	-0.0013
Average Faculty Salary	4 Year Retention Rate	0.0161
Median Family Income	4 Year Retention Rate	0.0011
Percent of First Generation Students	4 Year Retention Rate	-0.1150
Share of Undergraduate Women	4 Year Retention Rate	0.0659
Median SAT Verbal Score	4 Year Retention Rate	0.0002
Median SAT Math Score	4 Year Retention Rate	0.0006

Table 2. Institutional Characteristics and Student Success Model with 4 Year Retention as the response (Significant Covariates only shown)

Table 3

Institutional Characteristics & Student Success Well-Being Model		
Institutional Characteristics (Predictors)	**Student Success (Response)**	**Beta**
Median Family Income	4 Year Completion Rate	0.0036
Cost of Attendance	4 Year Completion Rate	0.0072
Instructional Expenditure Per Full-time Student	4 Year Completion Rate	0.004
Share of Undergraduate Women	4 Year Completion Rate	**0.317**
Percentage of Financially Independent Students	4 Year Completion Rate	**-0.1488**
Median SAT Verbal Score	4 Year Completion Rate	**0.0004**
Median SAT Math Score	4 Year Completion Rate	**0.0007**

Table 3. Institutional Characteristics and Student Success Model with 4 Year Completion as the Response (Significant Covariates only shown)

Figure 3

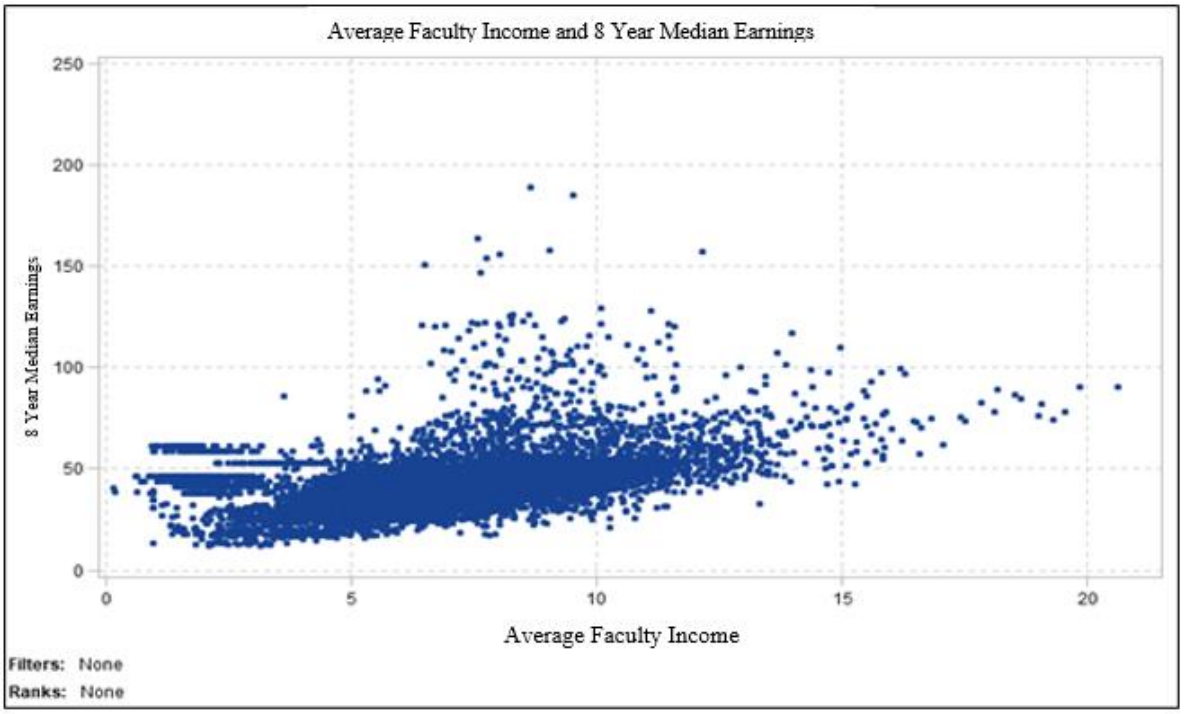

Figure 3. SAS® Visual Analytics Median Earnings after 8 Years and Average Faculty Salary

Table 4 shows the Model of institutional characteristics and Financial Well-being after 8 years post-graduation. Comparing this to Tables 2 and 3 we see that median family income is also positively associated with median earnings 8 years after graduation as it is for 4-year retention and 4-year completion rates in each institution. SAT Math score is also positively

associated with median earnings and student success metrics. We see an interesting divergence of results however for share of undergraduate women (Tables 2, 3 and 4).

Table 4

Student Success and Financial Well-Being Model		
Institutional Characteristics (Predictors)	Financial Well-Being (Response)	Beta
Share of Undergraduate Whites	Median Earnings 8 Years After Completion	-3.25
Share of Undergraduate Asians	Median Earnings 8 Years After Completion	10.1375
Average Faculty Salary	Median Earnings 8 Years After Completion	0.8023
Median Family Income	Median Earnings 8 Years After Completion	0.2193
Percentage of Financially Independent Students	Median Earnings 8 Years After Completion	17.097
Share of Undergraduate Women	Median Earnings 8 Years After Completion	-29.0905
Percentage of Pell Grant Recipients	Median Earnings 8 Years After Completion	-6.7817
Percentage of Student Loan Recipients	Median Earnings 8 Years After Completion	-3.5125
Admission Rate	Median Earnings 8 Years After Completion	-4.2612
Median SAT Verbal Score	Median Earnings 8 Years After Completion	-0.0305
Median SAT Math Score	Median Earnings 8 Years After Completion	0.0462

Table 4. Institutional Characteristics and Financial Well-Being Model with Median Earnings 8 Years after Completion as the Response (Significant Covariates shown only)

Figures 1 through 3 show some of our most interesting findings of our data exploration from using SAS Visual Analytics.

CONCLUSION

Answering our first research question we found that institutions with higher amounts of student success are related with a higher earnings of students that graduated from that institution. Secondly, we were able to determine many institutional characteristics that were statistically related to higher levels of student success at the analyzed institutions (See Tables 2 and 3). We also were successful in identifying institutional characteristics that are related to higher median earnings of their students. Comparatively, we find that higher median family income and SAT math scores are both related to student success and financial well-being. We also find that the while institutions with higher percentage of women have higher student success metrics, these institutions have a student body that earns less 8 years after graduation.

Overall, this paper presents the utility of SAS® software in importing, merging, and cleaning large complex datasets of institutions collected over many years using many built in Base SAS® procedures and data structures. PROC GENMOD contains a useful way to model complex longitudinal data where repeated measurements are expected and where there is a high possibility of finding large amounts of missing data points.

SAS Visual Analytics is also a powerful software suite that allows direct analysis and easy visualization at even large datasets with repeated longitudinal measurements over the

years. This paper thus has demonstrated the use of SAS Visual Analytics as a complementary component to analysis of large datasets from a big-data source.

RECOMMENDATIONS

In United States education system is divided by the type of institutions. The division segregates the schools into two major groups of public institutions and private institutions. Public institutions run through the government funds majorly. Whereas, the private institutes operate according to their own organizational structure that's either for-profit or not-for-profit.

Although, College Scorecard data includes all types of higher educational institutions but for this project we restricted our analysis to only public institutions. As our focus was to review the effects of institutional characteristics on student success and financial well-being for only publicly funded schools. However, the given model can be used to analyze the similar characteristics for each other type of institution as well or it can be used to analyze the difference between three types of institutions; public, private for profit and private not for profit.

The dataset of this project can be used by institutions to rank themselves among their peers as well as competitors. Therefore, this project can provide a strong base to institutions to strengthen their weak institutional characteristics strategically so that they eventually enhance their performance.

ACKNOLWEDGEMENTS

This project consumed huge amount of research and dedication. Still, implementation would not have been possible without the support of many individuals within our organization. We would like to specially thank the entire team of USF Institutional Research and Analytics and Dr. Shabnam Mehra, Director Institutional Research and Analytics – Office of Decision Support, University of South Florida, for continuous provision of expertise and technical support in the implementation. We would also like to thank Sreekanth Cherapati & Ramya Thiagarajan two Statistical Data Analysts - Office of Decision Support, University of South Florida, for their expert guidance in compiling this project and recommendation for choosing this dataset. Nevertheless, we express our gratitude towards our families and colleagues for their kind co-operation and encouragement which helped us in completion of this project.

REFERENCES

College Scorecard Data. October 30, 2018. "Download the Data". Accessed November 5, 2018 https://collegescorecard.ed.gov/data/

Sridharama, Selbaratnam. April 11, 2010. "Dropping Automatically Variables with Only Missing Values". *Proceedings of the SAS Global 2010 Conference,* Seattle, WA: SAS Global Forum. Available at http://support.sas.com/resources/papers/proceedings10/TOC.html

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please direct correspondence to:

Pratik Patel

Email: pratik1@usf.edu

Maham Khan

Email: khan10@usf.edu

Varsha Sharma

Email: varshasharma@usf.edu

Shreyas Dalvi

Email: shreyasdalvi@usf.edu

Advisor: Dr. Shabnam Mehra

Email: smehra@usf.edu