

Where Did My Students Go?

Stephanie R. Thompson, Datamum

ABSTRACT

Many freshmen leave their first college and go on to attend another institution. Some of these students are even successful in earning degrees elsewhere. As there is more focus on college graduation rates, this paper shows how the power of SAS[®] can pull in data from many disparate sources, including the National Student Clearinghouse, to answer questions on the minds of many institutional researchers. How do we use the data to answer questions such as “What would my graduation rate be if these students graduated at my institution instead of at another one?”, “What types of schools do students leave to attend?”, and “Are there certain characteristics of students who leave, and are they concentrated in certain programs?” The data-handling capabilities of SAS are perfect for this type of analysis, and this presentation walks you through the process.

INTRODUCTION

Graduation rate and persistence are two key measures used to compare colleges. These metrics have been incorporated into college rankings that are published yearly and have been around for many years. They are also part of the discussion regarding college accountability and establishing a government derived rating system. For the most part, these measures are internal to an institution. Occasionally, a state-wide system will generate them across multiple campuses. However, the numbers only tell part of the story.

Colleges are interested in using analytic techniques to identify students likely to leave so that interventions can be developed. The hope is that these students can be enticed to stay and complete their degree. These models can be challenging to develop due to the lack of predictor variables, missing data, dirty data, and a host of other issues. Sometimes, a college is left with looking at who has already left and using that to inform decision making.

Wondering what a graduation rate would be if students did not leave and succeed at another institution is another common question. While there is a certain percentage of students who decide college is not for them, at many talented students that leave do go on to graduate elsewhere. Knowing this can help to show if admissions policies are working and what thing could look like if more students were retained. This is not only for top tier schools – any can benefit from this type of analysis.

The National Student Clearinghouse (NSC) “serves as a single point of contact for the collection and timely exchange of accurate and comprehensive enrollment, degree, and certificate records on behalf of its more than 3,600 participating higher education institutions, which represent 98 percent of all students in public and private U.S. institutions (see references).” Participating institutions have access to information regarding enrollment at other participating institutions. This can be accessed online or in report for to see where departing students have enrolled. Degree information is also provided which can help to determine whether or not the student could have been successful at their first institution. For example, an electrical engineering student leaving a college and graduating with an electrical engineering degree elsewhere may have been successful at the first school. However, if that same engineering student earned a nursing degree and your school does not have a nursing program, this may have been a question of fit from the beginning.

Trying to answer these types of questions is where SAS comes in to play. The data to run these types of analyses can come from many disparate sources. The student record level reports from the NSC are provided in CSV format. Many student record systems store the data in relational databased such as Oracle. Other student data may be kept in Excel workbooks, Access tables, or other types of files and transactional systems. SAS provides the power to put it all together.

An overview of this process, as opposed to the results of the study, is the focus of this paper.

FIRST THINGS FIRST

Prior to preparing a file to submit to the NSC, the population of students to be analyzed must be determined. The information from the NSC can be used to answer any number of questions but the right students must be submitted first. Potential questions are shown in the list below:

- Where do first time full time freshman go if they do not persist and obtain a degree from my institution?
 - Are they enrolling in a 2-year or 4-year school?
 - Is the school public or private?
 - Is the school closer to their permanent home?
 - Is the tuition at the new school less or more than here?
 - Is the institution a competitor or distinctly different?
 - Did the student change majors or are they pursuing the same plan of study?
 - These questions can provide insight into possible reasons why a student left.
- What degree is earned first after leaving your institution?
 - Is it an Associates or Bachelors degree?
 - If the degree offered at my institution?
 - These question can help identify if students are earning degrees available at your college and could have been induced to stay, left to obtain a degree not available and show opportunities for new programs, or indicate that they stopped out of college altogether.
 - Degree data can also assist in calculating a projected graduation rate if those students would have persisted and earned their degree at your institution.
- Are bachelor degree earning students going on to pursue additional education and are they successful in earning a Masters or terminal degree (e.g., PhD, MD, or JD)?

Knowing the questions to be answered in advance will make sure the right students are submitted to the NSC.

PREPARING AN UPLOAD FILE

Before you can analyze the data from the NSC, you must first upload a file in a given format on the students you wish to receive information. The upload file may be created in Excel but the upload must be in text format. Detailed instructions are provided on the NSC website. Following the instructions will avoid delays in getting your data processed.

To obtain enrollment data for students who have left a given institution, the NSC does not use Social Security Number as the key to find students. Students are identified in the database using first name, last name, middle initial, name suffix, and date of birth. This can be challenging as name fields pulled from student transactional databases may have information in a slightly different from than what is needed. Maybe you have middle name and need to shorten to the initial. Thankfully, SAS has several functions that make this very simple to do including SUBSTRING. Some students have two names in the first name field and you need to determine if it is truly a double first name or if the second name is a middle name. Data quality can also be an issue. At times you will see a student name that you can almost guarantee is not their real name. These will need to be researched before preparing the file for the NSC. Without a single key to identify a student, each part of the multiple key is very important.

The birthdate can also be a challenge. The birthdate fields in some systems are actually date time fields and you need to extract the date only. The DATEPART function is best here. Data quality needs to be checked here as well. Just because a date is a valid date it does not mean it is an appropriate birthdate for a student. Determining age based on the birthdate and then reviewing extreme values will help you identify inaccurate dates. You can also look at the distribution of birth years to help you see if there are

issues. It is possible that the birth year could have been entered as 1895 instead of 1995. Correcting these errors will also insure as many records are matched as possible.

One last student identifier can be added to the record of each student for the use of the submitting institution. This identifier is not used by the NSC in finding records but is for you to use later in merging files or keeping students distinctly identified. You may not use Social Security Number here as this is a violation of student privacy regulations. This is stated very clearly on the NSC website. You will not want to use a nine digit number either as your identifier appears to be a Social Security Number. This is unfortunate as many student systems use a nine digit number as the student ID. This identifier can be up to 50 characters and multiple identifiers can be separated by a period. Choose something that will uniquely identify each student and can be easily created from data available for each student. The last 4 digits of a student ID along with student major at time of departure and campus may be sufficient. Using this optional field may make it easier on you when it is time to match up the output to your analysis data file to other data. Otherwise, you will need to use the name fields and birthdate.

Other student-level information required on the upload includes date to start looking for the student in the NSC database, submitting college identifiers, and a record type indicator. The file also needs a header row and footer row. Again, these details are on the NSC website.

An additional suggestion to consider when pulling the data to submit to the NSC is whether this is also a good time to gather additional information that will be used in later analyses once the results are returned. Fields such as gender, race, ethnicity, major enrolled when departed, year level when departed, GPA at departure, semester departed, and other student characteristics might be in the files already being accessed and so make the later analysis faster. Although these data points are not sent to the NSC, you can create a base data set to use later to merge with your returned results.

OVERVIEW OF THE DATA FILE FROM THE NSC

Once you upload your file, a conformation email will be sent. Another will be received when the file has been processed. If the file does not process an email stating so is sent. If you receive this email then you need to determine the issue, correct it, and re-submit the file. Forgetting the footer is a common mistake but it does prevent the file from processing.

Now that your data has been processed, what do you receive from the NSC? Three files are generated by the NSC for each upload. One is in HTML format and the other two in CSV. A link to sample reports is in the references section of this paper. The Control Report, in HTML format, summarizes how many students were submitted, how many were in their database, and whether or not data are blocked for certain students. The report also lists how many colleges in your state with more than 1,000 students do not participate in the NSC. This will help to see possible schools within your state where students may have gone but are not reported in the output file.

The Aggregate Report shows a list of all participating schools where submitted students have enrolled as well as summary statistics. This file also contains information regarding where blocked records are from. Information on degrees earned is also part of the summary.

The file of most interest to the SAS user is the Detail Report. For particular requests (the NSC has different types of reports that can be requested and the output is determined by the information in the header row of the upload file), the file may contain multiple rows per student. There will be a record for each college and reporting period from the date you requested NSC to search for records. If that date overlaps attendance at your institution, it will be included in the Detail Report. Any student not found in the NSC database will be flagged as such and will only have one row in the file. This file will also include the student identifier information submitted with you data file.

Beginning and ending enrollment dates are given in the Detail Report in the format YYYYMMDD. These can be converted to SAS dates by using the input function: `input(enrollment_begin, yymmdd8.)`. Having SAS dates makes thing much easier when it comes time for analysis. Other data manipulation might be necessary to match the data to your records or to make analysis easier.

CREATING AN ANALYSIS DATASET

Now that you have the data from the NSC, you can combine it with other student data to create an analysis dataset. You may use the file created when you pulled the students to send to the NSC or you may need to create it at this time. If you included a student identifier in the upload file and kept it in your data file it will be easy to merge the data in a DATA STEP, with PROC SQL, or another method.

Depending on the type of analysis you are going to run, you may need to have one row per student in your analysis dataset or keep it in multiple rows. Oftentimes, multiple datasets are created to analyze different parts of the information. One dataset may be to analyze degrees earned, another for concurrent enrollments, another for a deep dive into the first college a student attended after departure, and even one for students who were not found in the NSC database.

Once the datasets are created, you can use your favorite SAS tool to run your analysis.

CONCLUSION

The Power of SAS is what makes these types of analyses possible. Without the ability to pull data from multiple sources, merge it together, and manipulate it into a useful structure you could not begin to answer questions with your data. The many options for handling variables and changing them into the format needed are also part of the process. Lastly, the high power analytic tools and methods for visualizing data from SAS makes it easy for the analyst to answer the questions to support data driven decision making and provide output and visualizations that everyone involved in the process can understand.

REFERENCES

National Student Clearinghouse, "About the National Student Clearinghouse," <http://nscnews.org/about-nsc/>.

National Student Clearinghouse, "Sample Report," http://www.studentclearinghouse.org/colleges/files/ST_SampleReport.pdf.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Stephanie R. Thompson
Datamum
901-326-0030
Stephanie@datamum.com
<http://www.datamum.com>

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.