Talk Title: Almost Too Easy: A New Way to Wrangle IPEDS Data
Webinar Date/Time: August 27, 2019, 2pm EST

This script is intended to accompany presentation slides.

Presenters:
Kate Aloisio (KA), Smith College (kaloisio@smith.edu)
Kathy Foley (KF), Smith College (kfoley@smith.edu)
Emma Morgan (EM), Tufts University (emma.morgan@tufts.edu)

Introductions: (Slide 2)

KA: Hi everyone, I'm Kate Aloisio, the Assistant Director of Institutional Research at Smith College. Thank you [again] for joining us. We are very excited to be sharing this project with you and we were thrilled to be asked to share this talk through a webinar. We are also happy to now have our third collaborator with us who wasn't able to make it to the AIR Forum.

KF: Hello everyone, I'm Kathy Foley and I'm the one who held down the fort back home while Kate and Emma presented in Denver. I'm the Associate Director for Analytics in Smith College's Office of Institutional Research. For those who may be unfamiliar, Smith is a private liberal arts college for women located in western Massachusetts. We have around 3,000 students, of which about 2,500 are undergraduates.

EM: Hello everyone, I'm Emma Morgan, Senior Data Analyst in the Office of Institutional Research at Tufts University. Tufts is a mid-sized private research university in Massachusetts. Our main campus is located in Medford just outside Boston. We have a health sciences campus and a school of fine arts that are both located in Boston, and the Tufts Cummings School of Veterinary Medicine is located in Grafton, about 45 miles outside the city. In Fall 2018 we had close to 12,000 students, half undergraduate and half graduate/professional students. Our mascot is Jumbo the Elephant.

Before I start, we would like to thank AIR for inviting us to give this talk and share our work with the IR community. We also want to thank Tufts University and Smith College for their support throughout this project. Our collaboration and time spent on this project would not have been possible without the support of our institutions.

Why Compile IPEDS?: (Slide 3)

EM: This project and our partnership working together have been ongoing for four years. At the start, we were interested in IPEDS data for its peer comparison potential. IPEDS collects data from our institutions and provides a wealth of information. Both of our institutions had an interest in being able to more easily access this information. Each of us also has an interest in R, and we recognized the potential for R to make IPEDS data more useable. From the outset we recognized that our interests in having easier access to IPEDS data were not unique to Smith or Tufts: but were likely shared by other institutions.

There are many uses for IPEDS data. The scope of available data and relative consistency over time and across institutions makes it useful for showing trends over time and comparisons with peer institutions. I invite you to think of how you have used or would like to use IPEDS data in your own work.

The challenge of using IPEDS data is getting all the information you need into a format that is easy to work with. Our goal is to get IPEDS data in front of stakeholders. This should be relatively easy since the data are available, but unfortunately, it always takes longer and requires more work than we would like. Through the IPEDS data center, it's possible to download data from a single survey and year for all institutions, but this limits trends over time. Alternatively, you can select a subset of schools and metrics and download a file with multiple years of data. This format is still difficult to work with, and it requires a fair amount of effort to clean the variables names and reshape the data to be useable. Because the user selects specific schools and metrics, this method is also inflexible to changes in data needs.

For those of you unfamiliar with data downloads from IPEDS or who have blocked out the traumatic memories, we will be showing what data looks like when we download it from IPEDS and what we have done to try and make this better, more friendly, and usable for everyone in this room.

Use Case Scenario: (Slide 5)

EM: To illustrate what we've done, we're going to take a walk through an IR scenario that is likely familiar to many of us. Your boss approaches you one morning with this question: "Could you compare our institution to ten specific peer schools on select key metrics from IPEDS?"

(Slide 6)
You smile and say "Sure thing!" and rush off into the IPEDS weeds, full of optimism and confidence that you know exactly what you're doing.

(Slide 7)
In no time at all, you find yourself in the IPEDS data center. You select your schools, check the boxes for the metrics you need, and download a year of IPEDS data for those specific schools. You take a little time to clean up the downloaded file and produce a lovely chart using the software of your choice. Feeling confident in your job well done, you take the product to your boss.

(Slide 8)
And then your boss says "This is fantastic! Now can you show the trend for 10 years? And can you add one more peer school?"
(Slide 9)
At this point, we know that this simple request to add another school and show trends over time is going to require us to go back into the data center and repeat the entire process, all the while hoping that we won't select too many variables and have to split into multiple requests. You mentally collect yourself before starting over.

This is exactly the situation we're trying to avoid. We want to save this poor elephant from death by salt and sand. Kathy, Kate and I began working on this project to figure out how we could streamline and improve the ways we work with IPEDS data.

Before we can save our elephant friend, let's go on a little field trip to the IPEDS Data Center. We don't want this to be too traumatic; know that we're in this together and, if you've ever struggled with the data center, you're not alone.

Downloading IPEDS Data using "Compare Institutions": (Slide 10)

As I mentioned earlier, there are two main ways to pull data from the IPEDS data center. One option is to use "Compare Institutions:" Take the time to select each school in your list, search for and select each variable, choose the years you need for each variable, and hope that you don't make any errors. I can't speak for any of you, but I have spent many hours trying to use this "compare institutions" tool only to realize I made a terrible, terrible mistake (or a minor mistake) that requires me to start over. At a certain point, these stops and starts become too much, and it seems like a better option to just download ALL the data so we don't have to keep returning for every change.

(Slide 11)
Our second download option is to download data as "Complete Surveys." With this option, we can get a single year of data from a given survey (or at least a large section of a survey) for all reporting institutions. At first glance, this seems like it could be a good option: Complete surveys have data for all institutions, so adding an additional peer or looking at an addition metric is no longer an issue.

(Slide 12)
However, each complete data file has a single year in isolation, which makes it difficult to look at trends

over time. The data is also not user-friendly: Variable names are not readable human text; we have these odd imputation variables that start with “X”, and many of the variables require a separate codebook to understand. Years of experience in IR might embed some of these variable names and value labels in our minds, but this is not something that is easy and quick to use.

(Slide 13)
These are the options available to us as of now. Our elephant has raised himself out of the sand, but he’s still struggling to find solid ground.

There is a Better Way (slide 14):

Neither “compare institutions” nor “download complete surveys” gives us quite what we want. My colleagues at Smith and I were not satisfied with these options, and we were motivated to find a better solution. We have partnered to write an R script that produces a single longitudinal CSV for each IPEDS survey. Our method starts with the complete survey files, so we have information for all institutions. We also leverage the dictionary excel files for each survey and year to get the metadata we need. Our code combines multiple years of complete data files, does the same thing for the corresponding dictionary files, and then combines the data and dictionary to create readable column headings and value labels. We’re showing this to give you an idea of what is happening behind the scenes on our end, but you won’t need to worry about the details.

(Slide 15)
The important part is that what starts out looking like this….

(Slide 16)
...ends up looking like this.

Note that the top view is just 2017 data directly from the complete data file. If we look closely at the final product in the bottom half of the screen, we notice a few key differences. The second screenshot has multiple years of data. We can see this with the second and third columns, which have been added as part of the compiling process. The column FILE_NAME in the third column refers to the name of the complete data file downloaded from IPEDS. You will see that we use the revised versions when available. Since this is compiled admissions data, we pull from the adm admissions files beginning in 2014 when it became its own survey and from the institutional characteristics files in earlier years.

The field ACAD_YEAR is also added during our process. You can think of this as the fiscal year or academic year spring of the data. This is not the submission year; we wanted the data year so that we could connect data from the same time period across surveys. You can read more about this in the FAQ page of the shiny app.

We also see that the column headers have changed and are now understandable. Rather than APPLCN we now see “Applicants total”. We’ve also applied value labels where they exist. The column ADMCON7 has been renamed “Admission test scores”. This variable had value labels with 1 being “Required”’; we see the value replaced with the label in our compiled data.

We’ve also removed the imputation columns starting with “X” since we aren’t using these for our analyses or visualizations.

(Slide 18)
Although we began with the complete data files, you likely don’t need data for all 7,000 institutions. Towards this end we have created a web app that allows you, the IR professional, to subset our compiled CSV data files to include only the institutions of interest to you. You can provide a list of which institutions you want and download a more manageable file with only the data you want.

(Slide 19)
Hopefully, I didn’t lose anyone when I mentioned earlier that we are doing this work in R. Although R is

our tool working behind the scenes to compile data and create this web app, you do not need to know any R to use this tool.

Kate is now going to show us a demonstration of the IPEDS Data Compiler.

Demo:

KA: Welcome to our IPEDS Data Compiler app. It is a Shiny app hosted on shinyapps.io and if you are not familiar with either of those tools then all you need to know is that it is a website accessible through any internet browser.

First off, this app is a work in progress and none of us have app development in our background so we welcome feedback and questions and can be easily reached at iwdapplication@gmail.com

Currently, the app has two tabs the first being the 4-step process to obtain the data and the second being the Frequently Asked Questions.

[Click on FAQ]

We have attempted to anticipate questions about using the app, the resulting file, and about IPEDS in general, but if your question is not found here, please send us an email and we may add it in the future.

[Click on Compiler]

Now we are going to go through the four steps.

First, we are going to upload our list of peers as a CSV. The only variable that has to be included in the file is a column called UNITID that contains the institutions NCES ID.

When you click on Browse [Click on Browse] you will navigate to your CSV and select it [Select CSV].

Then a preview of your list of peers will appear. Again, the only variable that you must include is UNITID. If you include Institution, we will place that variable at the beginning of the longitudinal file. Otherwise, any variables that are in your list of peers will be merged onto the back of the file. So, you will not lose any information from the file you upload.

You can see here that we have uploaded a file that contains 10 peer institutions.

Next, in step 2, we select which one of the 36 IPEDS surveys we would like to compile. For this demonstration, we are going to select Applications, admissions, enrollees, and test scores. [Select Applications, admissions, enrollees, and test scores]

Then in Step 3, we are going to press the “Dominate the World” button and patiently wait. The waiting time depends on the size of your peer list and which survey you select. Completions for many institutions will take a few minutes. [Select Dominate the World]

Once your file loads, a preview of the first 10 columns will appear. The first will be the UNITID and the second will be the name of the institution if you have included it in your list of peers CSV. The next three variables are ones we created, and the remainder of the columns are the fields from the survey. You can start to see here that we have the Dictionary variable names and a version of the non-number fields with the original value and the code’s readable label.

I will pause to note that those are the only two big cleaning steps that we took. In our attempt to make this as flexible as possible we have tried to make no analytical decisions and leave that to the institution to decide. We also currently are not adding any calculated variables.

Lastly, in step 4 select Download CSV to save the file. We have set it up that the file will come down as the IPEDS table name underscore compiled, you can, of course, change this to whatever works for you. [Select Download csv]

Now that you have your file you can bring it into your analytic tool of choice. For this demonstration, we will open it in Tableau. [Open Tableau workbook]

Now you can see all of the variable names. And you can start working with the file. So, for instance, you can build a Comet viz that looks at measures at two points in time. Here we have Admit Rate on the y-axis and Yield on the x-axis. Each line and point is an institution and you can see the 10 peer institutions.

But oh no, I've forgotten to add Smith College. How incredibly foolish. I'm going to have to pull all of the data again. But since I only forgot to add a peer all I now need to do is add Smith College to my list of peers.

First, I'm going to add Smith College's unit id to my list of peers. [Open peer list, under unitid add 167835] I'm also going to add Smith College into the institution field so my Viz will have the institution name but as you will see I don't have to have any of the other information. [Under Institution add "Smith College", save csv]

Now in the app, I am going to go back to step 1 and click on "Browse" [Click on Browse] and repoint it to my updated peer list. Now in the preview table, I have 11 entries and the text has been updated to 11 peer institutions.

We are going to keep the same survey in Step 2 and Select Dominate the World in Step 3. [Click on Dominate the World]

Once the preview loads, we can now see that row 7 contains Smith College and now we can download and save the file [Click on Download csv]

Then we are going to return to the Viz and select Refresh [Return to Viz and select Refresh] and now Smith appears with its own comet and now you have more time to
dedicate to analyzing the data, or other hobbies you may enjoy.

[Advance the Slide]

Future Design Enhancements:

KF: We're pleased to share this tool with you all. But we'd like to note again that this is still a work in progress. We already have some ideas of our own for ways to improve this in the future and we welcome suggestions from those who are using the app as well. Here are just a few of the ideas we are considering for future iterations:

(1) First off, we'd like to allow the app user to indicate which specific years of data they want. Right now, you simply get all the years that we have compiled (usually about 10) and if you don't want that many you have to filter them out of the resulting CSV on your end.

(2) Now as I mentioned the files currently contain only about 10 years of data each, but we'd like to be able to go even further back, historically. The thing that stopped us is that prior to a certain date the IPEDS dictionary files download as HTML and not csvs and none of us has experience working with HTML files in R. But if anyone out there has expertise in that area we'd love to partner with you to be able to expand these longitudinal files further back in time.

(3) Third, we'd like to expand the options for specifying your peer institution list. Currently, you need to provide a list of specific UNITIDs to subset the giant files, but we understand that not everyone

has a pre-defined set of peer institutions, so we are considering adding functionality to allow the user to select their peers via the app based on IPEDS metrics. For example, perhaps you want to download the IPEDS admissions survey data for only those schools with an admit rate similar to your institution's. Additionally, we may look into building into the app the capacity to allow the user to manually specify institutions by name.

(4) Finally, we'd like to automate the refreshing of the IPEDS data when new surveys are released. Because the current version of the app is merely sub setting files that we have previously compiled, cleaned, and stored, this iteration of the app relies on the three of us to periodically update those compiled files with new data. In the future, we think it would be worthwhile to rewrite the app so that it is scraping data directly from the NCES site, compiling, cleaning and sub setting the file on the spot.

Wanna Use it?
KF: OK, so do you want to use it? Now that you've seen the app in action and heard about where we are thinking of taking it in the future, we'd love for you to try it out and provide us with feedback. All you need is your CSV peer list of UNITIDs and this link.

And if you have colleagues who'd like to use it, you can feel free to pass the link on to them. It's our hope that the app, along with the FAQs, is self-explanatory enough that any IR professional can get started with it, but if your colleagues are interested in this presentation, AIR Members will be able to access the recording on the AIR website.

Also, Kate and I will be presenting this work in person at Northeast AIR in Newport RI in November so if you or your colleagues will be there you can come see us then.

Questions?
KF: Now we're happy to take any questions you may have for us at this point.