Instant Comparison: IPEDS Survey Files Working for You

Emma Morgan
Tufts University
Office of Institutional Research & Evaluation
June 3, 2016
Glancing ahead (and in the review)

Special thanks to Kathleen Foley and Kate Aloisio in the Office of Institutional Research at Smith College for sharing original code!
Motivating Question

How can we use publicly available data from IPEDS to create longitudinal peer comparison data sets?
Challenge: Downloading data

Look up an institution

1. Select Institutions

How would you like to select institutions to include in your data file/report?
- By Names or UnitIDs
- By Groups
- By Variables
- By Uploading a File

Enter either an institution name or UnitID (or a comma separated list of UnitIDs) in the text box below. As you begin typing, a list of matching institutions will appear. You can select a single institution by clicking on it from the list, or, if you want all institutions on the list, click "Select".

Institution Name

Tufts
160148: Tufts University, Medford, MA

Select
Custom data files

Step 1: Select your institutions
Custom data files

Step 2: Select (a) year and (b)variables of interest
## Custom data files

### Step 3: Download the data file

<table>
<thead>
<tr>
<th>Year</th>
<th>Institution</th>
<th>Undergraduate Total</th>
<th>Degree/Certificate Seeking</th>
<th>First-Time</th>
<th>Non-Degree/Certificate Seeking</th>
<th>Continuing</th>
<th>Total Men</th>
<th>Total Women</th>
<th>Total Time Men</th>
<th>Total Time Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>Boston College</td>
<td>14117</td>
<td>6559</td>
<td>7558</td>
<td>13024</td>
<td>5505</td>
<td>7115</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Boston College</td>
<td>9856</td>
<td>4576</td>
<td>5280</td>
<td>9526</td>
<td>4416</td>
<td>5120</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Boston College</td>
<td>5460</td>
<td>2558</td>
<td>5090</td>
<td>5545</td>
<td>436</td>
<td>5512</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Boston College</td>
<td>2295</td>
<td>1016</td>
<td>1279</td>
<td>2184</td>
<td>1013</td>
<td>1271</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Boston College</td>
<td>7140</td>
<td>3577</td>
<td>3913</td>
<td>7021</td>
<td>3010</td>
<td>7761</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Boston College</td>
<td>142</td>
<td>69</td>
<td>73</td>
<td>142</td>
<td>69</td>
<td>73</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Boston College</td>
<td>7046</td>
<td>3068</td>
<td>3738</td>
<td>6939</td>
<td>3231</td>
<td>6688</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Boston College</td>
<td>371</td>
<td>138</td>
<td>133</td>
<td>371</td>
<td>138</td>
<td>133</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Boston College</td>
<td>4461</td>
<td>1933</td>
<td>2528</td>
<td>3480</td>
<td>1499</td>
<td>1939</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Brown University</td>
<td>9181</td>
<td>4489</td>
<td>4692</td>
<td>8795</td>
<td>4097</td>
<td>1399</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Brown University</td>
<td>6540</td>
<td>3120</td>
<td>3390</td>
<td>6355</td>
<td>3049</td>
<td>5206</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Brown University</td>
<td>6264</td>
<td>3057</td>
<td>3207</td>
<td>6120</td>
<td>3041</td>
<td>3200</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Brown University</td>
<td>1859</td>
<td>764</td>
<td>795</td>
<td>1559</td>
<td>764</td>
<td>795</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Brown University</td>
<td>4765</td>
<td>2233</td>
<td>2422</td>
<td>4682</td>
<td>2277</td>
<td>2405</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Brown University</td>
<td>55</td>
<td>33</td>
<td>22</td>
<td>55</td>
<td>33</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Brown University</td>
<td>6050</td>
<td>2289</td>
<td>2390</td>
<td>6020</td>
<td>2244</td>
<td>2333</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Brown University</td>
<td>340</td>
<td>185</td>
<td>154</td>
<td>340</td>
<td>185</td>
<td>154</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Columbia University</td>
<td>2593</td>
<td>1339</td>
<td>1254</td>
<td>2593</td>
<td>1339</td>
<td>1254</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Columbia University</td>
<td>2796</td>
<td>1356</td>
<td>1393</td>
<td>2796</td>
<td>1356</td>
<td>1393</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Columbia University</td>
<td>8100</td>
<td>4247</td>
<td>3853</td>
<td>7496</td>
<td>3926</td>
<td>7578</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Columbia University</td>
<td>3533</td>
<td>1795</td>
<td>1798</td>
<td>3533</td>
<td>1795</td>
<td>1798</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Columbia University</td>
<td>6947</td>
<td>3925</td>
<td>3022</td>
<td>6947</td>
<td>3925</td>
<td>3022</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Columbia University</td>
<td>598</td>
<td>331</td>
<td>267</td>
<td>598</td>
<td>331</td>
<td>267</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Columbia University</td>
<td>5982</td>
<td>3201</td>
<td>2781</td>
<td>5982</td>
<td>3201</td>
<td>2781</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Columbia University</td>
<td>15489</td>
<td>8549</td>
<td>11040</td>
<td>15489</td>
<td>8549</td>
<td>11040</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Columbia University</td>
<td>14062</td>
<td>7024</td>
<td>7028</td>
<td>14062</td>
<td>7024</td>
<td>7028</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Cornell University</td>
<td>14055</td>
<td>6979</td>
<td>7076</td>
<td>14162</td>
<td>6976</td>
<td>7066</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Cornell University</td>
<td>3225</td>
<td>1577</td>
<td>1648</td>
<td>3225</td>
<td>1577</td>
<td>1648</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Cornell University</td>
<td>10370</td>
<td>5492</td>
<td>5568</td>
<td>10370</td>
<td>5492</td>
<td>5568</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Cornell University</td>
<td>954</td>
<td>476</td>
<td>478</td>
<td>954</td>
<td>476</td>
<td>478</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Cornell University</td>
<td>10416</td>
<td>5327</td>
<td>5289</td>
<td>10416</td>
<td>5327</td>
<td>5289</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Cornell University</td>
<td>87</td>
<td>45</td>
<td>42</td>
<td>87</td>
<td>45</td>
<td>42</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>Cornell University</td>
<td>7097</td>
<td>3448</td>
<td>3549</td>
<td>7097</td>
<td>3448</td>
<td>3549</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Custom data files

Step 4:
Repeat for every year of interest and set of variables
Custom data files

Challenges & frustrations

• What if we use different comparison institutions?
  • Example: Engineering has a different set of comparison institutions than is traditionally used; grad/professional schools might be interested in different comparison institutions; we might want the flexibility of looking at different institutions?

• What if we are interested in different variables?
  • Example: I downloaded total enrollment and full-time enrollment by gender, but now I also am interested in enrollment by full-time/part-time status and race/ethnicity

• What if we need additional years of data?
Alternative: Complete data files

Step 1: Choose year and survey

**Years & Surveys**

- **2014**
- Fall Enrollment

Data files are available in ZIP format.
Complete data files

Step 2: Download data file and dictionary

Complete Data Files

Data files are available in ZIP format.

<table>
<thead>
<tr>
<th>Year</th>
<th>Survey</th>
<th>Title</th>
<th>Data File</th>
<th>Stata Data File</th>
<th>Programs</th>
<th>Dictionary</th>
</tr>
</thead>
</table>
Step 4:
Repeat for every year of interest
Complete data files

Challenges & frustrations

• One file per year
  • Longitudinal comparisons necessitate compiling data files

• Size of complete survey files
  • Complete survey files are huge and can be unwieldly to work with
  • For any institution, much of this data will not provide useful peer comparisons
    • Example: Enrollment at Tufts

• Column headers
  • Options: constantly reference the dictionary, or take time to rename

• Certain data columns have additional reference tables
  • Example: Student Level (EFALEVEL)
Complete data files: Example

Enrollment by race/ethnicity, gender, attendance status, and level of student: Fall 2014
Compiling Data
R to the rescue
Function: IPEDS_clean_merge()
Goals of the program

- Create a more manageable data file by keeping only institutions and metrics that are of interest
- Add institution names in addition to UNITID for easier reference
- Replace column headers with text description
- Replace any reference columns (i.e. student level for fall enrollment) with reference description
  - Fall Enrollment: EFALEVEL 1 → “All Students, Total”
- Merge across years to create longitudinal data set
- Automate so that files can be updated when new data are released
Don’t fear the code!
Getting started with R and R Studio

R version 3.2.2 (2015-08-14) -- "Fire Safety"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

Command line only – not very conducive to learning or exploring
Getting started with R and R Studio

View and edit code

See what is loaded and what functions and/or data are ready for use

Command Line

Load packages, get help, see graphs, etc.
Necessary input

IPEDS_clean_merge <- function(inputDirectory, varDictionary, peerFile, outputDirectory, outputName, datedDirectory, freq_table, lookup_dict) {

Necessary input

• **inputDirectory**
  - Filepath to folder containing complete survey files (.csv)

• **varDictionary**
  - “varlist” page from data dictionary, saved as .csv and read into R

• **peerFile**
  - List of comparison institution names and UNITID (IPEDS identifier)

• **outputDirectory**
  - Filepath to destination folder of compiled dataset

• **outputName**
  - Name of the compiled file

• **datedDirectory**
  - For record keeping only; location of files from every day the code is run

• **freq_table** (True/False)
  - Does the dictionary have a “Frequencies” tab with additional lookup values?

• **lookup_dict** (optional)
  - If an additional lookup table exists, data dictionary “Frequencies” tab saved as .csv and read into R
### Input: varDictionary

**Sample Format**

<table>
<thead>
<tr>
<th>varnumber</th>
<th>varname</th>
<th>DataType</th>
<th>Fieldwidth</th>
<th>format</th>
<th>imputationvar</th>
<th>varTitle</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UNITID</td>
<td>N</td>
<td>6</td>
<td>Cont</td>
<td></td>
<td>Unique identification number of the institution</td>
</tr>
<tr>
<td>20166</td>
<td>EFALEVEL</td>
<td>N</td>
<td>2</td>
<td>Disc</td>
<td></td>
<td>Level of student</td>
</tr>
<tr>
<td>21985</td>
<td>LINE</td>
<td>N</td>
<td>2</td>
<td>Disc</td>
<td></td>
<td>Level of student (original line number on survey form)</td>
</tr>
<tr>
<td>21990</td>
<td>SECTION</td>
<td>N</td>
<td>1</td>
<td>Disc</td>
<td></td>
<td>Attendance status of student</td>
</tr>
<tr>
<td>21991</td>
<td>LSTUDY</td>
<td>N</td>
<td>1</td>
<td>Disc</td>
<td></td>
<td>Level of student</td>
</tr>
<tr>
<td>20286</td>
<td>EFTOTLT</td>
<td>N</td>
<td>6</td>
<td>Cont</td>
<td>XEFTOTLT</td>
<td>Grand total</td>
</tr>
<tr>
<td>20241</td>
<td>EFTOTLM</td>
<td>N</td>
<td>6</td>
<td>Cont</td>
<td>XEFTOTLM</td>
<td>Grand total men</td>
</tr>
<tr>
<td>20246</td>
<td>EFTOTLW</td>
<td>N</td>
<td>6</td>
<td>Cont</td>
<td>XEFTOTLW</td>
<td>Grand total women</td>
</tr>
<tr>
<td>24432</td>
<td>EFAIANT</td>
<td>N</td>
<td>6</td>
<td>Cont</td>
<td>XEFAIANT</td>
<td>American Indian or Alaska Native total</td>
</tr>
<tr>
<td>24434</td>
<td>EFAIANM</td>
<td>N</td>
<td>6</td>
<td>Cont</td>
<td>XEFAIANM</td>
<td>American Indian or Alaska Native men</td>
</tr>
<tr>
<td>24436</td>
<td>EFAIANW</td>
<td>N</td>
<td>6</td>
<td>Cont</td>
<td>XEFAIANW</td>
<td>American Indian or Alaska Native women</td>
</tr>
</tbody>
</table>

Source: File documentation for enrollment by race/ethnicity, gender, attendance status, and level of student: Fall 2014, “varlist”
Necessary input

- **inputDirectory**
  - Filepath to folder containing complete survey files (.csv)
- **varDictionary**
  - “varlist” page from data dictionary, saved as .csv and read into R
- **peerFile**
  - List of comparison institution names and UNITID (IPEDS identifier)
- **outputDirectory**
  - Filepath to destination folder of compiled dataset
- **outputName**
  - Name of the compiled file
- **datedDirectory**
  - For record keeping only; location of files from every day the code is run
- **freq_table** (True/False)
  - Does the dictionary have a “Frequencies” tab with additional lookup values?
- **lookup_dict** (optional)
  - If an additional lookup table exists, data dictionary “Frequencies” tab saved as .csv and read into R
<table>
<thead>
<tr>
<th>varnumber</th>
<th>varname</th>
<th>codevalue</th>
<th>Valuelabel</th>
</tr>
</thead>
<tbody>
<tr>
<td>20166</td>
<td>EFALEVEL</td>
<td>1</td>
<td>All students total</td>
</tr>
<tr>
<td>20166</td>
<td>EFALEVEL</td>
<td>2</td>
<td>All students, Undergraduate total</td>
</tr>
<tr>
<td>20166</td>
<td>EFALEVEL</td>
<td>3</td>
<td>All students, Undergraduate, Degree/certificate-seeking total</td>
</tr>
<tr>
<td>20166</td>
<td>EFALEVEL</td>
<td>4</td>
<td>All students, Undergraduate, Degree/certificate-seeking, First-time</td>
</tr>
<tr>
<td>20166</td>
<td>EFALEVEL</td>
<td>5</td>
<td>All students, Undergraduate, Other degree/certificate-seeking</td>
</tr>
<tr>
<td>20166</td>
<td>EFALEVEL</td>
<td>19</td>
<td>All students, Undergraduate, Other degree/certificate-seeking, Transfer-ins</td>
</tr>
<tr>
<td>20166</td>
<td>EFALEVEL</td>
<td>20</td>
<td>All students, Undergraduate, Other degree/certificate-seeking, Continuing</td>
</tr>
<tr>
<td>20166</td>
<td>EFALEVEL</td>
<td>11</td>
<td>All students, Undergraduate, Non-degree/certificate-seeking</td>
</tr>
<tr>
<td>20166</td>
<td>EFALEVEL</td>
<td>12</td>
<td>All students, Graduate</td>
</tr>
<tr>
<td>20166</td>
<td>EFALEVEL</td>
<td>21</td>
<td>Full-time students total</td>
</tr>
<tr>
<td>20166</td>
<td>EFALEVEL</td>
<td>22</td>
<td>Full-time students, Undergraduate total</td>
</tr>
</tbody>
</table>

…continuing for all values of EFALEVEL, LINE, SECTION, LSTUDY

Source: File documentation for enrollment by race/ethnicity, gender, attendance status, and level of student: Fall 2014, “frequencies”
Function: IPEDS_clean_merge

1. Specify where to find data files (inputDirectory)
2. Set up a dummy that will hold the “cleaned” data files from each year until it’s ready to compile

```r
IPEDS_clean_merge <- function(inputDirectory, varDictionary, peerFile, 
outputDirectory, outputName, datedDirectory, freq_table, lookup_dict) {
  # Specify where R will find your file set
  setwd(inputDirectory)
  # Set up necessary blank lists
  # List to hold survey data from various years
  ds_list <- list()
  # For each file in directory:
  # (1) read in the file
  # Create list of rows to keep based on peers/comparison schools
  # Create list of columns to drop (imputation columns or blank fields for Tufts)
  # Store “good” data in file_list
```
3. For each data file...
   a) Read in the data file
   b) Identify which columns to keep
      • Depending on your purposes, you might change this
      • At Tufts, we delete
         • (a) columns for imputation variables
         • (b) columns that Tufts has left blank
   c) Identify which rows to keep
      Keep rows for Tufts and comparison institutions
3. For each data file (cont)…

d) Add survey year

Columns for fall and fiscal year, generated using the function IPEDS_FY that generates year from the survey file name

e) Add institution name

Use peerFile to lookup institution names based on UNITID

f) Store the cleaned data as an element of the list

```r
# We now want to add columns for Fiscal Year, Fall, and Institution Name
# Create placeholders first
ds_good$FY <- ""
ds_good$Fall <- ""
ds_good$Institution_Name <- ""

# Add FY and "Fall" for joining purposes in Tableau
# FY is based on the type of survey file and when data is collected
# IPEDS_FY function is defined below
survey_details <- IPEDS_FY(filename)
ds_good$FY <- survey_details[1]
ds_good$Fall <- ds_good$FY-1
ds_good$'Survey Code' <- survey_details[2]
ds_good$'Survey Title' <- survey_details[3]
survey <- survey_details[2]

# Add Institution Name
for (i in 1:nrow(ds_good)) {
p <- which(ds_good$UNITID[i] == peerFile$unitid)
ds_good$Institution_Name[i] <- as.matrix(peerFile$'institution name')[p]
}

# Store the good info in our new list
ds_list[[i]] <- assign(paste("ds",i,sep=""), ds_good)
}
4. Compile all years of data into a single dataset
   Each variable column now includes entries for each institution in each year
   ```r
   full_ds <- as.data.frame(rbind.fill(matrix(ds_list)))
   ```
Function: IPEDS_clean_merge

4. Compile all years of data into a single dataset
   Each variable column now includes entries for each institution in each year

   \[ \text{full_ds <- as.data.frame(rbind.fill.matrix(ds_list))} \]
Function: IPEDS_clean_merge

5. If the survey dictionary included a “frequency” table with additional lookup values...
   a. Generate reference tables for these variables using the frequency table
   b. Create additional columns as necessary with descriptions of coded values

```r
if (freq_table==TRUE) {
    full_ds <- lookup(full_ds, lookup_dict)
}
```
Function: IPEDS_clean_merge

6. Replace column name variable codes with descriptive titles
   Use the variable dictionary to substitute full text descriptions for each variable

7. Set the output directory

8. Export the compiled data set from R as a .csv

```r
# Replace variable codes with full header titles
for (m in 1:ncol(full_ds)) {
  if (names(full_ds)[m] %in% varDictionary$varname) {
    a <- which(names(full_ds)[m] == as.matrix(varDictionary$varname))
    names(full_ds)[m] <- as.matrix(varDictionary$varTitle)[a]
  }
}

# Set directory to where you want the file to output
setwd(outputDirectory)

# Write full .csv to output directory
write.csv(full_ds, paste(outputName, ".csv", sep=""), row.names = FALSE, na="")

# Set dated directory to save copies of files
setwd(datedirectory)
write.csv(full_ds, file = paste(outputName, Sys.Date(), ".csv", sep=""), row.names = FALSE, na="")
}
Your longitudinal peer comparison data file is ready to use!
Continuing Challenges

- Format of IPEDS data files may change
  - Example: Undergraduate admissions
    - Pre-2014, undergraduate admissions statistics included in Institutional Characteristics data file
    - In 2014, undergraduate admissions released as independent data file
    - The variable names are consistent, so we can still merge on these fields

- Automation requires checking for consistency

- Variables may change
  - Example: Race/ethnicity categorization before and after 2008
  - Older documentation is available as a web link, but is not downloadable as an Excel file
  - For historical data, it may be necessary to create a compiled dictionary if older variable names differ from what appears in the current documentation file
Possible Next Steps

- Use IPEDS directory file to generate peer lists
  - Carnegie Classification, size, region, etc.
- Extend to other data sources
  - College Scorecard
- Use the Shiny package in R to make the program more interactive and user-friendly
  - User can select data file directory, choose the correct dictionary and specify where to export files from a dialogue box rather than specifying all file paths in text
Questions?

Emma Morgan
Research Analyst, Tufts University
emma.morgan@tufts.edu

Special thanks to Kathleen Foley (Associate Director for Analytics) and Kate Aloisio (Research Analyst) from the Office of Institutional Research at Smith College for their ideas and code that jumpstarted this project.